International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 05 | May-2016 www.irjet.net p-ISSN: 2395-0072

Speech to Text Converter Using Gaussian Mixture Model(GMM)

Virendra Chauhan[1], Shobhana Dwivedi[2], Pooja Karale[3], Prof. S.M. Potdar[4]

[1,2,3]B.E. Student [4]Assitant Professor
[1,2,3,4]Department of Electronics and Telecommunication Engineering
[1,2,3,4]Sinhgad academy of Engineering, Pune 443001

Abstract - : *The fundamental of speech dates back in time wherein humans began to communicate with one other. Over the time it was realized that it is an effective medium of commutation. Later on, it acquired various forms and thus different languages came into existence. On close analysis it has been found that there are over hundreds of different languages; English being the most widely used. Accordingly, most the formal talks are in English. Most of the communication devices these days like security devices, home appliances, mobile phones, ATM machines, computers and hotels use speech processing. Our project focuses on establishing an interface between these two. The human computer interface has been developed in order to communicate and interact with ones who are suffering from different types of disabilities. Speech-to-Text Conversion (STT) system is advantageous for deaf and dumb people. It is also used in our day to day lives. The main aim of our system is to convert input speech signals into text as output for the disable students in the educational as well as industrial fields. This paper is presented to extract features of the speech signal by Mel-Frequency Cepstral Coefficients(MFCC) for multiple isolated words. Gaussian Mixture Model (GMM) is used to train the audio files to get the spoken word recognized. Database is created by storing the speech signal in MATLAB.*

Key Words: Feature Extraction, MFCC , Gaussian Mixture Model(GMM), Expectation-Maximization(EM), Maximum Liklehood Estimation (ML).

1.INTRODUCTION

We communicate with each other in many ways such as expression, eye contact, gesture, and speech. The simple mode of communication among people is speech and also the most natural and efficient form of exchanging information amongst them in speech. Speech-to-text conversion system is extensively used in many applications. In the field of education, Speech-to-text conversion system or speech recognition system is more productive for deaf and dumb students. Speech recognition is a challenging part in speech processing systems. The significant part of the system is Feature Extraction. There are different types of feature extractions methods. In recent years of research, many feature extraction techniques are used such as Linear Predictive Coding(LPC), Linear Discriminant Analysis(LDA), Independent Component Analysis (ICA), Principal Component Analysis(PCA), Cepstral Analysis and Mel-frequency cepstral(MFCCs), Kernal based feature extraction, Wavelet Transform and spectral subtraction. Mel Frequency is based on the characteristics of the human ear's hearing, which uses a discontinuous frequency unit to reproduce the human acoustic system. Mel-Frequency scale is used to extract features of the input speech signal. These cepstral features provide the accuracy of recognition to be systematic for speech recognition and emotion recognition system. Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Vector Quantization (VQ), Artificial Neural Network (ANN), are various techniques which are used by the researchers in recognition. Amongst all the algorithms, GMM speech recognition algorithm is most superior in many applications. At present, Speech-to-text convertor systems is mostly used in many mobile phones, computers and control systems. Accordingly, Cepstral Coefficients (MFCC) method is used. This Speech-to-text convertor systems are more useful in our day-to-day activities . In the paper, GMM and MFCC are implemented by using MATLAB.

2. METHODOLOGY

2.1 Silence and noise removal

Speech signal is divided into voiced and unvoiced signals. Voiced signals are periodic and one-third part of the speech signal is voiced signal which is important for intelligibility. Unvoiced signals are non-periodic, undesirable sound signals produced by the vocal tract. Unvoiced signals can also be considered as noise or the silence part of the speech signal. Silence can be removed by setting a threshold voltage which is dependent on background noise.

2.2 MFCC

The most easiest and prevalent method to extract spectral features is calculating the Mel-Frequency Cepstral Coefficients (MFCC) from human voice. It is one of the most popular methods of feature extraction used in speech recognition systems. It is based on frequency domain using the Mel scale which is based on the human ear scale. Time domain features are less accurate than the frequency domain features. The main aim of feature extraction is to reduce the size of the speech signal before the recognition of the signal. Steps involved in feature extraction are pre-emphasis, framing, windowing, fast fourier transform, Mel-frequency filtering, Logarithmic function and Discrete Cosine Transform etc.

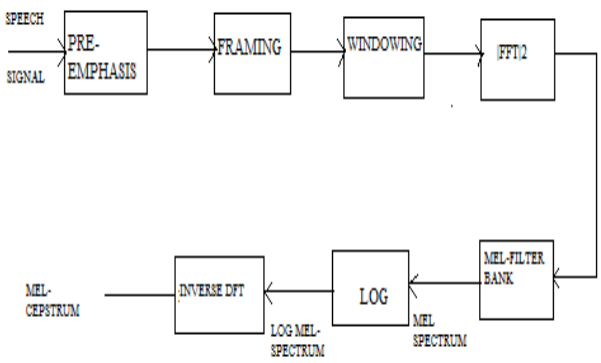


Fig 1 : Block diagram of MFCC

The first step in MFCC is pre-emphasis which is used to boost the high frequencies of a speech signal which are lost during speech production. Pre-emphasis is needed because high frequency components of the speech signal have small amplitude with respect to low frequency components. Therefore higher frequencies are artificially boosted in order to increase the signal-to- noise ratio. Next, is framing which is used to block the frames obtained by analog to digital conversion (ADC) of speech signal . The number of samples in each frame is chosen as 256 and the number of samples overlapping between adjacent frames is 128. Overlapping frames are used to acquire the information from the boundaries of the frame. Due to discontinuities at the start and the end of the frame causes undesirable effects in the frequency response, so windowing is used to eliminate the discontinuities at the edges. Hamming window is used which introduces least amount of distortion. Generalized hamming window equation is

$$w(n) = \alpha - \beta cos\left(\frac{2\pi n}{N-1}\right)$$

After windowing, Fast Fourier Transform(FFT) is measured for every frame to extract the frequency components of the signal in time domain. Speech signal does not follow linear frequency scale used in FFT. Hence Mel-scale is used for feature extraction which is directly proportional to the logarithm of linear frequency. Equation is used to convert linear scale frequency into Mel-scale frequency.

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

Triangular bandpass filters are used to extract the spectral envelope. 20 filters are used. Log is applied to the absolute magnitude of the coefficients of which is obtained after Mel-scale conversion. Discrete cosine transform(DCT) converts the Mel-frequency domain into time domain.

$$C(k) = \sum_{m=0}^{M-1} S(m) cos(\pi k (m + 1/2)/M), 0 \leq k < K$$

The value of K ranges between 8 and 13. We choose K as 13 and hence we obtain 13 coefficients for each frame.

2.3 GAUSSIAN MIXTURE MODEL

The Gaussian Mixture Model(GMM) is a parametric probability density function which is represented as a weighted sum of Gaussian component densities. It is used as a parametric model of probability distribution of measuring features in biometric systems. Gaussian Mixture Model(GMM) is used as a classifier to compare the features extracted from the MFCC with the stored templates. Gaussian Mixture Model is represented by its Gaussian distribution and each Gaussian distribution is calculated by its mean, variance and weight of the Gaussian distribution. Gaussian Mixture density is weighted sum of M component densities and can be expressed:

$$p(\bar{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\bar{x})$$

bi (x) – component densities, that can be written:

$$b_i(\bar{x}) = \frac{1}{2\pi^{D/2}|\sum_i|^{1/2}} e^{-1/2(x^2-\mu_i)' \Sigma_i^{-1}(x^2-\mu_i)}$$

where i = mean vector

Gaussian Mixture Model is described by the mean vectors, co-variance matrices and mixture weights from all component densities. These parameters are represented by the notation:

$$\lambda = \{p_i, \mu_i \quad i = 1,2,3\,.$$

Every speaker is shown by his GMM and is referred to his model . Plenty of techniques are available for calculating the parameters of GMM. One of the most popular methods is Maximum Likelihood(ML) estimation. It finds out the model parameters which maximize the likelihood of GMM. For T training vectors {x1,. . . . .,xT} the GMM likelihood is given as:

$$P(X|t) = \prod_{t=1}^{T} P(x_t|\lambda)$$

The above equation is a non-linear function of so the iterative Expectation-Minimization(EM) algorithm is used for training and matching. Thus the following parameters are calculated in the iterations:

Mixture weight : $\overline{w}_i = \frac{1}{T}\sum_{t=1}^{T} P(i/$

Mean : $\overline{\mu}_i = \frac{\sum_{t=1}^{T} p(i|\overline{x}_t}{\sum_{t=1}^{T} p(i|\overline{x}}$

Variance : = $\frac{\sum_{t=1}^{T} p(i|\overline{x}_t,\lambda)(x_t-}{\sum_{t=1}^{T} p(i|\overline{x}_t,\lambda)}$

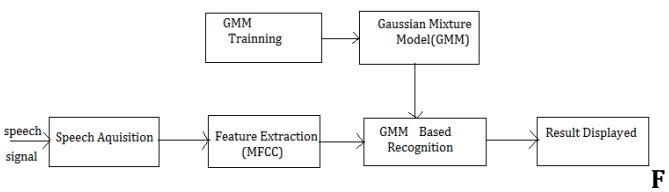
The a posteriori probability for component is given as follows:

$$P(x_t|\lambda) = \frac{wip(x_t|\mu i, \sigma i^2)}{\sum_{k=1}^{M} wk\ p(x_t|\mu i, \sigma i^2)}$$

These iterative steps are carried out for matching purposes in real-time and the Euclidean distance is found out between various database, hence a correct match is found.

3. Implementation

The general flow of processing and recognition of the speech signal is shown in fig(2)



F

ig 2 : General block diagram of speech signal processing

The speech signal is recognized by using Gaussian Mixture Model. The speech signal is recorded by using 16-bit Pulse code modulation with a sampling rate of 8KHz and it is stored as a wave file by using sound recorder software in MATLAB. .wav files are converted into speech samples by using MATLAB software's wavered command. The silence part of the speech as well as background noise is removed at the initial stage of processing. A threshold voltage is set to remove the noise and the silence part which is totally dependent on environment with less ambient noise. Then the features of the speech signal is extracted by the MFCC block. The total number of samples chosen in a frame is 256 and overlapping samples with the adjacent frame will be 128. We acquire MFCC cepstral coefficients at the output of MFCC block. In GMM, K-mean algorithm is used to obtain a cluster number specific to each observation vector and sets the centriod of the observation vector. After clustering the model, it returns one centriod for each of the cluster K and refers to the cluster number closest to it. K-mean algorithm is described as the squared distances between each observation vector and its centroids. In the training section parameters of GMM model are produced iteratively by expectation-maximization(EM) algorithm. Euclidean distance is found out between observation vector and its cluster centriods to match the spoken word with the present database. The word which is recognized is displayed as text in the output.

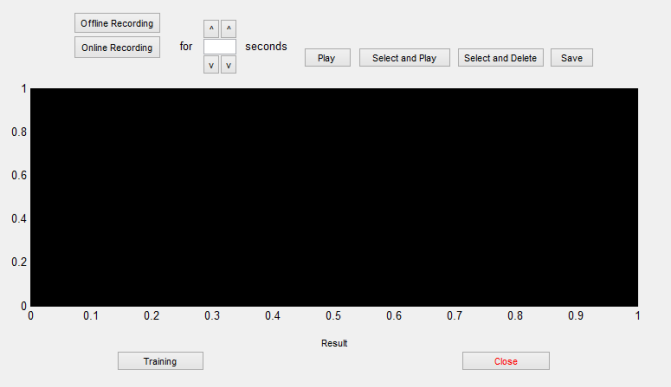


Fig 3 : Graphical user interface

A Graphical User Interface is used for the user to interact with the system with various buttons on it such as online and offline recording, play, training, select and play, close etc. There are two buttons for recording the speech signal which are online and offline recording. In online recording, the time interval is given in which the user has to speak up a word and the speech signal graph will be displayed as well as the spoken word will be displayed as result in text if the spoken word matches with the database present. In offline recording the pre-recorded speech signal is used as input and if it matches with the database present then the output then the result is displayed . Before recording the speech signal, the database stored is trained by clicking the training button.

4. RESULTS

For acquiring the results the speech signal is recoded. The system is trained for multiple words such as Samosa, Dosa , Tea etc. The results for the word Samosa are shown. The speech signal which is recorded for the word Samosa is shown in fig(4).

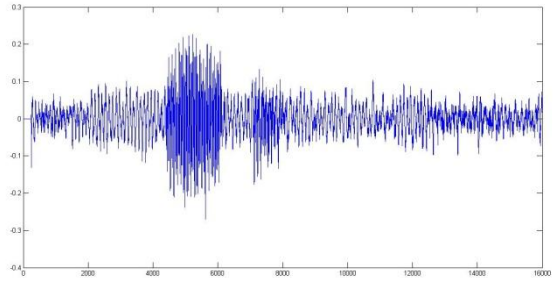


Fig 4 : Speech signal before silence removal

The next fig(5) shows the speech signal after silence and noise removal.

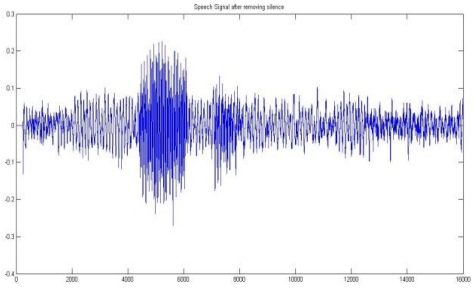


Fig 5 : Speech signal after silence removal

The Mel-frequency cepstral coefficients are plotted for the word Samosa.

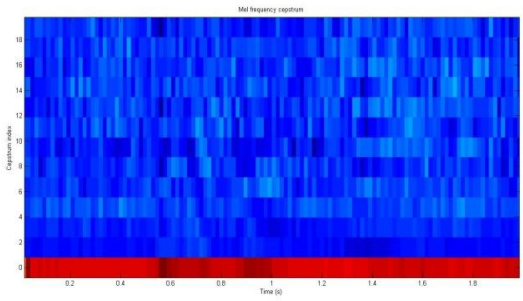


Fig 6 : Mel-frequency cepstrum of word Samosa

And at last the output is displayed in text of the spoken word which is shown in the fig(7).

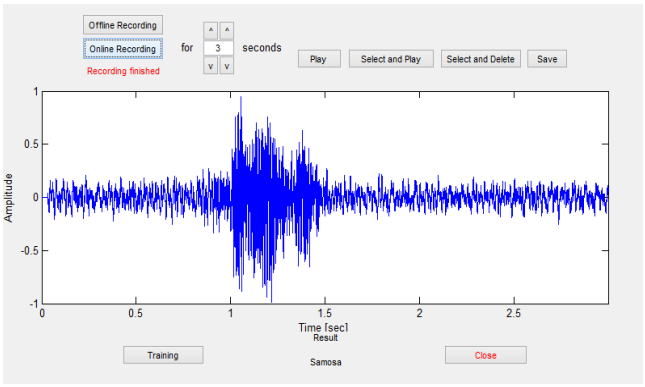


Fig 7 : Result displayed in text as Samosa

The percentage of recognition is shown in the table for multiple words. The accuracy is higher than any other algorithms used.

Table -1: Testing and accuracy results

Train Data	Number of test	Number of correct test	Error	Percentage of accuracy
Samosa	30	21	8	70
Dosa	30	20	10	66.67
Tea	30	23	7	76.67

6. Future Scope

1. Implementation of hardware can be done by using the DSP processors in real-time applications.
2. It can be used in hotels for giving the menu e.g. Samosa, Dosa, Tea etc.

5. CONCLUSIONS

Thus we are able to recognize multiple words such as Samosa, Dosa, Tea and is converted into text by using this paper. This system is suitable with an environment with less ambient noise. The system provides good performance with respect to other systems. It can be concluded that GMM provides more accuracy.

REFERENCES

[1] Douglas A. Reynolds, and Richard C.Rose, "Robust TextIndependent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1, pp. 74-77, 1995.

[2] Reference Book : Speech and Audio Processing(By Dr. Shaila D. Apte, WILEY INDIA Edition).

[3] Manan Vyas, "A Gaussian Mixture Model Based Speech Recognition System Using MATLAB", Signal and Image Processing: An International Journal(SIPU) Vol.4,No.4, August 2013.

[4] Om Prakash Prabhakar, and Navneet Kumar Sahu, "Performance Improvement of Human Voice Recognition System using Gaussian Mixture Model", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014.

[5] Ellis, Daniel. "An introduction to signal processing for speech." The Handbook of Phonetic Science, ed. Hardcastle and Laver, 2nd ed, 2009.