ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 5, Issue 3, March 2016

Speech to Text Conversion Using HMM

G.P.S. Prasanthi[1], K. Sirisha[2], G. Ramya[3],B. Padma[4]
Assistant Professor[1], Student[2, 3, 4]
[1,2,3,4] Department of Electronics and Communication Engineering,
[1,2,3,4] Gayatri Vidya Parishad College of Engineering for Women,
Visakhapatnam, Andhra Pradesh, India

***Abstract*—** "Real time speech to text" can be defined as accurate conversion of words that represents uttered word instantly after speaking." The speech-to-text conversion can provide data entry options for deaf students. The system is also used to find the disorder rate of persons affected with Parkinson's disease by calculating the efficiency of pronunciation. The system takes the speech at run time through a microphone and processes the sampled speech to recognize the uttered text. In the training phase, the uttered digits are recorded using the PCM modulation technique with a sampling rate of 8 KHz and saved as a wave file. MATLAB software uses the wavread command to convert the wav files to speech samples. A HMM Model is used for speech recognition, which converts the speech to text. In this *hidden* Markov model, the state is not directly visible, but output, dependent on the state, is visible. It uses baum-welch algorithm. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states.

***Index Terms*—** *PCM Modulation technique, Microphone, HMM Model, Baum-Welch Algorithm.*

I. INTRODUCTION

Basically, the mode of communication between humans takes place in several ways such as facial expressions, gestures, eye contact and speech. The most important and efficient way of communication is speech [1]. Speech to text conversion is very advantageous and used in various applications areas. It is very useful to deaf students and physically handicapped persons. By calculating the efficiency of conversion, it can be used to find the disorder rate of persons affected with Parkinson's disease. By using this efficiency, one can improve their pronunciation skills. Basically, speech to text conversion (STT) system is distinguished into two types, such as speaker dependent and speaker independent systems [2]. The proposed design is dealing with speaker dependent system. Feature extraction is the important part of speech recognition. There are many methods to extract a feature which includes Principal Component Analysis (PCA), Linear Discriminate Analysis (LDA), Independent Component Analysis (ICA), Linear Predictive Coding (LPC), Cepstral Analysis and Mel-frequency cepstral coefficients (MFCCs) [3].

A Mel-frequency cepstral coefficient (MFCCs) is used in this design. Mel-frequency cepstral features provide the rate of recognition to be efficient for speech recognition as well as emotion recognition system through speech [4]. During recognition various methods are used. Out of those Hidden Markov Model (HMM) is widely used in recognition. In this system MFCC and HMM techniques are designed in MATLAB.

II. LITERATURE SURVEY

Su Myat Mon, Hla Myo Tun, Designed a "Speech-To-Text(STT) System Using Hidden Markov Model(HMM)"[5].The existed technique converts speech to text using MFCC and HMM techniques. But, HMM technique is using only forward procedure to predict the words[5]. So, the prediction efficiency of a particular word was less. The proposed design is using MFCC technique for feature extraction. Also, the HMM is using forward-backward procedure and baum-welch algorithm. The forward-backward procedure gives better prediction of states and the best path can be identified using viterbi algorithm. The baum-welch algorithm gives the idea about iterations and log likelihood of states. This is a large process, future work of conversion may be fast and simple by using modern techniques.

III. SPEECH TO TEXT CONVERSION USING HMM

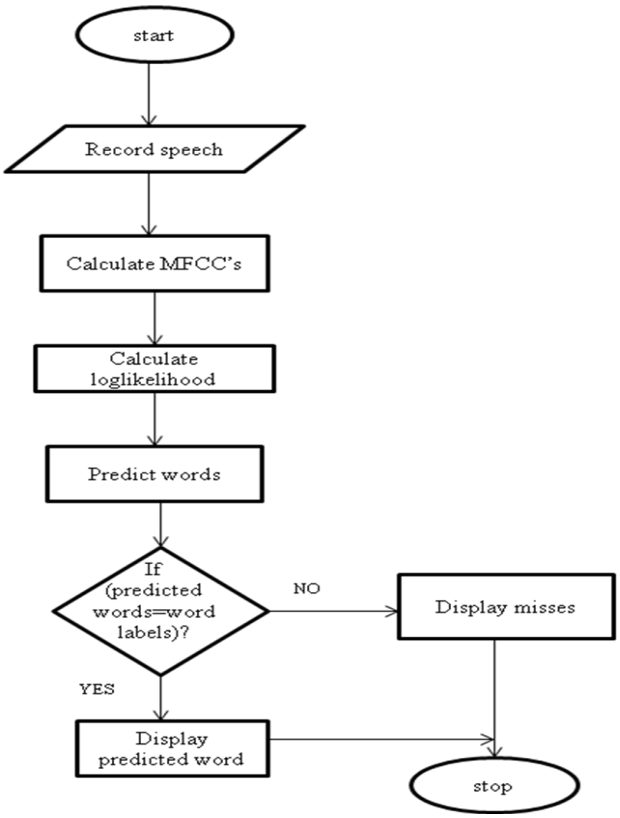

Fig 3.1. Process Sequence

1. MFCC

Feature extraction stage is the most important one in the entire process, since it is responsible for extracting relevant information from the speech frames, as feature parameters or vectors. The aim of feature extraction is to reduce the data size of the speech signal before pattern classification or recognition. The steps of Mel frequency Cepstral Coefficients (MFCCs) calculation are– framing, windowing, Discrete Fourier Transform (DFT), Mel frequency filtering, logarithmic function and Discrete Cosine Transform (DCT).Fig.2 shows the block diagram of MFCC process.

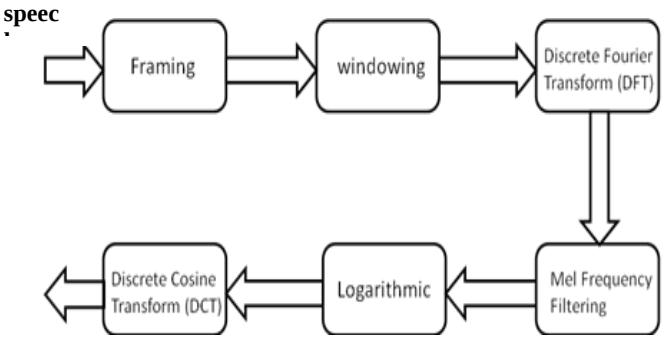

Fig 3.2. Block Diagram of MFCC

FRAMING: The speech signal is divided into a sequence of frames where each frame can be analyzed independently and represented by a single feature vector. Each frame is extracted from speech for every 20-30ms frame time length. So, the speech samples are taken to be continuous. Overlapping of frames is useful to avoid loss of information.

WINDOWING: In order to reduce the discontinuities of the speech signal at the edges of each frame, a tapered window is applied to each frame. The most commonly used window is Hamming window.

DFT: Discrete Fourier Transform (DFT) is used as the Fast Fourier Transform (FFT) algorithm. FFT converts each frame of N samples from the time domain into the frequency domain. The calculation in frequency domain is easier when compared to time domain.

MEL FREQUENCY FILTERING: The voice signal does not follow the linear scale but the frequency range in FFT is so wide. It is perceptual scale that helps to simulate the way human ear works. It corresponds to better resolution at low frequencies and less resolution at high frequencies.

LOGARITHMIC FUNCTION: Logarithmic transformation is applied to the absolute magnitude of the coefficients obtained after Mel-scale conversion. The absolute magnitude operation removes the phase information, making feature extraction less sensitive to speaker dependent variations.

DCT: Discrete cosine transform (DCT) converts the Mel filtered spectrum back into the time domain since the Mel

Frequency Cepstral Coefficients are used as the time index in recognition stage.

2. HMM

A Hidden Markov Model is a Finite State Machine having a fixed number of states. As the speech is a random process, the goal of hmm is to find the parameters of stochastic process in a well defined manner. Hmm process cannot be observed but but can be observed only through another set of stochastic processes that produce the sequence of observations.HMM presents a best way of quantifying speech patterns.

***SPEECH RECOGNITION*:**

A HMM is characterized by 3 matrices viz., A, B and PI.
A - Transition Probability matrix (N x N)
B - Observation symbol Probability Distribution matrix (N x M)
PI - Initial State Distribution matrix (N x 1)
where
N =Number of states in the HMM
M = Number of Observation symbols
we can apply HMM for speech recognition by using following steps:

1. Recursive procedures like Forward and Backward Procedures exist which can compute P (O|L), probability of observation sequence.

Forward procedure:

Initialization

$$\alpha_1(i) = \pi_i b_i(o_1), \quad 1 \le i \le N$$

Induction

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}), \quad 1 \le t \le T-1, 1 \le j \le N$$

Termination

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

Backward procedure:

Initialization

$$\beta_T(i) = 1 \quad 1 \le i \le N$$

Induction

$$\beta_T(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad T-1 \le t \le 1, 1 \le i \le N$$

2. The state occupation probability t(sj) is the probability of occupying state sj at time t given the sequence of observations $O_1, O_2 \ldots\ldots O_N$.
3.Baum-welch algorithm for parameter re-estimation.

.

Gaussian distribution:

- The Gaussian (or Normal) distribution is the most common (and easily analysed) continuous distribution
- Gaussian distribution was used to model the state–output distribution.
- If words were represented by a scalar we would model them with a Gaussian distribution.
- speaker and gender differences tend to create multiple modes in the data. To address this problem, using a mixture of Gaussians(GMM)

The Gaussian is described by two parameters:

1)The mean μ (location)
2)The variance σ2 (dispersion)

- maximise the likelihood of the data given these alignments

IV. EXPERIMENTAL RESULTS

In this HMM-based speech to text conversion system, seven audio files such as apple, banana, kiwi, lime, orange, peach, pineapple are modelled in HMM. The original signal at the sampling rate of 8 kHz is taken. The output is shown here:

```
Miss 1: Predicted pineapple, but was lime.
Miss 2: Predicted kiwi, but was peach.
Miss 3: Predicted banana, but was pineapple.

mcr =

    0.0286
```

Here, mcr is known as misses' classification rate and the respective misses are shown below during the training process.

Miss 1:

```
Recognized word as lime!
Recognized word as lime!
Recognized word as pineapple!
Recognized word as lime!
Recognized word as lime!
Recognized word as lime!
Recognized word as lime!
```

Miss 2:

```
Recognized word as peach!
Recognized word as peach!
Recognized word as peach!
Recognized word as peach!
Recognized word as peach!
Recognized word as kiwi!
Recognized word as peach!
```

Miss 3:

```
Recognized word as pineapple!
Recognized word as pineapple!
Recognized word as pineapple!
Recognized word as banana!
Recognized word as pineapple!
Recognized word as pineapple!
Recognized word as pineapple!
```

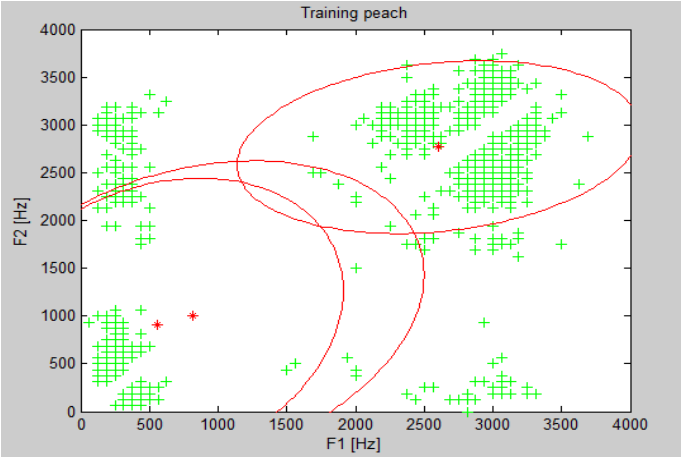

Fig 4.1. Training plot for peach

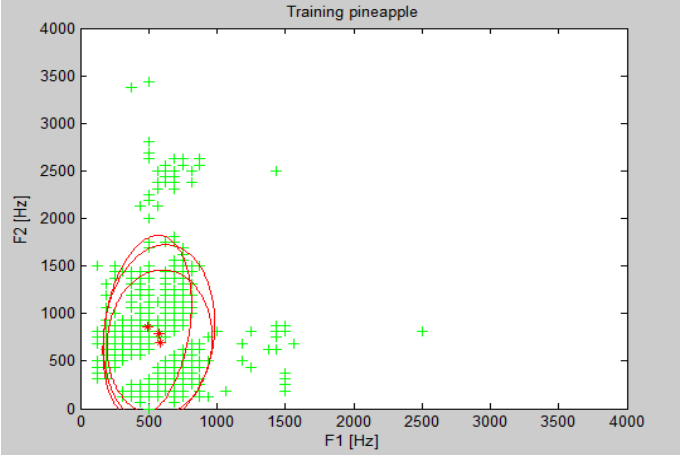

Fig 4.2. Training plot pineapple

Train data	Number of tests	Number of correct tests	Misses	Percentage of accuracy
Apple	15	15	0	100
Banana	15	15	0	100
Kiwi	15	15	0	100
Lime	15	14	1	93.33
Orange	15	15	0	100
Peach	15	14	1	93.33
Pineapple	15	14	1	93.33

Table 4. 1: Percentage accuracy for three states of HMM

V. CONCLUSION

This Speech- to-Text conversion system is implemented by using the MFCC for feature extraction and HMM as the recognizers. In audio folder, 105 audio files are recorded and these are analyzed to get feature vectors. These features are initially modeling in the HMM. After that, the test spoken word is addressed by baum-welch algorithm of HMM. From the simulation results, it can be clearly seen that the average recognition rate of 97.14% is achieved by the number of states (N=3).

REFERENCES

[1] Santosh K.Gaikwad, 'A review on speech recognition techniques', International Journal of Computer Applications, Volume 10– No.3, November 2010

[2] NishantAllawadi,'Speech-to-Text System for Phonebook Automation', Computer Science And Engineering Department Thapar University,June 2012.

[3] Sanjivani S.Bhabad, 'An overview of technical progress in speech recognition', International Journal of advanced research in computer science and software Engineering, Volume 3, Issuse 3, March 2013

[4] Akshay S. Utane, 'Emotion Recognition Through Speech Using Gaussian Mixture Model And Hidden Markov Model', International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4, April 2013.

[5] Su Myat Mon, Hla Myo Tun, 'Speech-To-Text(STT) System Using Hidden Markov Model(HMM)', International Journal Of Scientific & Technology Research Volume 4,Issue 06, June 2015.