Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM)

Su Myat Mon, Hla Myo Tun

Abstract: Speech is an easiest way to communicate with each other. Speech processing is widely used in many applications like security devices, household appliances, cellular phones, ATM machines and computers. The human computer interface has been developed to communicate or interact conveniently for one who is suffering from some kind of disabilities. Speech-to-Text Conversion (STT) systems have a lot of benefits for the deaf or dumb people and find their applications in our daily lives. In the same way, the aim of the system is to convert the input speech signals into the text output for the deaf or dumb students in the educational fields. This paper presents an approach to extract features by using Mel Frequency Cepstral Coefficients (MFCC) from the speech signals of isolated spoken words. And, Hidden Markov Model (HMM) method is applied to train and test the audio files to get the recognized spoken word. The speech database is created by using MATLAB.Then, the original speech signals are preprocessed and these speech samples are extracted to the feature vectors which are used as the observation sequences of the Hidden Markov Model (HMM) recognizer. The feature vectors are analyzed in the HMM depending on the number of states.

Keywords: Speech Recognition, End Point Detection, MFCC, HMM, MATLAB

1. Introduction

Human interact with each other in several ways such as facial expression, eye contact, gesture, mainly speech. The speech is primary mode of communication among human being and also the most natural and efficient form of exchanging information among human in speech [1]. Speech-to-text conversion (STT) system is widely used in many application areas. In the educational field, STT or speech recognition system is the most effective on deaf or dumb students. The recognition of speech is one the most challenges in speech processing. Speech Recognition can be defined as the process of converting speech signal to a sequence of words by means of Algorithm implemented as a computer program [1]. Basically, speech to text conversion (STT) system is distinguished into two types, such as speaker dependent and speaker independent systems [2]. This paper presents the speaker dependent speech recognition system. Speech recognition is very complexity case when processing on randomly varying analogue signal such as speech signals. Thus, in speech recognition system, feature extraction is the main part of the system. There are various methods of feature extractions. In recent researches, many feature extraction techniques are commonly used such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), Linear Predictive Coding (LPC), Cepstral Analysis and Mel-frequency cepstral (MFCCs), Kernal based feature extraction based approach, Wavelet Transform and spectral subtraction [3]. In this paper, Mel Frequency Cepstral Coefficients (MFCC) method is used. It is based on the characteristics of the human ear's hearing, which uses a nonlinear frequency unit to simulate the human auditory system. Mel frequency scale is widely used to extract features of the speech. Mel-frequency cepstral features provide the rate of recognition to be efficient for speech recognition as well as emotion recognition system through speech [4]. Moreover, Vector Quantization (VQ), Artificial Neural Network (ANN), Hidden Markov Model (HMM), Dynamic Time Warping (DTW) and various techniques are used by the researchers in recognition. Among them, HMM recognizer is currently dominant in many applications. Nowadays, STT system is fluently used in many control systems, mobile phones, computers and so forth. Therefore, speech recognition system is more and more popular and useful in our daily lives. In the system, MFCC and HMM are implemented by MATLAB.

2. Methodology

A. End Point Detection

Classification of speech into voiced or unvoiced sounds provides a useful basis for subsequent processing. A three-way classification into silence/unvoiced/voiced extends the possible range of further processing to tasks such as stop consonant identification and endpoint detection for isolated utterances [5]. In noisy environment, speech samples containing unwanted signals and background noise are removed by end point detection method. End point detection method is based on the short-term log energy and short-term zero crossing rate [6]. The logarithmic short-term energy and zero crossing rates are calculated in the following equation [1] and [2]

$$E_{log}=\sum_{n=1}^{N}\log(s(n)^2\,) \qquad [1]$$

$$ZCR=\frac{1}{2}\sum_{n=1}^{N}|sgn[s(n+1)]-sgn[s(n)]| \qquad [2]$$

$$sgn[s(n)]=\begin{cases}+1 & s(n)\geq 0\\ -1 & s(n)<0\end{cases}$$

Wheres(n) is the speech signal, E_{log} is the logarithmic short-term energy and ZCR is the short-term zero crossing rate.

B. Mel Frequency Cepstral Coefficient (MFCC)

Feature extraction is the most important part of the entire system. 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.1 shows the block diagram of MFCC process.

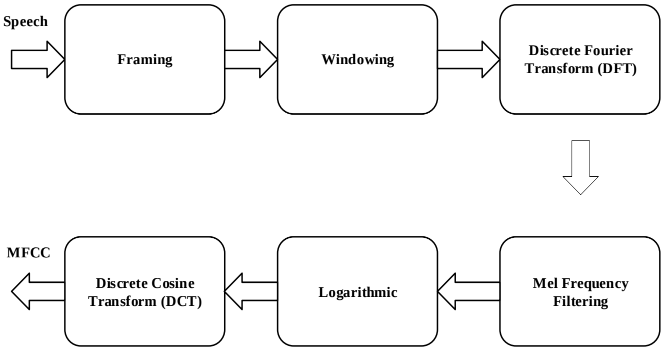

Speech
Framing
Windowing
Discrete Fourier Transform (DFT)
Mel Frequency Filtering
Logarithmic
Discrete Cosine Transform (DCT)
MFCC

Fig .1. *Block diagram of MFCC*

Framing: It is the first step of the MFCC. It is the process of blocking of the speech samples obtained from the analogue to digital conversion (ADC) of the spoken word, into the number of frame signal with 20- 40ms frame time length. Overlapping is needed to avoid loss of information. Windowing: In order to reduce the discontinuities at the start and end of the frame or to be smooth of the first and last points in the frame, windowing function is used. 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 is more precise in frequency domain rather than in time domain. Mel frequency filtering: The voice signal does not follow the linear scale and 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 at high. Logarithmic function: Logarithmic transformation is applied to the absolute magnitude of the coefficients obtained after Mel-scale conversion. The absolute magnitude operation discards 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.

C. Hidden Markov Model Recognizer

In recognition or classification of the speech signal, there are many approaches to recognize the test audio file. The methodologies of speech recognition are: ANN, GMM, DTW, HMM, Fuzzy logic and various other methods. Among them, HMM techniques are widely used in many applications than any other ones. There are four types of HMM model used in speech processing. Details of HMM models are given in [7]. The phonemes in speech follow the left to rightsequences, so the structure of HMM is a left-to-right structure. The states of HMM model represent the word or acoustic phonemes in speech recognition. The number of states of HMM model is randomly chosen to model. The choice of the number of states causes to change the feature vectors or observations. It affects the recognition rate or accuracy of speech recognition.The most flexible and successful approach to speech recognition so far has been Hidden Markov Models (HMMs). HMM is the popular statistical tool for modeling a wide range of time series data. In speech recognition area, HMM has been applied with great success to problem such as part of speech classification [1]. HMM word model λ is composed of initial state probability (π), state transition probability (A) and symbol emission probability (B). In HMM-based speech recognition system, there exist three main problems called evaluation, decoding and learning problems. The training and testing algorithm of HMM are discussed in details [8]. The probability of observations or likelihood given the model determines the expected recognized word. It is calculated by the following equation [3]

$$L=P(O|\lambda)=\sum_{i=1}^{N}\alpha_T(i) \qquad [3]$$

Where P(O|λ) is the probability of observations addressed by forward algorithm. N is the number of states and $\alpha_T(i)$ is the forward variables with the length of observations. The highest probability of observations determines the recognized spoken word.

3. Implementation

The flowchart of speech to text conversion is illustrated in Fig .2. To convert input speech to text output, the four main steps are developed by using MATLAB.These steps are speech database, preprocessing, feature extraction and recognition. Firstly, five audio files are recorded with the help of computer. Each audio file contains ten different pronunciation audio files. So, there are total of fifty audio files are recorded in speech database. The speech signals at low frequencies have more energy than at high frequencies. Therefore, the energies of signal are necessary to be boost at high frequencies. According to the saturation of environment, the unwanted noise may affect the recognition rate worse. This problem can be overcome by end point detection method. After preprocessing stage is finished, the speech samples are extracted to features or coefficients by the use of Mel Frequency Cepstral Coefficient (MFCC). Finally, these MFCC coefficients are used as the input of Hidden Markov Model (HMM) recognizer to classify the desired spoken word. The desired text output can be generated by HMM method even if the test audio file is included in the existing speech database.

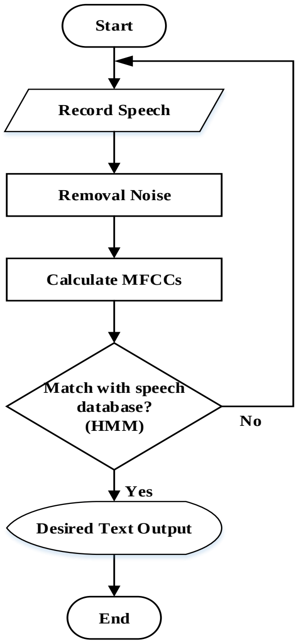
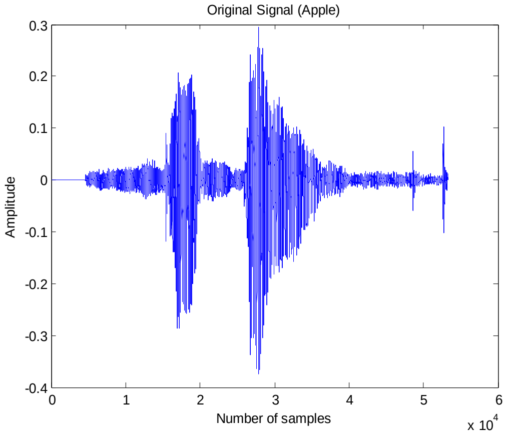
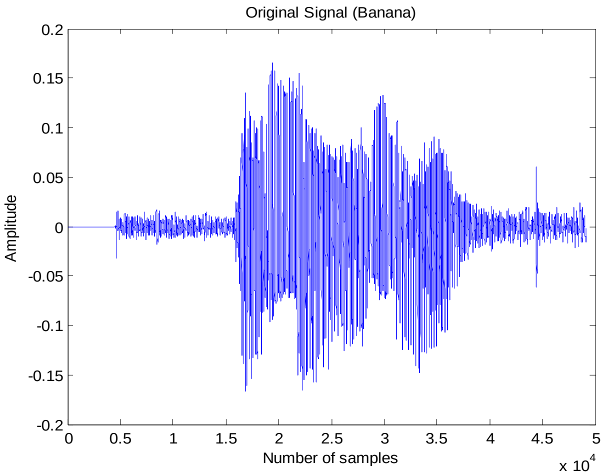


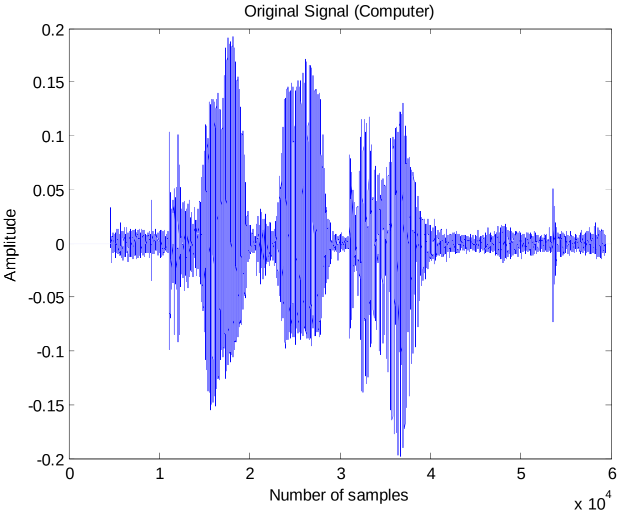
Fig .2. *Flowchart of speech to text conversion*

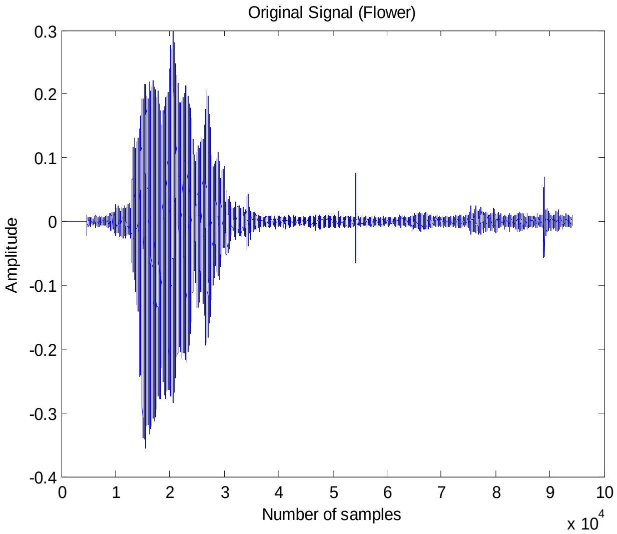
4. Simulation Results

In this HMM-based speech to text conversion system, five audio files such as apple, banana, computer, flower and key are modeled in HMM. The original signal at the sampling rate of 8 kHz are demonstrated in Fig .3.









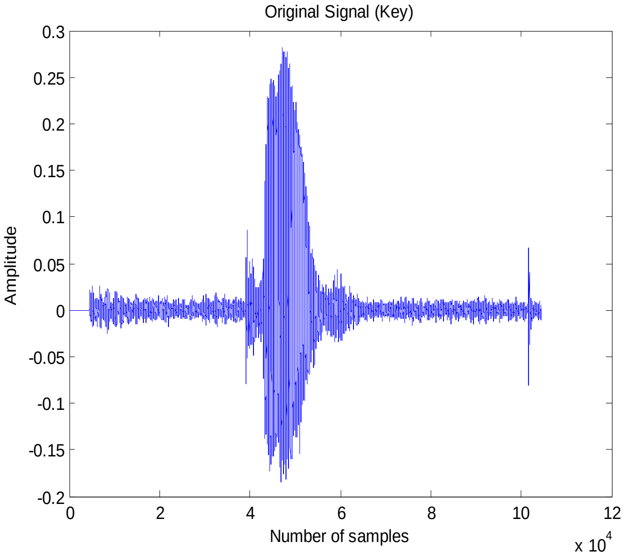


Fig .3 *Amplitude versus number of samples of five original signals*

In recognition, the more the number of states in HMM, the better the recognition rate or accuracy. Tables 1, 2 and 3 show the percentages of recognition rate for speech to text conversion.

Table .1 *Table of Percentage Accuracy for three states of HMM (N=3)*

Train data	Number of test	Number of correct test	Error	Percentage of Accuracy
Apple	50	31	19	62%
Banana	50	32	18	64%
Computer	50	33	17	66%
Flower	50	31	19	62%
Key	50	30	20	60%

Table .2 *Table of Percentage Accuracy for four states of HMM (N=4)*

Train data	Number of test	Number of correct test	Error	Percentage of Accuracy
Apple	50	37	13	74%
Banana	50	36	14	72%
Computer	50	36	14	72%
Flower	50	39	11	78%
Key	50	35	15	70%

Table .3 *Table of Percentage Accuracy for five states of HMM (N=5)*

Train data	Number of test	Number of correct test	Error	Percentage of Accuracy
Apple	50	43	7	86%
Banana	50	42	8	84%
Computer	50	45	5	90%
Flower	50	46	4	92%
Key	50	43	7	86%

In Table .1, the percentage of recognition rate for apple and flower is 62 %. For banana, the recognition rate is slightly increased to 64% and the recognition rate of computer have the best result of 66%.Whereas, the accuracy of key has the least of 60%.For number of states (N=4),the percentage of recognition rate is increased around 70 for all audio files. This is shown in Table .2. According to the Table .3, the number of states (N=5) gives the better accuracy than any other states. The recognition rate of individual spoken word is nearly from 84 to 92%.

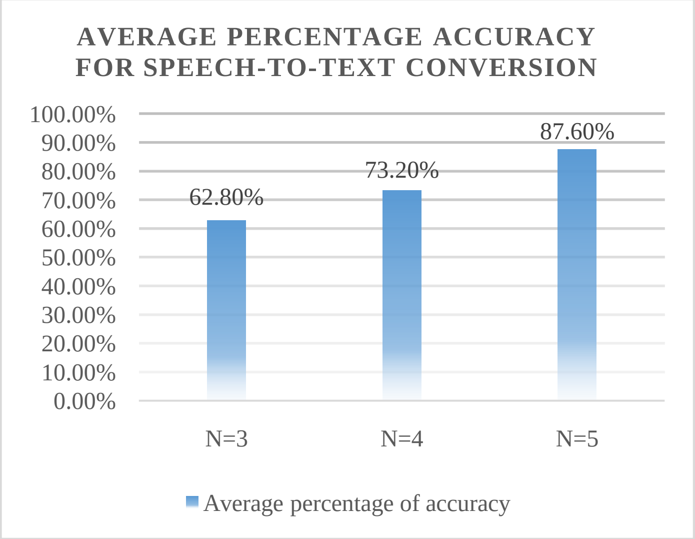


Fig .4 *Average Recognition Rates of Speech-to-Text Conversion System*

The average percentage accuracy or recognition rate for the system is illustrated in Fig .4.At the number of state (N=5), the average accuracy is about 87.6% as the most. It is better recognition rate than state three and four of HMM.

5. Conclusion

This Speech- to-Text conversion system is implemented by using the MFCC for feature extraction and HMM as the recognizers. In speech database, fifty audio files are recoded 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 forward algorithm of HMM. From the simulation results, it can be clearly seen that the average recognition rate of 87.6% achieved by the number of states (N=5) is better accuracy than any other states. But, if the number of states is too large, there are no enough observations per state to train the model. So, this may degrade the performance of the system. Thus, the choice of the number of states in the HMM also plays an important case in recognition. In this work, the performance of the system is more accurate and reliable by using end point detection algorithm in preprocessing stage.

Acknowledgement

The author would like to thank to Dr. Hla Myo Tun, Associate Professor and Head of the Department of Electronic Engineering, Mandalay Technological University for his help and for his guidance, support and encouragement.

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] Nitin N Lokhande, ‘Voice Activity Detection Algorithm for Speech Recognition Applications’, International Conference in Computational Intelligence (ICCIA), 2011

[6] S.A.R. Al-haddad, ‘Automatic Segmentation and Labeling for Continuous Number Recognition’, August 21-23, 2006 (pp221-224)

[7] D.B. Paul, ‘Speech Recognition Using Hidden Markov Models’, The Lincoln Laboratory Journal, Volume 3, Number 1 (l990).

[8] Mathew Magimai Doss,’Using Auxiliary Sources Of Knowledge For Automatic Speech Recognition’, Computer Science and Engineering ,2005