SPEECH RECOGNITION SYSTEM USING MFCC AND DTW

[1]INGYIN KHINE, [2]CHAW SU

[1,2]FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY, UNIVERSITY OF TECHNOLOGY (YATANARPON CYBER CITY),PYINOOLWIN,REPUBLIC OF THE UNION OF MYANMAR
E-mail: [1]ingyinkhine.igk.utycc@gmail.com, [2]drchawsu.it@gmail.com

Abstract - This paper presents a speech recognition system. The main goal of this system is to classify the speech of the speaker. In this system, there are three main parts – (1) pre-processing, (2) feature extractions and (3) classification. The input speech is preprocessed by Voice Active Detection (VAD).Features are extracted by Mel-Frequency Cepstral Coefficient (MFCC). The input speech is classified by Dynamic Time Wrapping (DTW). There are five types of speech objects and 32 speech signals for each object. The system is processed for both user-dependent and user-independent. The overall accuracy of training data testing is 100 % for both. However, the accuracy of the real-time data testing is 44 % and52 % for user-dependent and user-independent, respectively. Therefore, user-independent system is more accuracy and less error rate than user-independent one.

Index terms - DTW, MFCC,Speech recognition, User-Independent, VAD.

I. INTRODUCTION

Today is the "information age", and so, Information Technology (IT) has become a part of everyday lives. Similarly, speech also becomes an essential part for everyday lives because human interact with each other by using speech.That is, speech is the most important and natural way to communicate for human beings and so,speech processing is very crucial to recognize the speech or speaker. On the other hand, people can interact with not only each other but also everything through the speech. For example, people can control switch with the voice.

Speech recognition and speaker recognition are included in speech processing. Speaker recognition is the process of recognizing person from a spoken phrase. Speech recognition is the process of identifying words and phrase from spoken language[4].Speech recognition is very useful in many applications and environments in daily life. For example, it can be used in a car environment to voice control non critical operations, such as dialing a phone number.To make the daily chores easier, voice control could be helpful. With the voice, people could operate the light switch on/off, turn off/on the coffee machine or operate some other domestic appliances [7]. The system of this paper is intended to become a part of domestic appliance. Basic model for speech recognition is shown in Figure 1.

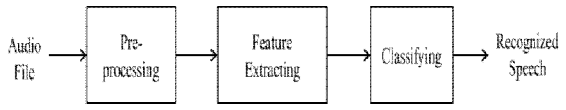


Figure 1: Block Diagram for Speech Recognition

There are basically three main steps for speech recognition: (1) pre-processing, (2) feature extracting and (3) classifying.

1. Pre-processing:It is the main stage for speaker recognition. It includes starting and ending point detection from the speech and the task of removing unvoiced part from the speech. It means that voice and unvoiced part in the speech can be separated in pre-processing stage. For pre-processing various methods are used such as Voice Activity Detection(VAD), Short Time Energy(STE), and Zero Crossing Rate (ZCR)[3].
2. Feature Extracting:Feature extraction is the main role of the speech recognition system. It is considered as the heart of the system. The work of this is to extract the features from the input speech (signal) which is spoken by different speakers[6]. There are many feature extraction techniques like Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), Relative Spectral (RASTA), Discrete Wavelet Transform (DWT), Wavelet Packet Transform (WPT), Probabilistic Linear Discriminate Analysis (PLDA), and Mel-Frequency Cepstral Coefficient (MFCC).
3. Classifying: This is the last stage for recognizing speakers. It is used for classifying different speech [3]. For classification, various techniques are available like Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Artificial Neural Network (ANN),Dynamic Time Warping (DTW), Vector Quantization (VQ) models, Support Vector Machine (SVM), and Nearest Neighbor (NN).

II. THEORETICAL BACKGROUND

Speech recognition is the process of identifying words and phrase from spoken language and converts them into machine readable format [4]. It is also a speech or speaker classification problem, which formally consists of three parts.

A. Pre-processing - VAD

Voice activity detection (VAD) refers to a type of methods which attempt to determine if a signal is speech or non-speech. In a noise-free scenario, the task is trivial, but it is also not a realistic scenario. The block diagram is displayed in Figure 2.

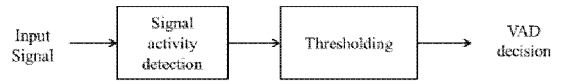


Figure 3: Block Diagram of VAD

The basic idea of algorithms is to

1. Calculate a set of features from the signal which are designed to analyze properties which differentiate speech and non-speech.
2. Merge the information from the features in a classifier, which returns the likelihood that the signal is speech.
3. Threshold the classifier output to determine whether the signal is speech or not.

VADs are used as a low-complexity pre-processing method, to save resources (e.g. complexity or bitrate) in the main task[8].

B. Feature Extraction– MFCC

Mel Frequency Cepstral Coefficients (MFCC) technique is the robust and dynamic technique for speech feature extraction. In Figure 3, block diagram of MFCC model is shown[1][4][5].

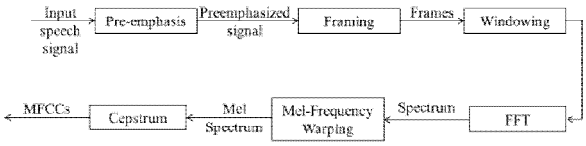


Figure 3: Block Diagram of MFCC

1. Pre-emphasis:This step processes the passing of signal through a high-pass filter which emphasizes higher frequencies.Mathematically, if x(n) is the speech signal passed through a high-pass filter, then the output signal y(n) is:

$$y(n) = x(n) - a * x(n-1) \qquad (1)$$

where, a is the pre-emphasis coefficient and its value is taken as 0.95.

2. Framing: It is the task of partitioning the speech signal into small segments in order toanalyze in the short instead of analyzing the entire signal at once. Beside, overlapping between the frames is done to prevent any information loss and to maintain a correlation between two adjacent frames. Overlapping is useful to produce continuity within frames.

3. Windowing: The next step is to window each individual frame. Each individual frame is multiplied by a window function to eliminate the spectral discontinuities of the signal, by taping the signal to zero at the starting and ending of each frame.if x(n) is a nth frame of a signal and the hamming window is w(n), then the result of windowing is the signal:

$$y(n) = x(n).w(n),\ 0 \le n \le N-1, \qquad (2)$$

where, N is the frame length.

4. Discrete Fourier Transformation:DFT is applied on each windowed frame in order to convert each frame of N samples from the time domain into the frequency domain. Fast Fourier Transform (FFT), which is speeding up the processing, is used because it is the fastest way to calculate the DFT.

5. Mel-frequency warping:The major work in this is to convert the frequency spectrum to Mel spectrum. This makes the spectral frequency characteristics of signal closely corresponding to the human perception. A subjective pitch is measured on a scale called the Mel scale for each tone of signal with an actual frequency.The Mel Frequency Scale is given by:

$$mel(f) = 2595 * \log10(1 + f / 700) \qquad (3)$$

6. Cepstrum: In this step, the Mel-Frequency Cepstrum Coefficients (MFCCs) are resulted by means of Discrete Cosine Transform (DCT). DCT converts the log Mel spectrum from frequency back to time domain. Transforming with DCT is required because FFT has been performed.

C. Classification–DTW

Dynamic time warping algorithm measures the similarity between two sequences which may vary in time or speed. This technique which is based on dynamic programming finds the optimal alignment between two times series if one time series may be warped non-linearly by stretching or shrinking it along its time axis. This warping between two times series can be used to find equivalently regions among the two times series or to find the similarity between two times series. Same user can give different utterances of same word which may differ in time. DTW resolves this problem by aligning the words properly and calculating the minimum distance between two words. It has been applied to temporal sequences of video, audio, and graphics information. Therefore, any data which may be become a linear sequence are often analyzed with DTW [4].

Distance metric is the particular metric to be used to perform the match operation.Two concepts exist under the distance metric:

1. Local distance: the difference or variability computed between two feature vectors of two different speech signals.
2. Global distance: the total difference or variability between two speech signals.

Suppose there are two numerical sequences (a_1, a_2,..., a_n) and (b_1, b_2,..., b_m). The length of the two

sequences can be different. The algorithm starts with local distances calculation between the elements of the two sequences using different types of distances. The most frequent used method for distance calculation is the absolute distance between the values of the two elements (Squared Euclidian distance). Those results stored in a matrix of distances with n lines and m columns of general term.

$$d_{ij} = (a_i - b_j)^2, \quad i = 1 \rightarrow n, j = 1 \rightarrow m \qquad (4)$$

Starting with local distances matrix, then the minimal distance matrix between sequences is determined using a dynamic programming algorithm and the following optimization criterion:

$$Y_{ij} = d_{ij} + \min (Y_{i-1,j-1}, Y_{i-1,j}, Y_{i,j-1}) \qquad (5)$$

where, Y_{ij} is the minimal distance between the subsequences $(a_1, a_2,..., a_i)$ and $(b_1, b_2,..., b_j)$.

A warping path is a path through minimal distance matrix from Y_{11} element to Y_{nm} element consisting of those Y_{ij} elements that have formed the Y_{nm}distance.

The global warp cost of the two sequences is defined as shown below:

$$GC = \frac{1}{p}\sum_{i=1}^{p} w_i \qquad (6)$$

where, w_i are those elements that belong to warping path, and p is the number of them [2].

III. EXPERIMENTAL RESULTS

The system consists of five different utterances. It is processed by both user-dependent and user-independent. There are nine speakers (included males, females, and child) for user-independent. For both, 32 voice signals of each utterance are trained in training dataset and totally 25 speech signals are tested. The recording configuration and training data requirements are described in Table 1 and 2, respectively.

The system is implemented by MATLAB programming language with three portions. Figure 4 shows the classification phase of the system.

Table.1Recording configuration

Types	User-dependent	User-independent
Audio Format Used	wave-format (.wav)	wave-format (.wav)
Sampling Rate	44.1 kHz	44.1 kHz
Number of Channels	1 (Mono-channel)	2 (Stereo-channel)
Recording Software	MatLab Code	Voice Recorder
Recording Hardware	Micro-phone (SM57)	Mobile Phone (Android)

Table.2 Training data requirements

Types	User-dependent	User-independent
No. of Speakers	Nine speakers	One speakers
Speaker	Three males Five females One child	One female
No. of Objects	Five objects	Five objects
Utterance	1) 2) 3) 4) 100 plus 5) orange	1) 2) 3) 4) 100 plus 5) orange
Feature Points	13 (for one file)	13 (for one file)

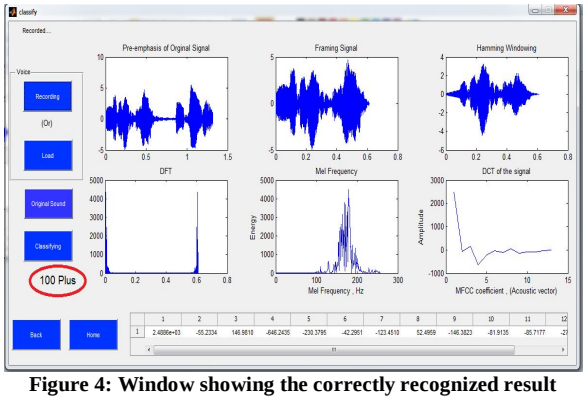


Figure 4: Window showing the correctly recognized result

The accuracy results are calculated by using the following equation:

$$\text{Accuracy} = \frac{\text{Total no. of correct recognition}}{\text{Total no. of testing files}} * 100\ \% \qquad (7)$$

Theaccuracy results of each objectfor both user-dependent and user-independentare shown in the following Table 3 and 4, respectively. The recognized results are measured on 160 training data. Figure 5 presents the bar chart of the accuracy results of testing data.

Table.3Accuracy results of testing data for user-dependent

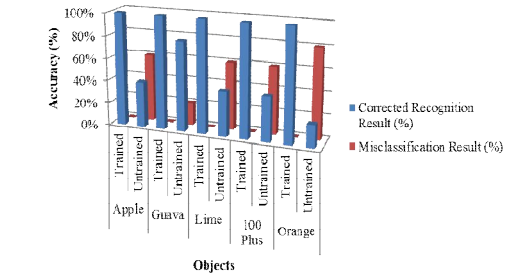
Objects	Testing Data	Corrected Recognition Result (%)	Misclassification Result (%)
Apple	Trained	100 %	0 %
	Untrained	40 %	60 %
Guava	Trained	100 %	0 %

	Untrained	80 %	20 %
Lime	Trained	100 %	0 %
	Untrained	40 %	60 %
100 Plus	Trained	100 %	0 %
	Untrained	40 %	60 %
Orange	Trained	100 %	0 %
	Untrained	20 %	80 %

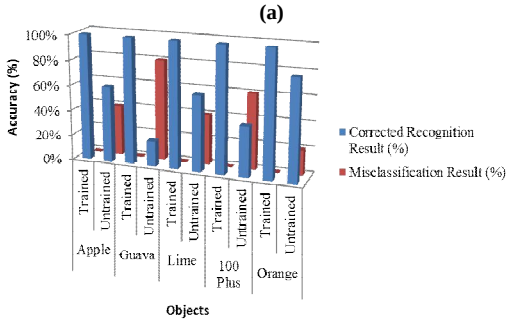
Table.4Accuracy results of testing data for user-independent

Objects	Testing Data	Corrected Recognition Result (%)	Misclassification Result (%)
Apple	Trained	100 %	0 %
	Untrained	60 %	40 %
Guava	Trained	100 %	0 %
	Untrained	20 %	80 %
Lime	Trained	100 %	0 %
	Untrained	60 %	40 %
100 Plus	Trained	100 %	0 %
	Untrained	40 %	60 %
Orange	Trained	100 %	0 %
	Untrained	80 %	20 %

According to the results, the guava object has the highest accuracy, but, the orange object has the highest error rate for user-dependent system. On the other hand,user-independent is opposite to user-dependent. In user-independent system, the orange is most accurate and the guava is least accurate.



(a)



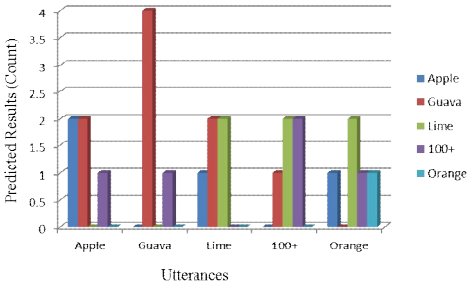
(b)

Figure 5: Bar chart for accuracy results of testing data(a) for user-dependent and (b) for user-independent

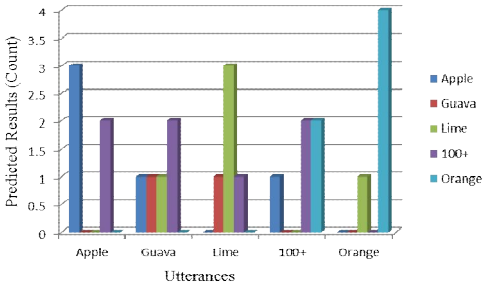
The overall accuracy of the training data for both is 100%. Nevertheless, it gives only 44% and 52% accuracy for real-time data of user-dependent and that of user-independent, respectively, corresponding to the results of Table 5. Figure 6 describes the bar chart of analysis of real-time data testing.

Table.5Confusion matrix for real-time data testing

Objects	User Types	Apple	Guava	Lime	100 Plus	Orange		Times
Apple	User-dependent	2	2	0	1	0		5
	User-independent	3	0	0	2	0		5
Guava	User-dependent	0	4	0	1	0		5
	User-independent	1	1	1	2	0		5
Lime	User-dependent	1	2	2	0	0		5
	User-independent	0	1	3	1	0		5
100 Plus	User-dependent	0	1	2	2	0		5
	User-independent	1	0	0	2	2		5
Orange	User-dependent	1	0	2	1	1		5
	User-independent	0	0	1	0	4		5



(a)



(b)

Figure 6: Bar chart for analysis of real-time data testing (a) for user-dependent and (b) for user-independent

CONCLUSION

Feature extraction and training processes are performed by using MFCC. The system is performed by user-dependent and user-independent. For user-independent, it consists of nine speakers. There are 160 input audio signals for five utterances and feature dataset is created.DTW is used to classify the feature points according to the feature dataset.This system is intended to apply in Domestics Pick and Place Robotic Arm Control System.It has two phases: training and testing.The time taken to process the training of feature points is 50s.The system got 100% accuracy for the training data.However, it has only 44 % accuracy in real-time data for the user-dependent data testing. As well as, the accuracy of real-time data testing in user-independent is only 52 %. Although the accuracy for guava object is highest in user-dependent, it has the least accuracy in user-independent. According to the experiment, the system for user-independent is more accurate than user-dependent. It has a challenge that the input speech signal length is quite different. Some speeches are long so that it has to speak very quickly. As a future work, the voices can be recorded with a specific environment using the suitable recording software. Moreover, the speech signals can be segmented by a phrase. It can be expected to get a better accuracy.

REFERENCES

[1] Dey .S and Kashyap .K: "A Dynamic-threshold Approach to Text-dependent Speaker Recognition using Principles of Immune System", IEEE, 2015.

[2] Dr. Sadiq J. Abou-Loukh and SamahMutasherGatea: "Spoken Word Recognition Using Slantlet Transform and Dynamic Time Warping", NUCEJ, 2011.

[3] Khushboo S. Desai and Pujara .H: "Speaker Recognition from the Mimicked Speech: A Review", IEEE, 2016.

[4] Kishori R. Ghule and Ratnadeep R. Deshmukh: "Automatic Speech Recognition System Using MFCC and DTW for Marathi Isolated Words", IJTEEE, 2015.

[5] Muda .L et.al: "Voice Recognition Algorithms using Mel-Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, 2010.

[6] Narang .S and Ms. Divya Gupta, "Speech Feature Extraction Techniques: A Review", IJCSMC, Vol. 4, Issue. 3, 2015, pg.107 – 114.

[7] Nilsson .M and Ejnarsson .M, "Speech Recognition using Hidden Markov Model performance evaluation in noisy environment", Master of Science in Electrical Engineering, 2002, pg.5 - 41.

[8] Tom Bäckström, "Voice Activity DetectionSpeech Processing", Aalto University, 2015.

★★★