Speech Command Recognition by using Dynamic Time Warping

Ei Mon Kyaw,MyatThandaKhin
University of Computer Studies, Yangon
eimonkyaw07@gmail.com

Abstract

Speech Recognition is a technology that can be useful in many applicationof our daily life. This proposed method present the speaker dependent speech recognition system. In this paperinclude the three steps of speech command recognition systems. Firstly, automatically crop the input command signal for removing silent part of the command .The second one is extract the feature vectors by using Mel frequency cepstralcoefficients (MFCCs) and then feature matching the command and predefined speech command by using Dynamic Time Warping (DTW). Before matching with DTW, the system sorts the predefined command for matching with command to improve performance of recognition. This system uses the ten predefined command for one username. This paper also compares the speech recognition system between the using of Multi-band Spectral Subtraction and Feature extractions in preprocessing method. Finally, the proposed system designs the "Information Retrieval" application using MATLAB.

Keywords*:Multi-band Spectral Subtraction, Feature Extraction, Feature Matching, Mel Frequency Cepstral Coefficient (MFCC), Dynamic Time Warping(DTW).*

1. Introduction

In computer science and electrical engineering, speech recognition (SR) is the translation of spoken words into text. It is also known as "automatic speech recognition" (ASR), "computer speech recognition", or just "speech to text" (STT). Some SR systems use "speaker-independent speech recognition" while others use "training" where an individual speaker reads sections of text into the SR system.[3] These systems analyze the person's specific voice and use it to fine-tune the recognition of that person's speech, resulting in more accurate transcription. Systems that do not use training are called "speaker-independent" systems and systems that use training are called "speaker-dependent" systems.[3]

Speech recognition applications include voice user interface such as voice dialing (e.g."Call home"), call routing (e.g. "I would like to make a collect call"), demotic appliance control, search (e.g. find a podcast where particular words were spoken), simple data entry (e.g., entering a credit card number), preparation of structured documents (e.g. a radiology report), speech-to-text processing (e.g., word processors or email), and aircraft (usually termed Direct Voice).[3]

The termvoice recognitionor speaker identification refers to identifying the speaker, rather than what they are saying. Recognizing the speaker can simplify the task of translating speech in systems that have been trained on a specific person's voice or it can be used to authenticate or verify the identity of a speaker as part of a security process.

The second section of this paper includes the pre-processing method and the third session presents feature matching technique. Fourth section represents experiment and results and last section includes the conclusion.

2.System overview of propose system

This System can allow the user to input to an application by just talking, while reducing the burden of clicking or pressing buttons on devices, and typing on keyboards. Firstly, automatic cropping the input signal for removing of silent feature in speech signal. After the cropping of input signal, system extract feature vector of the speech signal with Mel Frequency Cepstral Coefficient (MFCC). This method based on the important ideas of cepstrum and its coefficients that collectively represent the short-termpower spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Finally, recognize the signal with predefined command by using Dynamic Time Warping (DTW) Method. DTW is the matchingtechnique for recognition system that warp speech command signal vector to match with predefined command signal vector. The overview of this system is shown in Figure 1.

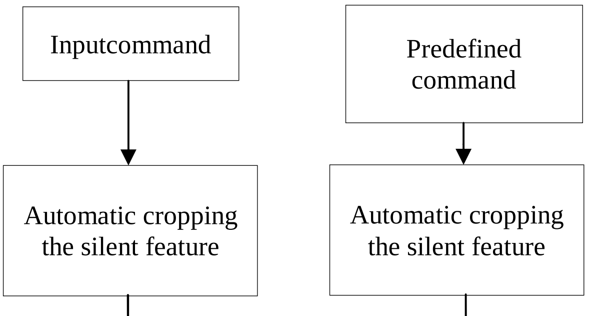

Inputcommand
Predefined command
Automatic cropping the silent feature
Automatic cropping the silent feature

Figure1. System Overview of the proposed system

3. Acoustic Pre-processing

Pre-processing is the very important task for speech recognition to produce better performance. Pre-processing system include Auto cropping method and Feature Extraction.

3.1 Removing silent part of speech signal

Firstly, system removes the silent part of from the speech command signal by performingthe automatic cropping to improve performance. This method remove the silent frame depend on the threshold. Silent have been removed from the signal with zero crossing rate andenergy vector. The “Figure 2”show the removal of silent feature from the original speech command signal.

3.2Feature Extraction (MFCC)

The extraction of the feature vectors of acoustic signals is an important task to produce a better recognition performance. MFCC isbased on the important ideas of cepstrum and take human perception sensitivity with respect to frequencies into consideration, and therefore are best for speech/speaker recognition.

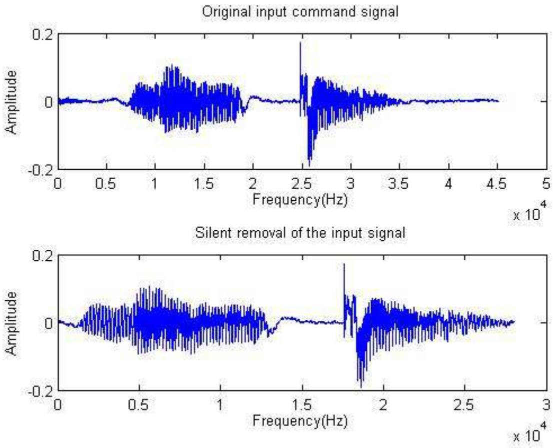


Figure 2. Automatic cropping the silent of the speech command

3.2.1 Pre-emphasis

The first stage of feature extraction is to boost the amount of energy in the high frequency. The speech signal s(n) is sent to high-pass filter:

$$s_2(n) = s(n) - a * s(n-1) \text{Eq(1)}$$

In “Eq(1)”, where $s_2(n)$ is the output signal and the value of a is usually between 0.9 and 1.0. The z-transform of the filter is shown in “Eq(2)”.

$$H(z) = 1 - a * z^{-1} \qquad \text{Eq(2)}$$

The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans. Moreover, it can also amplify the importance of high-frequency formats. The next example demonstrates the effect of pre-emphasis.

3.2.2Frame Blocking

The input speech command signal is segmented into frames of 20 to 30ms with optional overlap of 1/3 to 1/2 of the frame size. Usually the frame size(in term of sample points) is equal to power of two in order to facilitate the use of FFT.

3.2.3 Windowing

The simplest window is the rectangular window that can cause problems because it abruptly cut off of the signal at its boundaries. A common window used in MFCC extraction is the Hamming window, which shrinks the values of the signal toward zero at the window boundaries, avoiding discontinuities. The Hamming window equation is shown in “Eq. (3)”.

$$W(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L}\right) & 0 \le n \le L-1 \\ 0 & otherwise \end{cases}$$ Eq(3)

To extract the signal, this section multiply the value of signal at a time n, X(n) by the value of the window at a time n, W(n).It's shown in "Eq. (4)".

$$Y[n] = X(n) * W(n)$$ Eq(4)

where$Y[n]$ is the output signal, $X(n)$ is the input signal and $W(n)$ is Hamming window.

3.2.4 Fast Fourier Transform or FFT

Spectral analysis show that different timbres in speech signal corresponds to different energy distribution over frequencies. FFT perform to obtain the magnitude frequency response of each frame.

When FFT is using on a frame, the signal within a frame is periodic, and continuous when wrapping around. The Fourier Transform is used to convert the convolution of the glottal pulse U[n] and the impulse response H[n]in the time domain. This statement supports as shown in "Eq. (5)" below:

$$X(k) = \sum_{n=1}^{N} x(n)\omega_N^{(n-1)(k-1)}$$ Eq(5)

Where $\omega_N = e^{(-2\pi i)/N}$ is the Nth root of unity.

3.2.5 Mel Filter Bank

The results of the FFT will be information about the amount of energy at each frequency band. Human hearing is not equally sensitive at all frequency bands. It is less sensitive at higher frequencies, roughly above 1000 hertz. It turns out that modeling this property of human hearing during feature extraction improve speech recognition performance. The positions of these filters are equally spaced along the Mel frequency, which is related to the common liner frequency f by the following "Eq(6)"

$$mel(f) = 1125 * \ln\left(1 + \frac{f}{700}\right)$$ Eq(6)

3.2.6 Discrete Cosine Transform or DCT

In this stepapply DCT on the20 log energy E$_k$obtained from the mel filter bank to have mel-scale cepstral coefficient. The formula of DCT is shown in "Eq(7)"

$$C_m = \sum_{k=1}^{N} \cos\left(\frac{\pi(k-0.5)\pi}{2N}\right)E_k \quad m = 1,2,\ldots L$$ Eq(7)

Where N is the number of mel filters, L is the number of mel-scale cepstralcoefficients.The obtained features are similar to spectrum, thus it is referred to as the mel-scale cepstral coefficients, or MFCC.MFCC alone can be used as the feature for speech recognition. For better performance, this paper can add the log energy and perform delta operation.

3.2.7 Deltas and Energy

The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is aneed to add features related to the change in cepstral features over time. 13 delta or velocity features (12 cepstralfeatures plus energy), and 39 features a double delta or acceleration feature are added. The energy in a frame fora signal x in a window from time sample t1 to time sample t2, is represented as shown below in "Eq. (8)".

$$Energy = \sum X^2[t]$$ Eq(8)

Where X[t] = signal. Each of the 13 delta features represents the change between frames corresponding to cepstral or energy feature,while each of the 39 double delta features represents the change between frames in the corresponding deltafeatures.

3.3 Multi-band Spectral Subtraction

Spectral subtraction is a method for restoration of the power or the magnitude spectrum of a signal observed in additive noise, through subtraction of an estimate of the average noise spectrum from the noisy signal spectrum. Spectral subtraction algorithm is used for removing only for the white noise and multi band spectral subtraction is used for removal of both white noise and as well as colored noise.[9] Speech signal enhance by using Multi-band Spectral Subtraction. If $y(n)$, the noisy speech, is composed of the clean speech signal $s(n)$ and the uncorrelated additive noise signal $d\ (n)$, then:

$$\mathrm{y(n) = s(n) + d(n)}$$ Fig(9)

The power spectrum of the corrupted speech can be approximately estimated as:

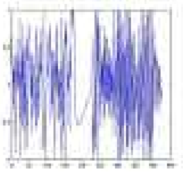
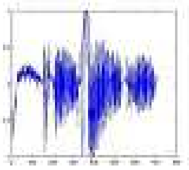
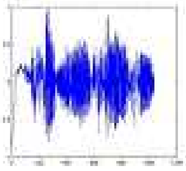
$$|\mathrm{Y}(\omega)|^2 \approx |\mathrm{S}(\omega)|^2 + |\mathrm{D}(\omega)|^2$$ Fig(10)

The multi-band spectral subtraction method is based on the assumption that the additive noise will be stationary and uncorrelated with the clean speech signal.This method split the noise and speech spectra into different frequency band.This proposed method use the initial .25s, window length 25ms and shift percentage 40% for good performance.

3.4 Sorting the Predefined Command

This method calculate the length of the predefined command signal after the extracting feature vectors and sorting depends on their length to recognize more efficiently and higher accuracy.

Table 1. Sorting the command depend on their length

	Length	signal
May Thu	424	
KhinSabal Han	712	
Su Myat Sandi	1040	

When the user command the speech and then system cropping the input speech and extract feature and then find the length of command. When system obtains the length of command and then system cuts the number of predefined command according to the length.

4. Feature Matching Method

Dynamic time warping (DTW) is an algorithmfor measuring similarity between two temporal sequences which may vary in time or speed.[1] For instance, similarities in walking patterns could be detected using DTW, even if one person was walking faster than the other, or if there were accelerations and decelerations during the course of an observation.

This technique also used to find 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.[1] This warping between two time series can then be used to find corresponding regions between the two time series or to determine the similarity between the two time series. To align two sequences using DTW, an n-by-m matrix where the (ith, jth) element of the matrix contains the distance D (x_i, y_j) between the two points xi and y_j is constructed. Then, the absolute distance between the values of two sequences is calculated using the Euclidean distance computation as shown in Eq.

$$\mathrm{D}(\mathrm{xi},\mathrm{yj}) = |\mathrm{xi} - \mathrm{yj}|\mathrm{Eq}(8)$$

Each matrix element (i, j) corresponds to the alignment between the points x_i and $y_{j.}$ Then, accumulated distance is measured by Eq.

$$\mathrm{D}(\mathrm{i},\mathrm{j}) = \min[\mathrm{D}(\mathrm{i}-1,\mathrm{j}-1),\mathrm{D}(\mathrm{i}-1,\mathrm{j}),\mathrm{D}(\mathrm{i},\mathrm{j}-1)] + \mathrm{D}(\mathrm{i},\mathrm{j})\mathrm{Eq}(9)$$

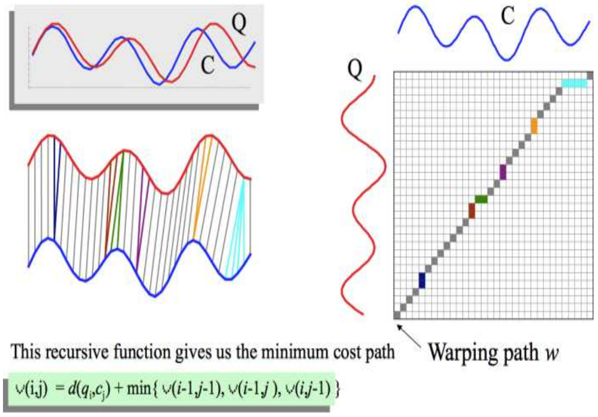
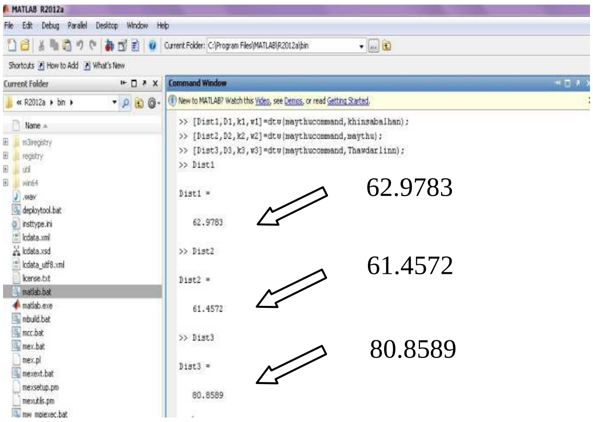

Q
C
C
Q
This recursive function gives us the minimum cost path
Warping path w
γ(i,j) = d(q_i,c_j) + min{ γ(i-1,j-1), γ(i-1,j), γ(i,j-1) }

Figure3. DTW Cost Computation between Two Time Series Q and C

DTW is the matching technique for recognition speech spectrum and its measuring similarity between two speech signals.This method is used to recognize speech command signal by comparing their feature vector with predefined command feature vector that are classify with command length. Before this method, this systemneed to specify the predefined command depending on the user command length. DTW warp speech command signal vector to match with predefined command signal vector and compute optional distance between two signals.

DTW find the minimum distance of the command signal and predefined signal and the figure present the minimum path between "maythucommand"and "khinsabalhan", "maythucommand" and "maythu" and "maythucommand" and "Thawdarlinn". DTW shows the pairs of minimum distance in the figure andchoose the minimum one for match.



```
>> [Dist1,D1,k1,w1]=dtw(maythucommand,khinsabalhan);
>> [Dist2,D2,k2,w2]=dtw(maythucommand,maythu);
>> [Dist3,D3,k3,w3]=dtw(maythucommand,Thawdarlinn);
>> Dist1

Dist1 =

   62.9783

>> Dist2

Dist2 =

   61.4572

>> Dist3

Dist3 =

   80.8589
```

Figure4. Matching Command Using Dynamic Time Warping

5. Experiment & Result

The proposed method and speech recognition system present several experiment. This system has speaker dependency.This research uses the 10 predefined commands for one user name. This system records thespeech commandusing audio recorder in matlaband then processing the following steps shown in system overview.

The Figure5 shows the silent removal of this recognition system.After that this paper extract the feature vectors using Mel Frequency Cepstral Coefficient (MFCC). Figure 6, Figure 7, Figure 8 and Figure 9 show the feature extraction by using MFCCs.

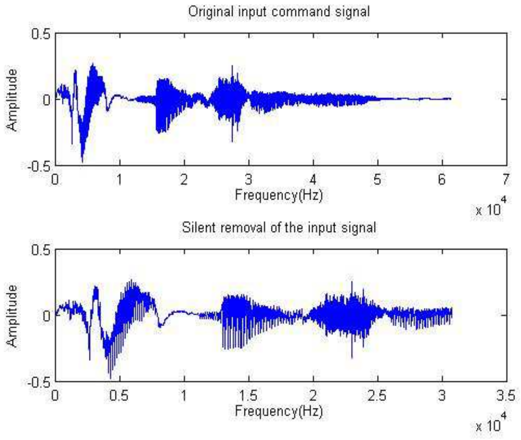


Figure 5.Auto Cropping the Input Command Signal

Finally, this system find the minimum distance between command's mel frequency coefficient andpredefined'smel frequency coefficient by using Dynamic Time Warping (DTW).

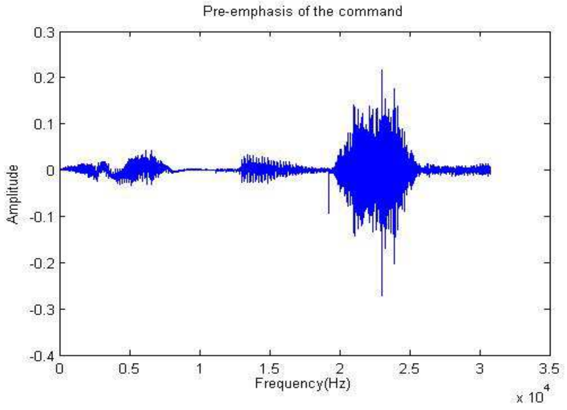


Figure6.Pre-emphasis the Command

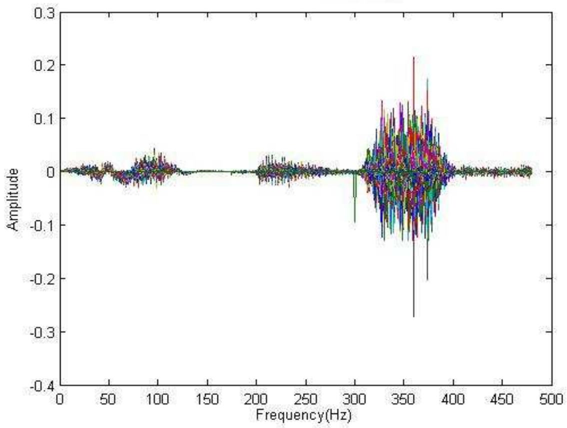


Figure 7. Hamming Window of the Signal

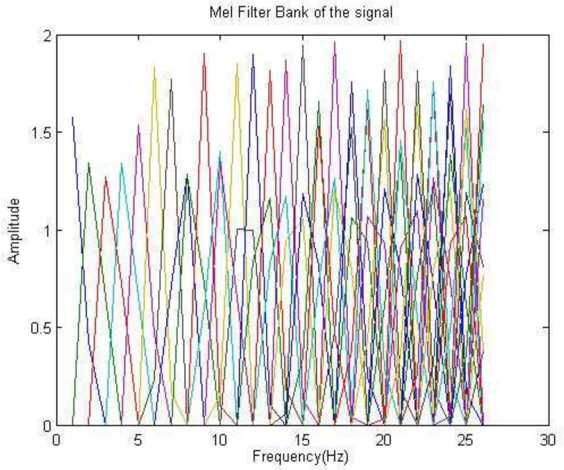


Figure 8. Mel Filter Bank

In this paper use to compares the two main methods before matching the command that aremulti-band spectral subtraction and feature extraction. This system uses the same pre-processing methodlike auto cropping and same amount of predefined command for matching and also same matching method DTW.

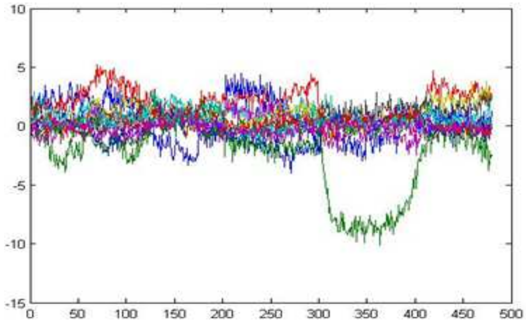


Figure 8. Mel Frequency Coefficient of the Command

The first method use the auto-cropping the command and then multi-band spectral subtraction and after that matching the command with DTW.The second one is automatically crop the commandand then extracting the feature vectors withMFCC. After extracting the feature, this system is matching the MFCC coefficient using DTW.

This system is testing for 100 person's name between feature extraction and MBSS with predefined command. This paper presents the accuracy and performance of the testing data is shown in Table 2.

Table 2. Comparison of Accuracy of Testing for 1000 Predefined Command

	Time	Accuracy
Speech recognition with MBSS	1812.5 s	50%
Speech recognition with Feature extraction	1086.2 s	90%

Table 2 show the comparison of two pre-processing method and their accuracy and time. The time complexity calculation is also included the pre-processing time. The feature extraction is more efficiency than the Muti-band Spectral Subtraction.

Finally, this research built the simple "Information Retrieval System" using MATLAB. This system recognizes the speech command of user name and displays their image. The final result GUI is shown in Figure 9.

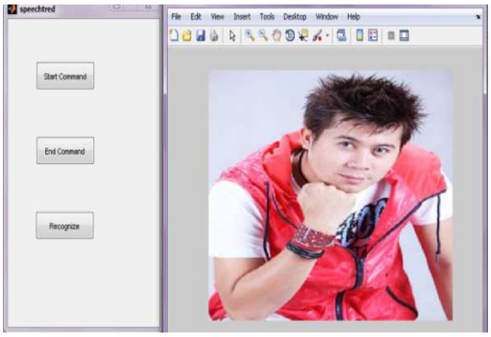


Figure9. The Result of the Recognition of Speech Command

6. Conclusion

This paper uses the automatic crop system for removingsilent part of speech signal to become higher performance and efficiency. This paper compares and contract with multi-band spectral and feature extraction with same pre-processing and same amount of testing predefined data before matching. This thesis present feature extraction isbest method for speech recognition than MBSS.Finally, this paper usesDynamic time warping that is described how to compare or match the command with already predefined command.

References

[1] DANIEL JURAFSKY & JAMES H.MARTIN, "SPEECH AND LANGUAGE PROCESSING".

[2] http://en.wikipedia.org/wiki/Dynamic_time_warping

[3] http://en.wikipedia.org/wiki/Speech_recognition

[4] http://mirlab.org/jang/books/audiosignalprocessing/speechFeatureMfcc.asp?title=12-2%20MFCC

[5] Palden Lama and MounikaNamburu, "Speech Recognition with Dynamic Time Warping using MATLAB",2005.

[6] Rajesh Makhijani, Ravindra Gupta, "Isolated Word Speech Recognition System Using Dynamic Time Warping", International Journal of Engineering Sciences & Emerging Technologies, Dec. 2013

[7] Sheng Li , MingXi Wan and SuPin Wang, "Multi-Band Spectral Subtraction Method for ElectrolarynxSpeech Enhancement", www.mdpi.com/journal/algorithms,2009

[8] Suma Swam and K.V Ramakrishnan. "An Efficient Speech Recognition System", Computer Science & Engineering: An International Journal (CSEIJ), Vol. 3, No. 4, August 2013

[9] SUNIL DEVDAS KAMATH, B.E, "A Multi-Band Spectral Subtraction Method for Speech Enhancement", The University Of Texas At Dallas, December 2001.