UNIVERSITI TEKNIKAL MALAYSIA MELAKA

FAKULTI KEJURUTERAAN ELEKTRIK
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LAPORAN PROJEK SARJANA MUDA

VOICE RECOGNITION ROBOTIC CAR

Clement Chia Kuan You

Bachelor of Mechatronics Engineering

May 2013

“I hereby declare that I have read through this report entitle “VOICE RECOGNITION ROBOTIC CAR” and found that it has comply the partial fulfilment for awarding the degree of Bachelor of Mechatronic Engineering”.

Signature : ....

Supervisor’s Name : ....

Date : ....

VOICE RECOGNITION ROBOTIC CAR

CLEMENT CHIA KUAN YOU

A report submitted in partial fulfilment of the requirements for the degree of Bachelor of Mechatronic Engineering

Faculty of Electrical Engineering
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

YEAR 2013

I declare that this report entitle “VOICE RECOGNITION ROBOTIC CAR” is the result of my own research except as cited in the references. The report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : ....

Name : ....

Date : ....

Specially dedicated to my beloved father, mother and sister

ACKNOWLEDGEMENT

In completing this project, I have received some assistance from my supervisor, other lecturers, family and also my friends. First of all, I would like to show my upmost gratitude to my supervisor Mr. Hairol Nizam bin Mohd Shah for giving the guidance and support by sharing his expertise and knowledge with me. He provided advice for me to explore some new idea in my project and solution for difficult problem. I am very thankful for his advices and guidance until the successfulness of the project.

In addition, I would like to give my recognition and thanks to my previous panels, Dr. Muhammad Fahmi bin Miskon and Dr. Mariam binti Md Ghazaly for their patience in correcting my proposal so that I was on the right track in completing this project. I was able to understand more about the method to apply fundamental knowledge into this project with their advices.

Moreover, I owe a debt of thanks to all those time, concern and supports were given by my parents and friends during the process of completing this report. I am thankful to everyone who always inspires me directly and indirectly during the milestone of completing my final year project.

Last but not least, I would like to give my biggest and sincerest thanks to God for providing me strength, wisdom and intelligent in completing this project so that I can produce the best outcomes for my final year project.

ABSTRACT

The idea for this project is to develop a voice recognition system that recognized five commands to control a robotic car. The focus area for this project is mainly on software parts which is the voice identification and recognition system. The aim of the system was not recognizing a lot of words but only isolated word and to demonstrate the system on a simple built robotic car. The system allows user to input voice commands through a microphone to control the movement of the car. Voice command is sent to computer and computer will process and compare the signal with signal stored in database using Vector Quantization (VQ) technique. Mel-wrapping filter bank in feature extraction was used to reduce the root mean square amplitude noise amplitude and improve signal to noise ratio. Experiment was conducted to compare the distance used to recognize voice command and signal to noise ratio (SNR) of the voice recognition system. Result showed that the SNR was improved and analysis was done on the experiment conducted. After that the output of the recognition system was transferred to Arduino UNO for demonstration of executing the command. The robotic car can be controlled by 5 basic voice command which is stop, forward, reverse, turn left and turn right by integrating source code in MATLAB with Arduino UNO microcontroller.

ABSTRAK

Idea untuk projek ini adalah untuk membina suatu sistem pengecaman suara yang dapat mengecam 5 arahan untuk mengawal kereta robot. Sistem ini membolehkan pengguna untuk input arahan-arahan suara melalui mikrofon untuk mengawal pergerakan kereta. Projek ini fokus terutamanya pada bahagian perisian iaitu pengenalan dan pengecaman suara. Sistem ini tidak bertujuan untuk mengecam banyak perkataan tetapi hanya perkataan tunggal sahaja dan untuk demonstrasi pada satu kereta robot yang mudah dibina. Arahan suara akan dihantar ke komputer dan komputer akan memproses dan membandingkan isyarat dengan isyarat yang disimpan dalam pangkalan data menggunakan teknik Pengkuantuman Vektor (VQ). Penapisan Mel-Pengumpulan yang merupakan salah satu langkah dalam pengekstrakan ciri isyarat telah digunakan untuk mengurangkan amplitude punca min persegi amplitud bunyi dan meningkatkan nisbah isyarat kepada hingar. Eksperimen telah dijalankan untuk membandingkan jarak yang digunakan dalam pengecaman arahan suara dan nisbah isyarat kepada hingar (SNR) sistem tersebut. Keputusan menunjukkan bahawa SNR telah bertambah baik dan analisis telah dijalankan pada eksperimen yang dilakukan. Selepas itu, output daripada sistem pengecaman telah dipindahkan ke Arduino UNO untuk demonstrasi melaksanakan arahan yang diberi. Kereta robot boleh dikawal oleh 5 arahan suara asas iaitu berhenti, ke hadapan, ke belakang, pusing kiri dan pusing kanan dengan mengintegrasikan kod programing dalam MATLAB dengan mikropengawal Arduino UNO.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

HMM	-	Hidden Makrov Model
ANN	-	Artificial Neural Network
DTW	-	Dynamic Time Warping
VQ	-	Vector Quantization
LBG	-	Linde-Buzo-Gray
MFCC	-	Mel Frequency Cepstral Coefficients
FFT	-	Fast Fourier Transform
DCT	-	Discrete Cosine Transform
SNR	-	Signal to Noise Ratio

LIST OF APPENDIXES

CHAPTER 1

INTRODUCTION

This chapter present about the motivation, problem statement, objectives and scopes of the project.

1.1 Motivation

Living in this century full of development, world’s economy, military, healthcare, entertainment and transportation has been changed by the advanced technology which exists among all of us. With today technology, there are different ways to control appliances and devices without going near to the controlling button on the devices such as using remote control. One of the ways of controlling devices is by using voice recognition technology.

When voice control is mentioned, speech recognition is the first word to be considered. The term "voice recognition" is used to refer as speech recognition where the recognition system is trained to a particular speaker, hence there is an element of speech recognition, which attempts to identify the person speaking or to recognize what is being said [1]. However, there are differences between voice recognition and speech recognition. Voice recognition is a system relates to identifying voice of a particular user based on his or her unique vocal sound. On the other hand, speech recognition identifies almost anybody’s spoken words in the correct sense and then converting them into machine-readable language.

It is easier and more convenient to control or command a device by just speaking to it and it somehow increase the efficiency and effectiveness of working as people is able to work in parallel through this technology. The “Voice Recognition Robotic Car” is able to provide users a hands-free robotic vehicle control. Hence, it saves the time and effort of moving something from a distance. Once the system is implemented in a real vehicle, physically disabled or elder people are able to control their own vehicle without the help of others. Therefore with this voice recognition technology, it will directly enhance the quality of human lifestyle.

1.2 Problem Statement

Nowadays, vehicles are very important in order to ease daily job and improve the quality of life. Most of the vehicles are not friendly for physically disabled or handicapped user. Besides that, some operation such as police, military, rescue operation need unmanned vehicle to do the job as the situation they face daily is dangerous and sometimes inaccessible by human. Such job with high risk needs control in distance like voice control instead of hand control so that job can be done without risking human life or limb.

In voice recognition system, although different recordings of the same words may include more or less the same sounds in the same order, the precise timing or the durations of each sub word within the word will not match. Therefore the efforts to recognize words by matching the speech to pre-recorded speech templates will give inaccurate results because there is no temporal alignment. Besides that, noise that occurred in a sample of speech would affect the accuracy of recognizing a voice signal. As noise energy in a signal is more than the energy of a signal, the signal to noise ratio (SNR) is decreased. Once SNR is lower, the accuracy of recognizing words can be degraded.

1.3 Project Objectives

- To apply Vector Quantization technique in voice recognition system
- To apply filter technique to reduce the noise and increase signal to noise ratio
- To evaluate signal to noise ratio of before and after filtering technique applied

1.4 Project Scopes

Several scopes have been highlighted to achieve the objectives in this project. This project focuses mainly on the pre-processing stage that extracts salient features of a speech signal and technique called Vector Quantization (VQ) which used to compare the feature vectors of speech signals. These techniques were applied for recognition of isolated word only. Instead of recognizing a lot of words, the system recognized 5 basic commands such as forward, reverse, turn left, turn right and stop. In addition, with the implementation of this technique, only the user's voice can be recognized and therefore security of this system is guaranteed. MATLAB was used for the development of software part and to process the signals and recognize voice command. Analysis was done by using MATLAB in comparing the signal to noise ratio and distance of the pre-recorded signal. Microcontroller which is Arduino UNO was used to interface the voice recognition system with the output.

CHAPTER 2

LITERATURE REVIEW

2.1 Chapter Overview

In completing this project, literature review had been done to research and study the theory behind each algorithm in programming. This chapter shows the comparison made to choose the best way in completing this project. Articles, journals, conferences, and previous project from internet had served as resources of literature review.

2.2 Case studies of the project

Speech is a natural source of interface for human–machine communication, as well as being one of the most natural interfaces for human–human communication [2]. Speech Recognition or Voice Recognition technology promises to change the interaction between human and machines (robots, computers etc.) in the future. This technology is still improving and scientists are still working hard to cope with the remaining limitation. Nowadays this technology has been introduced to many important areas.

There are two categories of speech recognition which are speaker dependent and speaker independent. Speaker dependent is a system that trained by the user who will use the system. This system only responds accurately to the user that trained the system. The advantage of speaker dependent system is that it can achieve higher command count and better accuracy than speaker independent system. Meanwhile system independent is a system that responds to a word regardless of who is the one that speaks. Due to this reason, the system needs to respond to different kind of speech patterns, inflection and enunciation's of the target word. Command count for speaker independent system is usually lower than speaker dependent system but the accuracy can be maintained within processing limits. Normally in the field of industry, speaker independent voice system is required compare to speaker dependent because more people's speech can be identified instead of limits it down to the one who trained the system.

The most general form of voice recognition can be done through feature analysis which usually leads to "speaker-independent" voice recognition. This method processes the voice input using Linear Predictive Coding (LPC) or Fourier Transform technique and then will try to find the characteristic similarities between the expected input and actual voice input. These similarities will be present for a wide range of speakers, and so the system need not be trained by each new user. This method will not waste time on finding the match between the actual voice input and a previously stored voice template. Speaker independent method can easily deal with types of speech difference but fail to handle pattern matching, which including speaking accents of different nationalities, and varying speed of delivery, volume, pitch and inflection [3].

Speech recognition style can be a constraint to speech recognition system. There are 3 styles of speech which are isolated, connected and continuous. Isolated type is the most common type of speech recognition available today. This type of recognition can only recognize words spoken separately, each word or command spoken must following a short pause so that system can identify the word spoken. The second style is connected style which is in the middle of isolated and continuous. This style allows user to speak multiple words and can identify words or phrases in a duration of 1.92 seconds. The last style is the continuous style which is similar to the natural conversation speech used in daily life. This type is extremely difficult for recognizer to shift through the text as the word tends to merge together. Continuous recognition can be found on the market and still under development.

Besides the types of recognition, there are some approaches of statistical speech recognition. The most popular technique is the Hidden Markov Models (HMM)[7]. There are others technique that used for speech recognition system such as Artificial Neural Network (ANN) and Dynamic Time Warping (DTW). In HMM- based speech recognition, the audio signal could be viewed as a piece-wise stationary signal. This allows assumption that speech is approximately a stationary process in a short duration of time. Thus, speech can be thought as a Markov Model for many states. In addition, HMMs are popular because they can be trained automatically and computationally feasible to use. In speech recognition HMM provide the simplest setup possible by outputting a sequence of n dimensional real-valued vectors every 10 milliseconds with n value is more than 10. The vectors would consist of Cepstral coefficients, which are obtained by taking a Fourier transform of a short-time window of speech and de-correlating the spectrum using a cosine transform, then taking the most significant (first) coefficients [3].

Another approach in speech recognition is the use of Artificial Neural Networks (ANN). NN are capable of solving more complicated recognition tasks than HMM, but when the recognition needs lots of vocabularies, it does not scale as well as HMM. ANN can handle low quality, noisy data and speaker independence rather than just being used in general-purpose speech recognition applications. Therefore, ANN is more accurate than HMM based systems but in the condition that the vocabulary to be recognized is limited and there is training data [3].

Dynamic Time Warping (DTW) is an algorithm that measures similarity between two sequences which may vary in time or speed [11]. For example, the walking patterns of two people are being observed. DTW is able to detect the similarities between the two people even if they are walking in different speed, or even if acceleration and deceleration occurs in the observation. Hence, DTW which can analyze any data that can be turned into a linear representation has been applied to video and audio. One of the important applications is the automatic speech recognition where DTW is used to deal with different speaking speed. Utterance of a same word in a same duration will have a different speech pattern because the parts of word cannot be spoken at the same rates each time. Hence, time alignment needed to be done in order to obtain the differences between two audio signals. It allows computer to compare two given sequences and search for the optimal match between the signals.

Dynamic Time Warping (DTW) which will give a temporal alignment while comparing pre-recorded sample with the input speech signal. This will increase the accuracy of the recognition process as the distance of these signals has been reduced to the minimum which eased the matching of the voice signal. The technique, Dynamic Time

Warping (DTW), was introduced to the data mining community by Berndt and Clifford (1994) [11].

Suppose there are two time series *Q* and *C*, of length *n* and *m* respectively, where:

$$Q = q1,q2,\ldots,qi,\ldots,qn \qquad (2.1)$$

$$C = c1,c2,\ldots,cj,\ldots,cm \qquad (2.2)$$

In order to align two sequences using DTW an *n*-by-*m* matrix need to be constructed where the (*i*th, jth) element of the matrix contains the distance $d(q\text{i},c\text{j})$ between the two points *q*i and *c*j (With Euclidean distance, $d(q\text{i},c\text{j}) = (q\text{i} - c\text{j})2$). Each matrix element (*i*,*j*) corresponds to the alignment between the points *q*i and *c*j. This is illustrated in Figure 1. A warping path *W*, is a contiguous (in the sense stated below) set of matrix elements that defines a mapping between *Q* and *C*. The *k*th element of *W* is defined as *w*k = (i,j)k so:

$$W = w1,\ w2,\ \ldots,wk,\ldots,wK \qquad (2.3)$$

$$\max(m,n) \qquad K < m+n-1$$

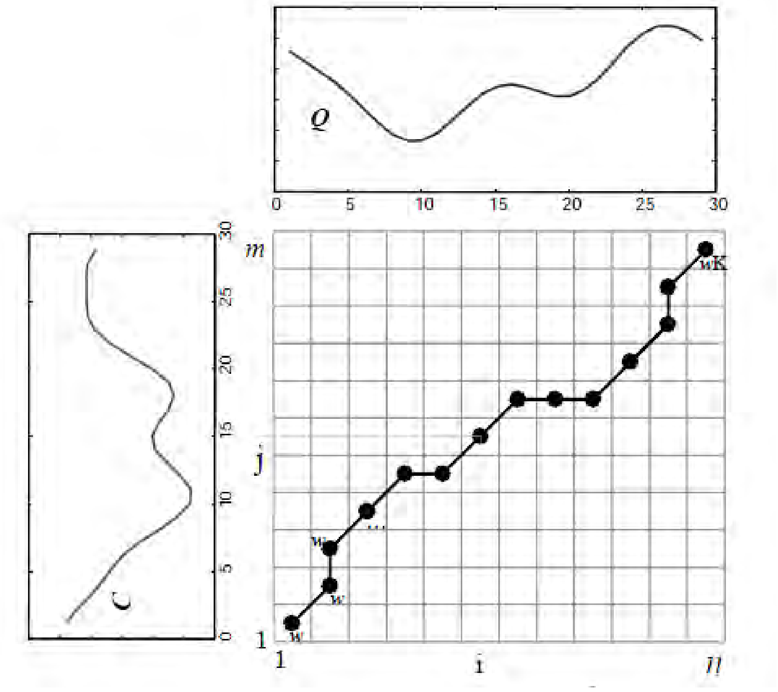

Figure 2.1: Example of warping path

There are several constraints where the warping path was subjected to. First constraint is the boundary conditions where it required the warping path to start and finish in diagonally opposite corner cells of the matrix. Second constraint is the continuity that restricts the allowable steps in the warping path to adjacent cells including the diagonally adjacent cells. The last constraint is the monotonicity where it forced the points in W to be monotonically spaced in time [11].

1. Boundary Conditions: w_1 = (1,1) and w_K= (*m*,*n*),

2. Continuity: Given w_k = (*a*,*b*) then w_{k-1} = (*a'*,*b'*) where $a–a' \leq 1$ and $b\text{-}b' \leq 1$.

3. Monotonicity: Given w_k = (*a*,*b*) then w_{k-1} = (*a'*,*b'*) where $a–a' \geq 0$ and $b\text{-}b' \geq 0$.

There are exponentially many warping paths that satisfy the condition above, the path that minimized the warping cost is focused.

$$DTW(Q,C) = \min \left\{ \sqrt{\sum_{k=1}^{K} \frac{w_k}{K}} \right. \qquad (2.4)$$

Where K in the denominator is used to compensate for the fact that warping paths may have different lengths.

Therefore this path can be found very efficiently using dynamic programming to evaluate the following recurrence which defines the cumulative distance g(*i*,*j*) as the distance *d*(*i*,*j*) found in the current cell and the minimum of the cumulative distances of the adjacent elements [11]:

$$(i,j) = d(qi,cj) + \min\{\ (i\text{-}1,j\text{-}1)\ ,\ (i\text{-}1,j\ )\ ,\ (i,j\text{-}1)\ \} \qquad (2.5)$$

Vector Quantization (VQ) is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a acoustic vector and can be represented by its center called a VQ codeword. The collection of a group of codeword was also called a codebook.

Figure 2.2 shows a conceptual diagram to illustrate this recognition process of using Vector Quantization. Example adapted from the journal showed only two speakers and two dimensions of the acoustic vector [15]. The circles refer to the acoustic vectors