International Journal of Engineering & Technology, 7 (3.14) (2018) 242-249

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Develop and Implementation of Voice Recognition Robotic Car

Hairol Nizam Mohd Shah[1*], Zalina Kamis[1], Mohd Fairus Abdollah[1], Mohd Shahrieel Mohd Aras[1], Faizil Wasbari[2], Nursabillilah Mohd Ali[1], Clement Chia Kuan You[1], Zairi Ismael Rizman[3]

[1]Center for Robotics and Industrial Automation, Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka, Malaysia
[2]Faculty of Mechanical Engineering, Universiti Teknikal Malaysia Melaka, Malaysia
[3]Faculty of Electrical Engineering, Universiti Teknologi MARA, 23000 Dungun, Terengganu, Malaysia
**Corresponding author E-mail: hnizam@utem.edu.my*

Abstract

The idea in this paper is to develop a voice recognition system that can recognized five commands to control a robotic car. The focus area is mainly on voice identification and recognition system. The aim of the system was not recognizing sentences but only isolated a word then demonstrates the action on a simple built robotic car. The system allows user to deliver voice commands through a microphone for control the movement of the car. Voice command is sent to computer and the process to 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 also improve signal to noise ratio. Result showed that 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.

Keywords: *Voice Recognition; Vector Quantization; Arduino.*

1. Introduction

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 [1-4]. 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.

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 [5]. 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.

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.

2. Related Work

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 [6]. 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 de-

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

pendent 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, 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 [7].

Besides the types of recognition, there are some approaches of statistical speech recognition. The most popular technique is the Hidden Markov Models (HMM) [8]. 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 [7].

Dynamic Time Warping (DTW) is an algorithm that measures similarity between two sequences, which may vary in time or speed [9]. Dynamic Time Warping (DTW) gives 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 in 1994.

Vector Quantization (VQ) is a process of mapping vectors from a large vector space to a finite number of regions in that space [10]. Each region is called an 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.

In the training phase, using the LBG algorithm can be used to generate speaker-specific VQ codebook for each known speaker by clustering his/her training acoustic vectors. The result codewords or centroids are shown in Figure 1 by black circles and black triangles for speaker 1 and 2 respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified as the speaker of the input utterance [11].

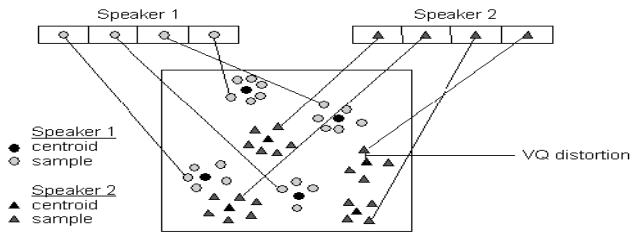


Fig. 1: Example illustration of vector quantization technique [11]

From the research by [12] in analysis of speech recognition technique has been made to show the comparison of HMM, ANN and DTW in terms of relevant variables, input and output. However, these techniques can only recognize the speech instead of recognizing the speaker. Therefore, Vector Quantization (VQ) technique was needed as this technique is more efficient in recognizing speaker than recognizing speech.

Adaptive Wiener Filter (WF) is a two-input technique, which provides a baseline for the performance of the in-line schemes [13]. The technique has a separate input for the noise and can provide complete noise cancellation, through the adaptive filtering process. Clearly, the scheme cannot be applied directly to the in-line noise reduction process. The design and performance of the filter depend on the complexity and speed of the adaptation algorithm.

A block diagram of the Adaptive Line Enhancer (ALE) was proposed by [14]. This is a modification of the standard line enhancer that removes narrowband noise from a broadband signal. In this implementation, the pitch detector provides a reference signal for voiced sounds and by applying a delay of one pitch period, the noise signal can be extracted as an error signal and used to characterize the filter parameters. The output of the filter provides the speech output.

The implementation proposed by [15] require a speech detector to decide, which of the two adaptive stages should be used at any particular time. The look-direction adaptation is used when speech is detected; otherwise the noise cancelling section is applied to adapt to the noise. This system thus relies on a speech detector for good performance. The ability of the system to perform limited dereverberation is claimed to be an asset for LPC-based speech recognizers [14].

Spectral Subtraction operates by making an estimate of the spectral magnitude during periods of no speech and subtracting this spectral estimate of the noise is from the subsequent speech spectral magnitude [15]. In common with the other techniques, spectral subtraction requires a speech detector. Because of its limitations, as noted above, leading and following frames are treated as speech. The technique is computationally expensive compared with the other techniques, due to the need to transform to and from the frequency domain. The technique is enjoying considerable attention, including the possible use of non-linear subtraction modifications have been proposed.

Based on the experiment conducted by [16] clearly show that the adaptive Weiner Filter provides the best noise reduction performance; the output signal-to-noise ratio for each speech sample approximates closer to the maximum possible value for that sample. The in-line processes do not perform as well, although the Spectral Subtraction method provides significant improvement for the lower input SNR values. As noted earlier, this method is somewhat more complex than the Adaptive Line Enhancer.

3. Methodology

There are two phase for the basic structure of the system, which is the training and testing phase. First and foremost, five command speech signals were recorded in .wav format as MATLAB can only read this format. Duration of each recorded speech was set to be constant which is 3 seconds for more efficient in comparing the distance for recognition. After speech was recorded, speech signals were read into MATLAB for feature extraction. Following steps after feature extraction was done was feature matching by applying vector quantization technique to determine the centroid or mean of the speech signals in acoustic vector form. Once the mean was determined, the mean was used to compare with the mean of the input speech signal from microphone. Euclidean distance was calculated and the speech signal was recognized based on the shortest distance calculated. These steps were described in details in the following section. Figure 2 shows the basic structure of the software and workflow of the software.

3.1. Feature Extraction

Feature extraction was done to obtain acoustic properties and speech data that used to define speaker individuality. Mel Frequency Cepstral Coefficients (MFCC) technique [17] was chosen to extract the feature of speech signals. Several steps in MFCC feature extraction was done and detailed in following section. All of the speech signal plotted in time domain was a raw data and has a great amount of data. It was difficult to analyze the voice characteristics and therefore feature extraction of the signal is needed.

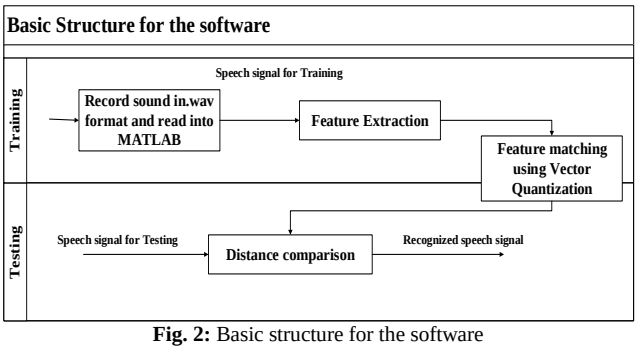


Fig. 2: Basic structure for the software

3.1.1. Step 1: Frame Blocking and Windowing

Command shown below was used in MATLAB to read the signal in order to plot.

[s fs nb] = wavread(''Folder1\voice'.wav')

The signal data was blocked in to frames of N samples. After framing speech signal, windowing was performed. The purpose of this step was to minimize speech signal's discontinuities at the beginning and end of each frame of speech signal. The concept was to minimize the spectral distortion as much as possible at the beginning and end of each frame. Window was defined as:

$$\omega(n), 0 \le n \le N-1$$

where N is the number of samples and with the Hamming window in (1).

$$\omega(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right); 0 \le n \le N-1 \qquad (1)$$

Frame of speech signal were then multiplied with Hamming window.

3.1.2. Step 2: Fast Fourier Transform (FFT)

Fast Fourier Transform (FFT) was done to convert framed speech signal's properties from time domain (linear) to frequency domain (logarithmic). A comparison was made between the energy level that plotted in linear and logarithmic. Power spectrum was computed using imagesc command and colorbar command to display the value in MATLAB. Figure 3 showed the linear and logarithmic power spectrum plot of one from the 5 signals. As shown in Figure 3, nothing obvious can be seen and analyzed in the linear power spectrum. It is a better view or more obvious by plotting power spectrum in logarithmic instead of linear as shown on the right hand side of Figure 3. Besides, the areas containing the highest level of energy were displayed in red and were located between 0.85 seconds to 1.30 seconds. The plot also showed that most of the energy was concentrated at lower frequency which is below 2000 Hz.

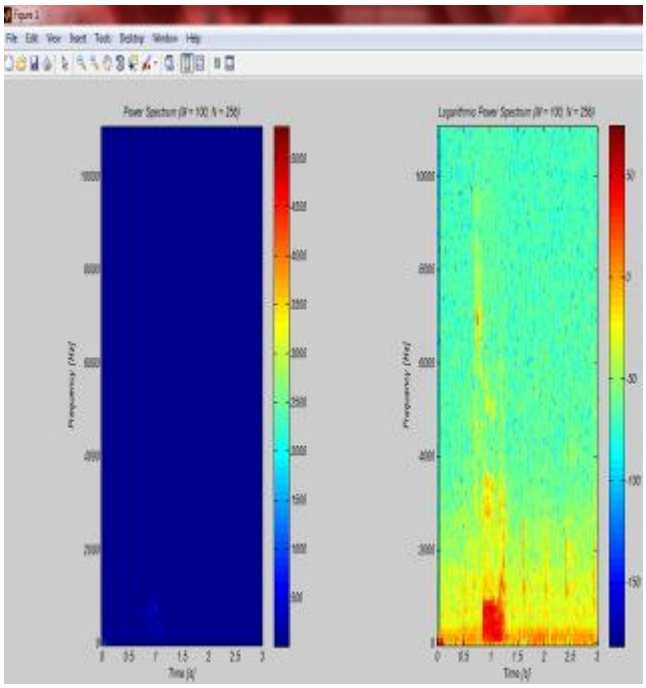


Fig. 3: Linear and logarithmic power spectrum plot

3.1.3. Step 3: Mel-Frequency Wrapping

Frequency, f that usually measured in Hz was measured on a scale called 'Mel' scale as properties of speech signals did not follow linear scale. For the reference of Mel scale, 1000Hz was defined as 1000 Mels. Approximated formula shown in (2) was used for the nonlinear transformation of frequency to Mel scale:

$$F_{mel} = 2595 * \log_{10} 1 + \frac{F}{700} \qquad (2)$$

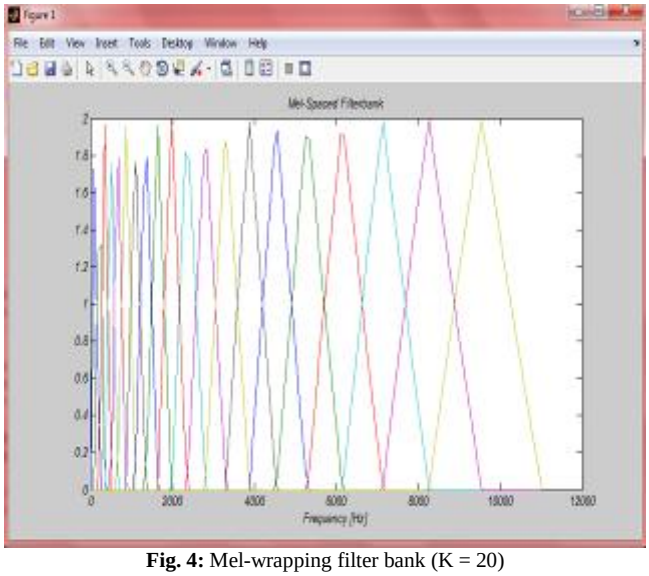


Fig. 4: Mel-wrapping filter bank (K = 20)

Figure 4 shows the Mel-wrapping filter bank which was the approach in simulating the speech signal. This filter bank was viewed as a group of filter in histogram that overlapped with each other in frequency domain. The number of Mel spectrum coefficients, K used in simulation was chosen to be 20. In addition, this filter bank has triangular band pass frequency response for each filter and the bandwidth was determined by constant Mel frequency interval as shown in Figure 4.

3.1.4. Step 4: Discrete Cosine Transform (DCT)

For the last step, Mel spectrum in logarithmic was converted back to time domain using Discrete Cosine Transform (DCT) and results was known as Mel Frequency Cepstral Coefficients (MFCC). The cepstral representation provided a better representation of speech signal's properties for analysis. Therefore, Mel power spectrum coefficients was represented in (3) and (4).

$$S_k, k = 1,2,....,k \tag{3}$$

where the MFCC was calculated in MATLAB using

$$\tilde{C}_n = \sum_{k=1}^{k} \left(\log \tilde{S}_k\right) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{k}\right], n = 1,2,....,k \tag{4}$$

which n = 0 was excluded as the coefficients, only carried very few of the speaker's specific information.

3.2. Feature Matching

In feature matching of speech signal, Vector Quantization (VQ) technique that included plotting VQ codebook and also implementing a well-known algorithm developed by Linde, Buzo and Gray which was called LBG algorithm was used.

3.3. LBG Algorithm

This algorithm was used to cluster acoustic vectors into codebook vectors. The procedure of how this algorithm works was shown below:

1. One vector codebook was picked which represent the centroid of every acoustic vectors
2. Size of the codebook was doubled and split based on the rule:

$$y_n^+ = y_n(1+\varepsilon)$$
$$y_n^- = y_n(1-\varepsilon)$$

where n starts from 1 to the chosen size of codebook and is a splitting parameter.

3. Codeword that closest to the chosen codebook in terms of similarity measurement was picked and vector was assigned to corresponding cell.
4. Codeword was updated by using the centroid of the acoustic vector assigned to the cell. This step also known as centroid update.
5. Step 3 and 4 was repeated until the average distance falls below preset range which is 10.
6. Step 2 to 4 was repeated until desired codebook size was reached.

The procedure was shown in Figure 5, where a clearer picture for the workflow of LBG algorithm.

3.4. Experiment Design

3.4.1. Repeatability Testing

In this experiment, each recognized command was tried for 20 times and then distance of each trial was recorded. Distance of each trials of recognition was then presented in a graph to evaluate the recognition system's repeatability. Figure 6 showed the workflow of this experiment

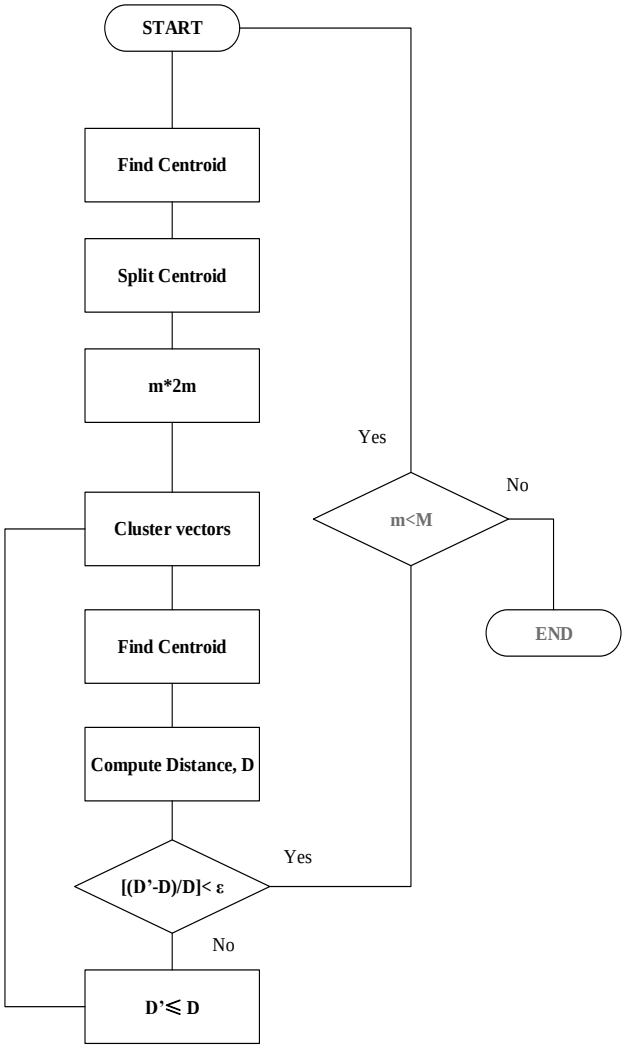


Fig. 5: Flow diagram of LBG algorithm

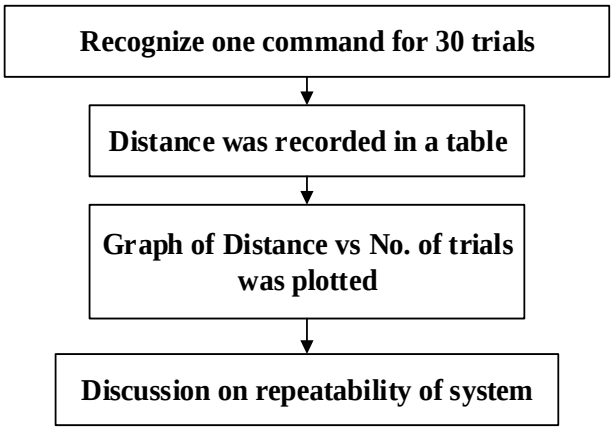


Fig. 6: Workflow of Experiment 1

3.4.2. Signal-to-Noise Ratio (SNR) Comparison

In this experiment, SNR for each speech signal stored in database was obtained using MATLAB. SNR before processed with Mel-wrapping filter bank was then obtained and recorded in a table for each command. After the signal processed with Mel-space filter bank, SNR of each signal was obtained again and recorded in the same table. Figure 7 showed the workflow of this experiment.

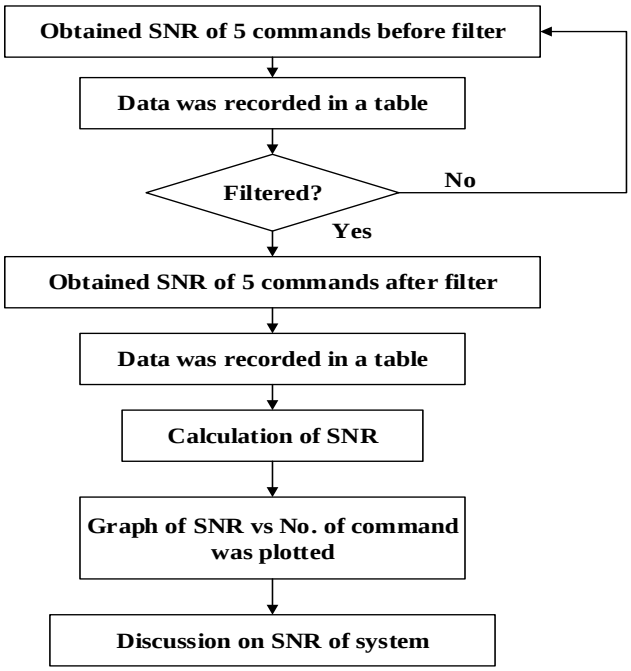


Fig. 7: Workflow of Experiment 2

4. Results and Discussion

4.1. Experimental Setup

The voice recognition system was integrated with a simple built robotic car, which controlled by Arduino UNO microcontroller as show in Figure 8 and 9. Once the system recognized a specific command in MATLAB, system printed an alphabet to Arduino and robotic car was controlled based on the programming code. The robotic car was built by two DC motor and using Arduino UNO microcontroller to control the movement. Once system recognized is "FORWARD" command, an alphabet "a" was sent to Arduino and following the condition where Arduino set motorLPin1 and motorRPin1 as HIGH. The following method was same as other commands for REVERSE, TURN RIGHT, TURN LEFT and STOP.

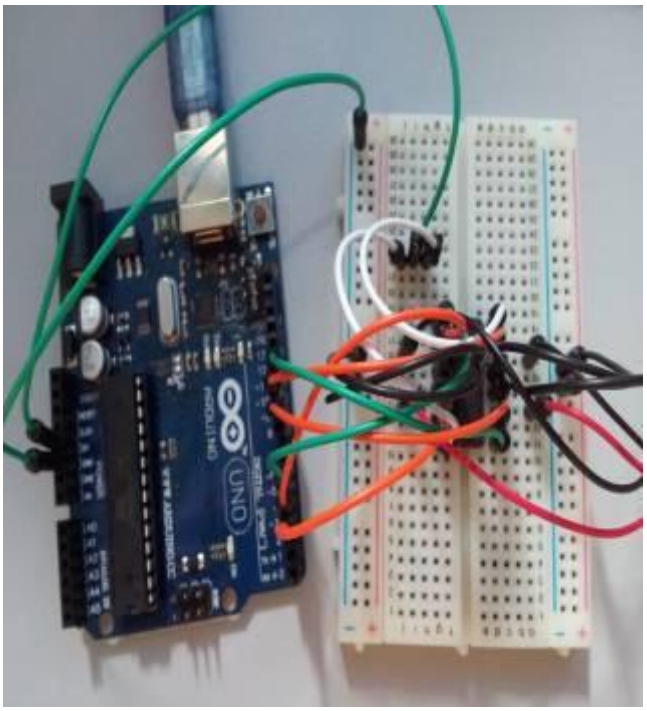


Fig. 8: Arduino UNO and H Bridge



Fig. 9: Simple built robotic car for demonstration

Table 1 showed the reference of which speech signal is which command that was used in this project. Results of the recognition were shown in following figure using MATLAB.

Table 1: References for speech signal

Speech Signal	Command
#1	FORWARD
#2	REVERSE
#3	TURN RIGHT
#4	TURN LEFT
#5	STOP

Figure 10 showed the result of recognition for this project. Menu was created to provide a button to start the recognition. Once button was clicked, recognition system asked for duration of the recordings and 3 seconds was entered for better recognition as speech signal stored in database was all recorded in 3 seconds. The system took some time to compute MFCC coefficients and VQ codebook of the input utterance as shown in Figure 9. After the progress was completed, distance was obtained by comparing the input utterance with all speech signal stored in database. As shown in Figure 10, the shortest distance from comparison was speech signal #1 and from Table 1, speech signal #1 was FORWARD command. Dialog function was added in source code to display the software simulation result. In Figure 11 to 14, the method of recognition is the same where speech signal with shortest distance after comparison, was taken to be the recognized command. Table 1 was referred to determine which speech signal is which command.

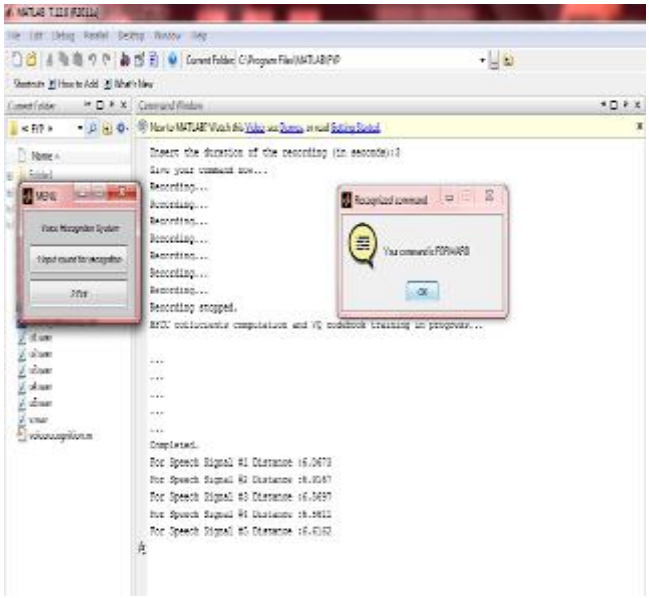


Fig. 10: Recognition of FORWARD command

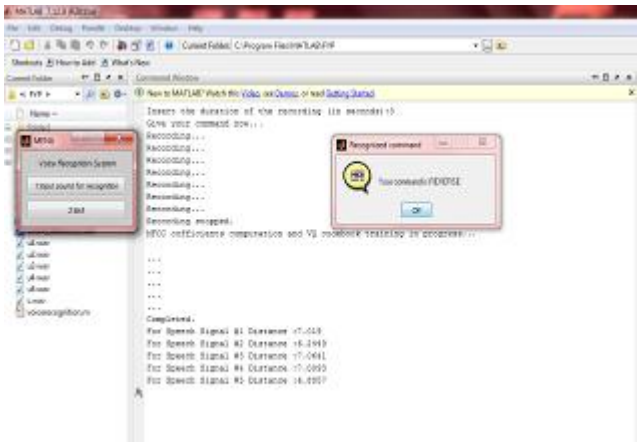


Fig. 11: Recognition of REVERSE command

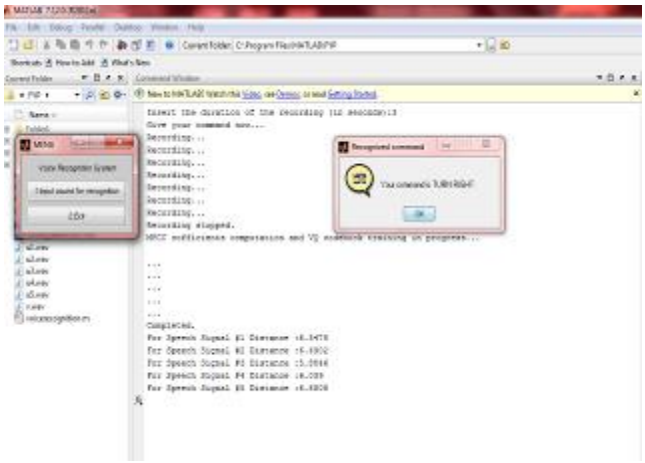


Fig. 12: Recognition of TURN RIGHT command

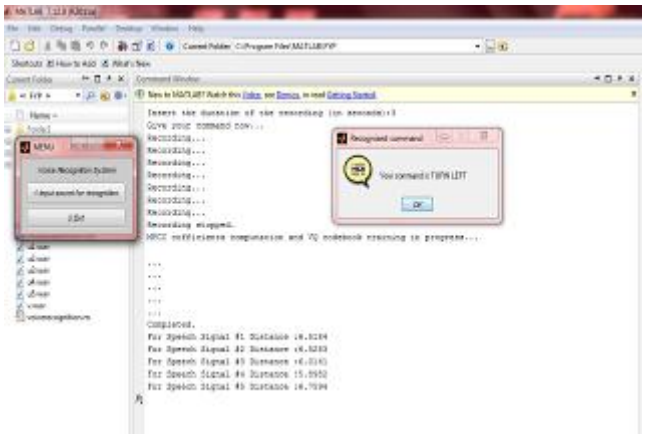


Fig. 13: Recognition of TURN LEFT command



Fig. 14: Recognition of STOP command

4.2. Repeatability Test

Table 2 showed the distance recorded for 20 trials of recognition. Distance recorded in Table 2 was the shortest distance after comparison between speech signal in database and input utterance that reflected which command given.

Table 2: Distance recorded for 20 trials of recognition

No. of Trials	Distance for recognition				
	Speech signal #1	Speech signal #2	Speech signal #3	Speech signal #4	Speech signal #5
1	6.0673	6.2449	5.8846	5.8982	5.6414
2	6.0982	6.235	5.8923	6.1012	5.6387
3	6.1104	6.2071	5.8822	5.9899	5.5998
4	6.1055	6.1998	5.9908	6.0422	5.6072
5	6.0587	6.214	6.0541	6.1288	5.5047
6	6.0462	6.2981	6.0123	6.0573	5.6328
7	5.9983	6.2573	5.998	6.0275	5.6332
8	6.0158	6.1877	5.8808	6.2043	5.6871
9	6.0873	6.2367	6.1203	5.9801	5.5989
10	6.1027	6.245	6.0238	6.0273	5.6071
11	6.1378	6.2755	5.7698	6.0586	5.6803
12	6.1572	6.1979	5.8098	6.0375	5.5587
13	6.0981	6.2654	6.0623	5.7929	5.5499
14	6.0654	6.3091	6.1002	5.8012	5.6293
15	6.0674	6.2733	5.9076	5.9099	5.6494
16	6.0891	6.2496	6.0963	5.8982	5.5879
17	6.0556	6.1769	5.8932	5.8849	5.5933
18	6.0389	6.2337	5.9902	5.9989	5.4989
19	6.1354	6.325	5.8803	6.0452	5.6328
20	6.1298	6.2977	6.0511	6.1658	5.6053

Figure 15 showed the result of that repeatability test presented in a graph. Different color of line represented different speech signal. Based on Figure 15, the graph showed that for 20 trials recognition of the command the shortest distance computed was consistent. The distance computed for each trial did not deviated much where the range is small as shown for speech signal #1, #2 and #5 in the graph. This reflected that the recognition's ability of this system processed high repeatability and can give consistent results. Speech signal #3 and #4 also had high repeatability but compared to the other 3 signal, the deviation is higher.

The reason that speech signals #3 and #4 had higher deviation was that the signal is "TURN RIGHT" and "TURN LEFT" command. These two command contained same utterance of word which is "TURN". The utterance of "TURN" might have affected the computation of distance and eventually affected the repeatability of recognition. Therefore, the repeatability of recognition was not very consistent for signal #3 and signal #4 compared to repeatability of recognition of other signals.

This problem happened to signal #1 and signal #2 when "FORWARD" command and "BACKWARD" command was decided to use. The recognition of these signals were unsuccessful as the system sometimes recognized "FORWARD" sometimes recognized "BACKWARD". This was due to the utterance or "WARD" which almost the same command that affect the computation of distance. "FORWARD" command was given but system computed the shortest distance for "BACKWRD" command. In order to cope with this, "REVERSE" command was used to replace signal #2 to avoid the uncertainty in recognition.

In conclusion for this experiment, distance deviation for 20 trials was exponentially inverse proportional to the repeatability of the recognition. The larger the distance deviation, the harder the comparison can be made accurately, which affect the accuracy of the system indirectly. The shorter the distance, the easier the comparison can be made and system result in high repeatability of recognition. Figure 16 showed the relationship of repeatability and distance that can be concluded from this experiment.

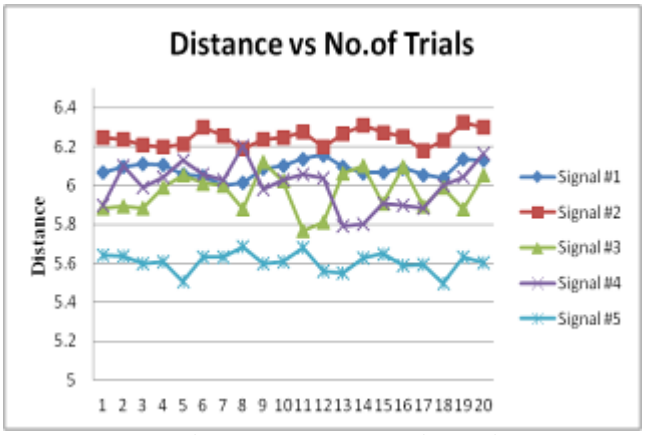


Fig. 15: Graph of distance versus number of trials of recognition

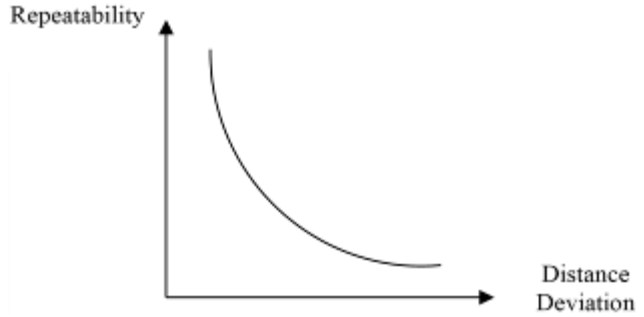


Fig. 16: Relationship of repeatability versus distance deviation

4.3. Signal-to-Noise Ratio (SNR) Comparison

Table 3 showed the comparison of SNR before applying Mel-wrapping filter bank and after Mel-wrapping filter bank.
In order to calculate the SNR, assumption has to be made which is the signal and noise was measured in the same impedance as in (5) to (7).

$$SNR_{dB} = 10\log_{10}\left(\frac{A_{Signal}}{A_{Noise}}\right)^2 \quad (5)$$

where A_{signal} = Root Mean Square Amplitude of Signal, A_{noise} = Root Mean Square Amplitude of Noise and A_T = Total Root Mean Square Amplitude.

$$A_T = \frac{A_{Tpeak}}{\sqrt{2}} \qquad A_{Signal} = \frac{A_{Speak}}{\sqrt{2}} \quad (6)$$

Getting the peak amplitude to calculate the root mean square amplitude and then the root mean square amplitude of noise can be assumed that:

$$A_{noise} = A_T - A_{Signal} \quad (7)$$

Figure 17 presents the graph of comparison of signal to noise ratio (SNR) for before and after applying the filter bank. Data was obtained from Table 3 and speech signal of which command can be referred from Table 1. For a signal to consider good, the SNR must be the higher the better where it means that the noise energy inside the signal is lesser than the signal energy. As shown in Table 3, the root mean square amplitude of every signal was increased after Mel-wrapping filter bank was applied compared to before the filter bank applied. When the amplitude of signal was higher and the amplitude of noise it lower, the SNR was improved. It is clearly seen that SNR was higher after the Mel-space filter was applied as shown in the graph.

Table 3: Comparison of signal to noise ratio

Speech Signal	Before Mel- wrapping filter bank				After Mel- wrapping filter bank			
	A_T	A_{signal}	A_{noise}	SNR(dB)	A_T	A_{signal}	A_{noise}	SNR(dB)
#1	0.4806	0.3085	0.1721	5.0695	0.4806	0.3675	0.1131	10.2359
#2	0.4033	0.3076	0.0957	10.1415	0.4033	0.3382	0.0651	14.3119
#3	0.7016	0.4638	0.2378	5.8024	0.7016	0.5148	0.1868	8.8052
#4	0.7016	0.4324	0.2692	4.1162	0.7016	0.5273	0.1743	9.6152
#5	0.7016	0.4659	0.2357	5.9187	0.7016	0.5082	0.1934	8.3916

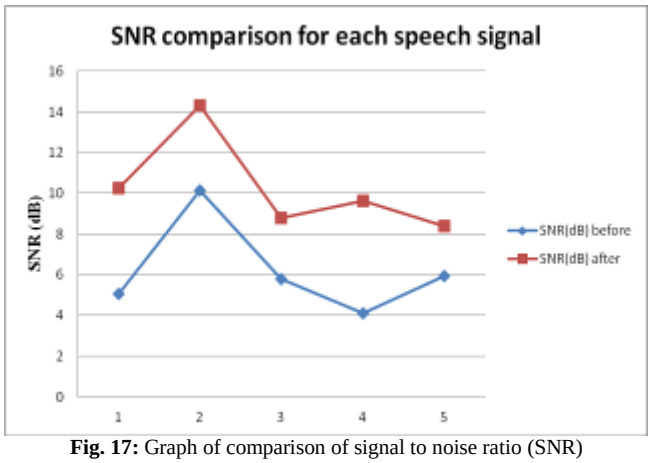


Fig. 17: Graph of comparison of signal to noise ratio (SNR)

5. Conclusion and Recommendations

In conclusion, the LBG algorithm used in VQ technique provided a simple and easy way for understanding and also to recognize a voice command. In addition, distance comparison for each trial of recognition provides consistent shortest distance and therefore, the system was concluded as provide high repeatability recognition. Possible error such as utterance of words had been solved by using different command with same meaning to avoid the uncertainties in recognizing each command.
Furthermore, the findings through signal to noise ratio (SNR) comparison of this paper proved that SNR of the speech signal was improved after the Mel-wrapping filter bank applied to filter the speech signal. Comparison on a graph also showed that the SNR after the filtering technique for every signal was higher than the SNR before any filtering technique. Assumption was made for the analysis to cope with possible error in calculation as the power of signal and power of noise cannot be obtained to calculate SNR using theoretical formula.
In future work, improvement can be made by implementing Artificial Neural Network (ANN) technique into the voice recognition system. The advantage of using ANN is that this technique can handle low quality, noisy data by training and testing the network for a couple of times. While, ANN can handle low quality or high noise power data, filtering technique can be removed from source code of the system to prevent code redundancy. However, this recommendation is optional and need to base on objectives set for a project.
On the contrary, for the real-time comparison of distance, Dynamic Time Warping (DTW) technique can be used. This technique function almost the same as Vector Quantization (VQ), where both technique compare signals and find the shortest distance to recognize the signal. However, for VQ technique data samples need to be convert to acoustic vector form in frequency domain for the comparison to be done but for DTW, comparison of distance can be made without converting signal to frequency domain. DTW compare signal in time domain and also provide real-time comparison.

Last but not least, instead of using cable, the serial communication between MATLAB programming code and Arduino UNO micro-controller can be replaced with wireless connection so that the movement of the robotic car will not be limited. For this improvement, XBee wireless RF module was recommended. The communication between software and hardware will become less limitation and this XBee chip help to ease the movement of robotic car.

Acknowledgement

The authors are grateful for the support granted by Center for Robotics and Industrial Automation, Universiti Teknikal Malaysia Melaka (UTeM) in conducting this research through grant PJP/2018/FKE(4C)/S01605 and Ministry of Higher Education.

References

[1] Hairol Nizam Mohd Shah, Mohd Zamzuri Ab Rashid, Zalina Kamis, Mohd Shahrieel Mohd Aras, Nursabillilah Mohd Ali, Faizil Wasbari, Tengku Muhammad Mahfuz Tengku Anuar, "Sign Detection Vision Based Mobile Robot Platform", Indonesian Journal of Electrical Engineering and Computer Science, vol 7(2), pp. 524-532, 2017.

[2] Hairol Nizam Mohd Shah, Mohd Zamzuri Ab Rashid, Zalina Kamis, Muhammad Nizam Kamarudin, Mohd Fairus Abdollah, Alias Khamis, "Implementation of Object Recognition Based on Type of Vehicle Entering Main Gate", Indonesian Journal of Electrical Engineering and Computer Science, vol 3(2), pp. 458-467, 2016.

[3] Hairol Nizam Mohd Shah, Mohd Zamzuri Ab Rashid, Nur Maisarah Mohd Sobran, Rozilawati Mohd Nor, Zalina Kamis, "Autonomous Mobile Robot Vision Based System: Human Detection By Color", Journal of Theoretical & Applied Information Technology, vol. 55(2), pp. 183-189, 2013.

[4] HNM Shah, MZA Rashid and YT Tam, "Develop and Implementation of Autonomous Vision Based Mobile Robot Following Human", International Journal of Advanced Science and Technology, vol. 51, pp. 81-91, 2013.

[5] T. Z. Qi and T. J. Moir, "Automotive Speech Control in a Non-Stationary Noisy Environment,"15th International Conference on Mechatronics and Machine Vision in Practice, pp. 374–377, 2008.

[6] Lawrence Rabiner, and Biing Hwang Juang, Fundamentals of Speech Recognition, Prentice Hall, New Jersey, 1993.

[7] P. Chopra and H. Dange, "Voice controlled robot", Department of Electronics Engineering, K.J. Somaiya College of Engineering, Vidyavihar, Mumbai, 2007.

[8] G. Talwar, R.F.Kubichek, and H.Liang, "Hiddenness Control of Hidden Markov Models and Application to Objective Speech Quality and Isolated- Word Speech Recognition," Fortieth Asilomar Conference on Signals, Systems and Computers, pp. 1076–1080, 2006.

[9] E. J. Keogh and M. J. Pazzani, "Scaling up Dynamic Time Warping to Massive Datasets 1 Introduction 2 The Dynamic Time Warping Algorithm," no. Derriere, 1998.

[10] P. Ch and S. Kumar, "Design Of An Automatic Speaker Recognition System Using MFCC , Vector Quantization And LBG Algorithm," vol. 3(8), pp. 2942–2954, 2011.

[11] D. Gupta and R. M. C, "Isolated Word Speech Recognition Using Vector Quantization (VQ)," vol. 2(5), pp. 164–168, 2012.

[12] M. Cowling and R. Sitte, "Analysis of Speech Recognition Techniques for use in a Non-Speech Sound Recognition System," pp. 16–20.

[13] Wiener N, "Extrapolation Interpolation and Smoothing of stationary Time Series", Wiley, USA pp. 10-15, 1949.

[14] Sambur, M.R. "Adaptive Noise Canceling for Speech Signals", LEEE Trans. Acoust .,.-- Speech, Signal Processing, vol. ASSP-26, pp.419-423, 1978.

[15] Van Compernolle, D. "Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings", lEEE Int. Conf. Acoust. Speech & Signal Processinq, Albuquerque, pp. 833-836, 1980.

[16] A.G. Maher, "A Comparison of Noise Reduction Techniques for Speech Recognition in Telecommunications Environments," vol. 5, 1992.

[17] Hairol Nizam Mohd Shah, Mohd Zamzuri Ab Rashid, Mohd Fairus Abdollah, Muhammad Nizam Kamarudin, Chow Kok Lin, Zalina Kamis, "Biometric voice recognition in security system", Indian Journal of Science and Technology, vol. 7(2), pp. 104-118, 2014.