Design and implementation of an automatic speech recognition based voice control system

Adnene Noughreche, Sabri Boulouma and Mohammed Benbaghdad

March 9, 2021

Design and implementation of an automatic speech recognition based voice control system

Noughreche Adnene[1], Boulouma Sabri[2*], Benbaghdad Mohammed[1]

Abstract

This paper deals with the study, design, and implementation of an automatic speech recognition-based voice control system for small vocabulary isolated words, the aim is to control the motion of a small mobile robot. The speech recognition algorithm was implemented using Mel frequency cepstral coefficients (MFCC) for feature extraction and the dynamic time warping (DTW) for pattern matching. The NI MyRIO-1900 embedded module and LabVIEW are used for the real-time implementation; it is built around an interactive Human Machine Interface (HMI). The system is designed to recognize eight commands, including a wakeup word. The developed ASR based control system has the ability to learn new commands, and to adapt to new users and even new languages. This system has an improved ability to recognize and execute commands with high accuracy in real-time. It was able to achieve a recognition accuracy of 99.75% for one speaker using a dictionary of 10 word sets and an accuracy of 96.375% for a set of three different speakers using a hybrid dictionary with 10 word sets, when the word sets in the hybrid dictionary is raised to 30, the designed system is able to achieve a recognition accuracy of 99.5 event with three different speakers.

Keywords

Mobile robot — Voice control — Speech recognition— MFCC— Dynamic time warping, NI-myRIO.

[1] Ecole Supérieure des Transmissions, Koléa, Tipaza, Algeria

[2] *Unité de développement des Equipements Solaires, UDES/Centre de Développement des Energies Renouvelables, CDER, Bou-Ismail, 42415, W. Tipaza, Algérie*

* *Corresponding author: benbagh.moh@gmail.com*

Introduction

Automatic speech recognition (ASR) allows a machine to understand human speech and accordingly, implement different actions and interactions depending on the intended application. ASR systems are ubiquitous and omnipresent, they are used to assist people with disabilities and meet their needs, they are used in automatic call processing and customer services, High-performance fighter aircraft, Data entry, voice dictation and speech transcription among many other applications [1,2]. Different types of ASR systems can be implemented with different levels of complexity. Generally, ASR systems are classified according to many criteria, namely. The number of speakers, the nature of the utterance, the vocabulary size and the spectral bandwidth [2,3].

In recent years, ASR has received great attention from researchers in industry and academia. This led to tremendous advancements in the field. From the view point of industry, many high-tech actors have developed their own ASR systems for mobiles (Apple Siri), for home assistant (Google home and amazon Alexa echo dot), and for PCs such as Microsoft Dragon and Microsoft Cortana. From the academic point of view, many research works have been carried out in this context. In [4], An ASR system for robotic applications was proposed using Linear predictive coding (LPC) and hidden Markov models (HMM). In [5], another robotic application for ASR was proposed, where the dynamic time warping (DTW) approach and Mel frequency cepstral coefficients (MFCC) method were used, the same commands were considered, the command dictionary is created offline, with only one user, the resulting accuracy was about 91%. In [6], MFCC feature extraction and DTW method were used for a communication System with hearing-impaired people based on isolated words of the Arabic language, the algorithm was able to achieve about 98% accuracy. Always with the Arabic language, an isolated words automatic Arabic speech recognition system was developed in [7], where the LPC and HMM methods were used for feature extraction, a new artificial neural network (ANN) based algorithm was used for classification. In [8], a speech recognition system was developed. Two methods LPC and DWT were used for feature extraction, and ANNs were used for pattern classification to recognize speaker-independent spoken isolated words from Malayam dialect, 50 speakers uttered 20 isolated words for each of them. Both the methods produce good recognition accuracy, but discrete wavelet transforms are found to be more suitable for recognizing speech because of their multi-resolution characteristics and efficient

time-frequency localization. In [9], a syllable-based ASR system was developed for the Indonesian language, the obtained recognition accuracy reached 78%. In [10], a detailed theory about feature extraction techniques like LPC and LPCC is examined. An isolated word ASR system was developed in [11] using LPC for feature extraction, Vector Quantization (VQ) method for classification, 12 words of Lithuanian language pronounced ten times by ten speakers were the database that the system was worked with. In [12], the authors proposed a new speech recognition approach for recognition and identifying Vietnamese voice commands on a smartphone using support vector machines (SVM) and convolutional neural networks (CNN).
In this work, the aim is to design and implement a dedicated customizable ASR system for voice control of mobile robot's motion. The remainder of this paper is organized as follows, in section 2, the theoretical background of the ASR algorithm is presented, in section 3, system implementation and key issues related to the software/hardware architecture are presented. in section 4, implementation, test results and discussion are presented. the paper terminates with a conclusion and perspectives for future work.

1. ASR Based Voice control system

1.1. Overview

An ASR system comprises a set of steps as shown in Figure 1. These steps include, the speech acquisition and digitization step, the speech preprocessing and enhancement step, the feature extraction step, and finally, the classification/matching step followed by the decision or output step.

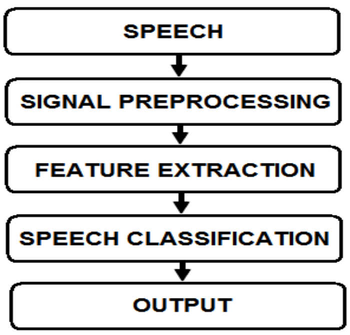

Figure 1: Basic structure of an ASR system

1.2. Speech Acquisition and preprocessing

Once the speech signal is acquired in a digital form, some processing operations should be applied. This step is necessary to get reliable speech features in the presence of different factors that may reduce the performance or the ASR system [14]. Speech preprocessing involves many tasks. First, the speech signal is pre-emphasized prior to any further processing. At this step, the high-frequency contents of the input signal are emphasized to flatten its spectrum [2] and to boost the energy amount in the high frequencies. This is achieved by a first order FIR filter given by:

$$s_{preemph}[n] = s[n] - as[n-1] \quad (1)$$

where the coefficient a is generally taken $a = 0.97$. [2]

The second step in preprocessing is global thresholding, this step aims to detect only intelligible speech signals and disregard background noise or silence. Global thresholding is carried out at the utterance level and is implemented according to the following equation:

$$s[n]_{th} = \begin{cases} s[n], & \text{if } E > E_{th} \\ 0 & \text{, otherwise} \end{cases} \quad (2)$$

where $E = \sum_{n=0}^{N-1} |s(n)|^2$ is the total energy of the utterance and E_{th} is a fixed energy threshold.
The third step in preprocessing is normalization, since speakers defer in speech loudness [5]. On the other hand, different microphones defer in their sensitivity to speech. Thus, for a more reliable ASR system, the speech signal is normalized according as folows [5]:

$$s[n]_N = \frac{s[n] - mean\{s[n]\}}{\max\left|s[n] - mean\{s[n]\}\right|} \quad (3)$$

1.3. Framing and windowing

Speech signal is inherently nonstationary. However, for short time frames of typically 20-30ms, it can be safely considered stationary [2]. Windowing is used to reduce spectral leakage and smoothen the FFT [2, 3], the Hamming window is widely used in speech applications, it is defined as follows:

$$w_{ham}[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \; 0 \le n \le N-1$$
(4)

Furthermore, to increase the number of frames, and enhance the spectral representation quality, an overlapping between frames is applied.

After framing and windowing of the speech signal,a voice activity detection (VAD) algorithm is applied, VAD tries to determine which frames contain speech and which frames do not. When a person speaks, there will be pauses, hesitations, etc. in his speech. So, those non-speech sounds or silences should be first removed prior to feature extraction [1, 5].

1.4. Feature extraction using MFCCs

The goal of feature extraction is to reduce the undesirable variability and help the classification process. The MFCC method is a widely used for this purpose. MFCCs used in speech recognition are based on frequency domain using the Mel frequency scale based on the human ear perception scale. The MFCC computation process is presented in Figure 2.

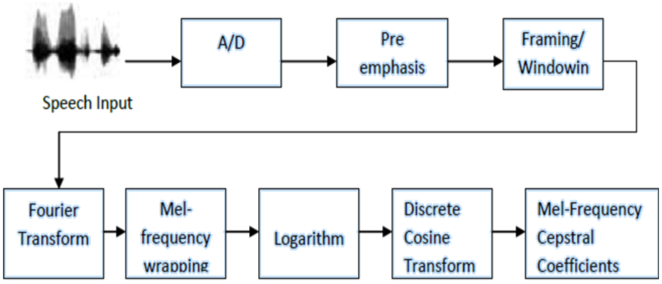

Figure 2: Main steps for the MFCC feature extraction

In the first step, the Discrete Fourier Transform (DFT or FFT) is taken for each frame as follows:

$$S[k]=\sum_{n=0}^{N-1} s[n]e^{-j\frac{2\pi nk}{N}},\ \text{k=0,1,...,N-1} \quad (5)$$

Notice that the FFT is computed for over N points, however only N/2 values are needed to avoid redundancy.
Now, the Mel filter bank is created using the Mel Frequency scale. In the Mel scale, the frequency spacing below 1 kHz is linear and the frequency spacing above 1 kHz is logarithmic. The Mel scale is computed from the fre quency (in Hertz) as follows:

$$f_{Mel}=2595\log\left(1+\frac{f}{700}\right) \quad (6)$$

A Mel filter bank consists of half overlapping triangular filters linearly distributed on the Mel-frequency scale as shown in Figure 3. The main parameters in the filter bank are, the number of triangular filters (between 20 and 40), lower frequency of the left-most filter, and higher frequency of the rightmost filter [1,5]. The maximum frequency should be below the Nyquist frequency $f_N = f_s/2$ while the minimum frequency should be selected above 100 Hz. Typical values used for a sample rate of 11025 Hz is $f_{max} = 5400\ Hz$ and $f_{min} = 100\ Hz$.

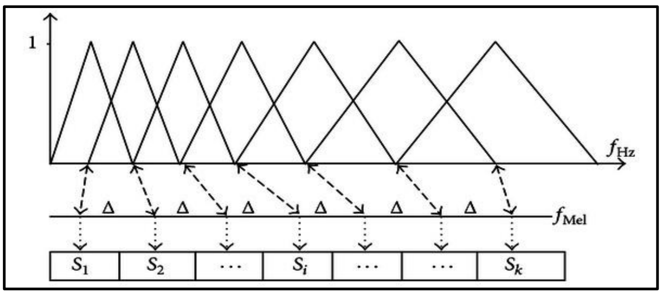

Figure 3: The Mel scale filter bank

If $S[k]$ is the power spectrum density with N samples, L is the number of filters. The output of each filter is given by the sum of every single filter's spectral components as:

$$R[l]=\sum_{k=0}^{N/2} S[k]M_l[k],\ \text{l=1,...,L-1},\ k=\frac{kf_s}{N} \quad (7)$$

Next, the logarithm of the square magnitude of the output of the filter bank is taken. Log energy computation is useful for capturing the dynamic variation in the human hearing system and having features less sensitive to input variations [5,12]. It is given as follows:

$$N[l]=\log_{10}\left[\sum_{k=0}^{N/2} S[k]M_l[k]\right] \quad (8)$$

Finally, the discrete cosine transform (DCT) is used to transform the lop power spectrum to time domain instead of the IDFT. The DCT leads highly uncorrelated features and a smooth spectrum which is useful to separate the vocal tract source and the excitation [3]. The DCT is given by:

$$C[i]=\sqrt{\frac{2}{L}}\sum_{l=1}^{L} N[l]\cos\left(\frac{\pi i}{L}\left(l-\frac{1}{2}\right)\right), i=1,...,C \quad (9)$$

where L is the total number of triangular Mel filters and C is the number of MFCCs of interest.
The set of MFCCs provides perceptually meaningful and smooth estimates of the speech spectrum over time. Since speech is inherently dynamic, it is reasonable to seek a representation that includes some aspects of the dynamic nature of the speech. The resulting parameter sets are called the delta cepstrum (first derivative) and the delta-delta cepstrum (second derivative). The delta cepstrum coefficients can be obtained by the difference of cepstral vectors, as:

$$d_i=\frac{\sum_{n=1}^{N} n(c_{n+i}-c_{n-i})}{2\sum_{n=1}^{N} n^2} \quad (10)$$

The delta-delta coefficients are calculated by applying the same formula on the delta coefficients.
In summary, the number of MFCCs considered in this work is 12 for MFCC, 12 delta MFCCs, and 12 delta-delta MFCCs. Besides, the logarithm of frame energy, the delta-energy and the delta-delta energy are also introduced as additional feature along with the delta and delta-delta MFCCs, which leads to 39 MFCCs per frame. This is summarized in table.1.

Table 1: Total number of MFCC features per frame

Feature name	# of elements
Mel Cepstral Coefficients	12
Delta Mel Cepstral Coefficients	12
Delta-Delta Mel Cepstral Coefficients	12
Energy Coefficient	1
Delta Energy Coefficient	1
Delta-Delta Energy Coefficient	1

1.5. Classification using DTW algorithm

The DTW algorithm uses two sequences of feature vectors and iteratively warps the time axis until there is an optimal path between the two sequences [5, 13]. Since a word can be said in different ways (Slow, fast, high pitch, low pitch), this method is very efficient in this aspect.
The first step in DTW is to create the cost matrix D, where the two sequences represent each axis, $X=[x_1,x_2,...,x_m]$, $Y=[y_1,y_2,...,y_m]$. This matrix is computed as follows:

$$D(i,j)=d(i,j)+\min\begin{Bmatrix} D(i-1,j) \\ D(i,j-1) \\ D(i-1,j-1) \end{Bmatrix} \quad (11)$$

When all elements in the matrix are calculated, we create a warp path, W = $(w_1, w_2, w_3, \dots, w_k)$ begins. This is done by backtracking and greedy search to minimize the distance given by:

$$Dist=\sum_{k=1}^{k=L} Dist(w_{ki},w_{kj}) \quad (12)$$

In summary, the ASR algorithm can be summarized in the following flowchart.

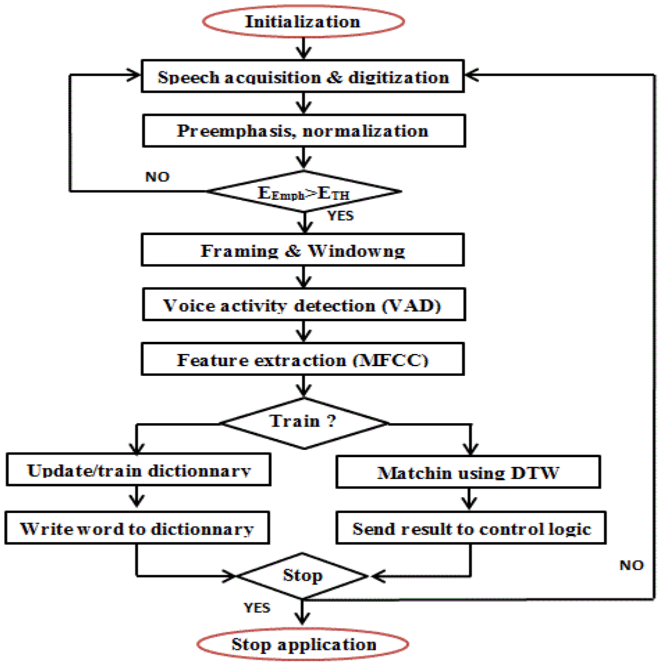

Figure 4: Flowchart of the ASR algorithm

1.6. Decision and control logic

After recognition, a command should be executed according the following control logic:
- The system always stores the previous command from the matching result.
- It reads the newly acquired utterance and if it matches one of the 8 commands it will carry on the corresponding tasks according to which recognized command.
- If the matched word is MyRIO, then the action corresponding to wake up will be executed regardless of the previous command.
- If the matched word is one of the six other commands, the corresponding command will be executed and its opposite command is cancelled.
- If the matched word is Stop, the system goes into a standby mode, and will disregard all subsequent commands until the wake-up word is said again.
- The current command (to be run) is set as the previous command for the next iteration and so on.
- The system runs according to this logic until the user hits the stop application button, when the application will be closed (no acquisition and no recognition).

2. Results and Discussion

2.1. Design specifications

In the developed system,the number of words used in this system is restricted to 8 words, which are the wake-up word, that we call "*myRIO*". The other commands are actions to be executed by the controlled system (vehicle or robot, etc), these words are respectively: *"forward", "backward", "turn right", "turn left", "lights on", "lights off", "stop"*, each word is repeated and saved 10 times in the dictionary to enhance the recognition system accuracy, regarding speech signal variability. The system will wait for 3 seconds until the user utters the word. In this system, we have considered a constant value for each word duration, which is 0.6 seconds. An energy thresholding level of 10 is used for capturing only the intelligible speech signal. This system is used to control the motion of a mobile robot and the status of the LEDs (light or turn off the LEDs).

2.2. Hardware architecture

LabVIEW and NI myRIO-1900 module are used to implement the system. This module uses the NI reconfigurable input output (RIO) industry standard which comprises a real-time processor and an FPGA target. The general specifications of the NI myRIO-1900 are summarized in table 2. [14]

Table 2: hardware specifications of the NI-myRIO module

NI myRIO-1900 specifications	
Processor	Xilinx Z-7010 667 processor 2 cores
Non voltatile memory	256 MB
DDR3 volatible memory	512 MB
DDR3 clock frequency	533 MHz
DDR3 databus width	16 bits
FPGA type	Xilinx Z-7010, 40 Mhz.
Operating system	Real-time Linux
Programming language	LabVIEW and C/C++

For the speech recognition part, the myRIO is the master controller, which receives the voice commands from the speaker, carries out the ASR task according to the flowchart of Figure.3 and sends the corresponding control signals to the motors of the mobile robot and the LEDs, this is shown Figure 4.

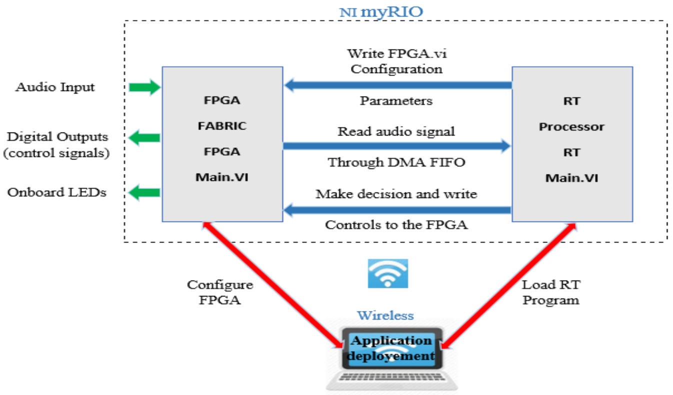

Figure 5: Architecture of the developed application

2.3. Software architecture

The developed application consists of two main segments of code: the FPGA configuration code that defines the personality of the FPGA fabric and RT Main.vi that implements the ASR and control algorithms.

The FPGA-Main program:

For the FPGA target configuration program is shown in figure 6, the audio signal is acquired as a mono (Left channel). A DMA FIFO – FPGA to RT (Target to Host) is used to transfer data from FPGA to RT processor, also the four inbuilt LEDs on myRIO are programmed, which indicates when the wake-up word is recorded. Besides, nine digital I/Os are programmed, these I/Os are further connected to LEDs on the breadboard to indicates the time interval when a word can be said. Also, seven digital I/Os are programmed to refer to the word was said by the user. As the wheeld robot is not fully customized for our voice control system, only four I/O corresponding of the four movements (Forward, Backward, Turn Right, Turn Left) are connected to an ENCODER HT12E to convert their Boolean values from parallel to serial then send it through wireless communication using the RF 433 MHz transmitter for the emission and RF 433 MHz receiver for the reception, then through a HT12D DECODER to the motors' drive of the mobile wheeled robot.

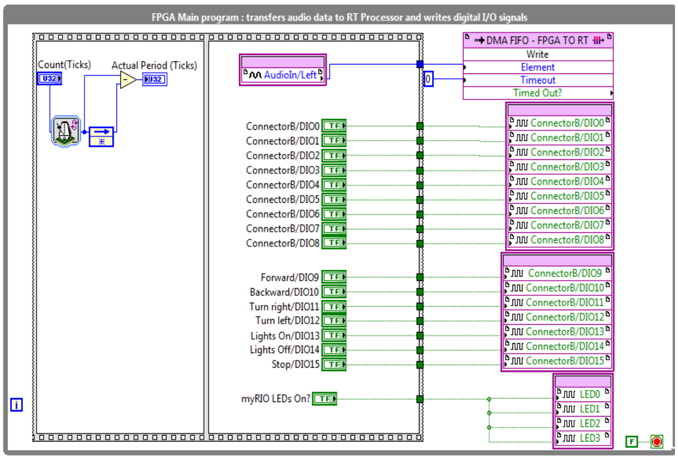

Figure 6: G-code for the FPGA target

The RT Main.vi code:

First, the program should open a reference to communicate to the FPGA target. It is then possible to read indicators and DMA-FIFOs and writes to controls and DMA-FIFOs in the FPGA target. The audio sampling rate is 11025 Hz, which is in the interval of appropriate sample rate for human speech. By dividing the internal clock rate of the FPGA (40MHz). the RT code contains the preprocessing, feature extraction, classification and the control logic subVIs. In the control logic, information about the matched word and the previous command is used to execute adequate action. The GUI (front panel of the RT Main.vi) for the application is shown in Figure 7.

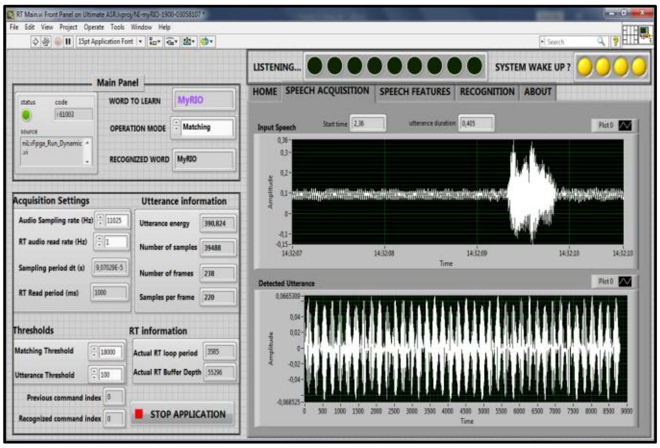

Figure 7: The developed graphical user interface

3. Results and Discussion

3.1. The experimental setup

The experimental setup in Figure 8. In order to test the commands issues by the MyRIO for motion control of a mobile robot, LEDs are used as indicators for each action. The speech signal entry is carried out using a close-talking microphone containing the effect of room acoustics and the background noise. The speech collected was sampled at 11025 Hz . In order to reduce the computational burden in the RT processor and to have a small DMA FIFO size for audio streaming, only a single audio channel is considered and send from the FPGA to the RT for processing and recognition.

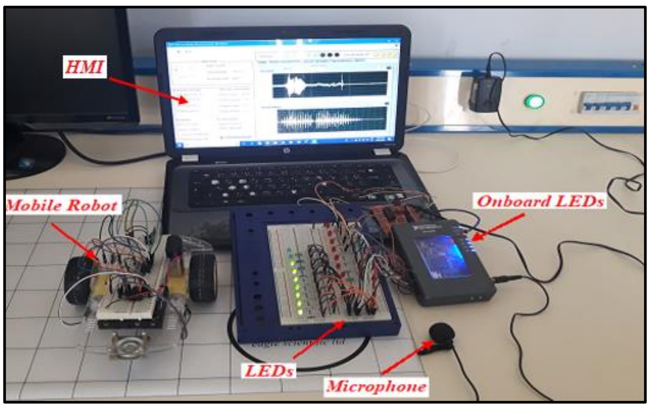

Figure 8: The experimental test setup

3.2. Learning and threshold setting

The system is first trained to learn the commands from the intended user by selecting the learning mode from the user interface as shown in Figure 9. the scrolling green LEDs show that the system is ready to accept the words. The created word dictionary is composed of 8-word patterns for ten sets, where each word pattern represents a matrix containing its feature vectors.

To set an appropriate matching threshold, a dictionary is first created, then the commands are uttered 50 time for each command and at each time the uttered command is matched with all the words within the dictionary, this allows as to see the variation of the average DTW cost between all different words. As a sample, the matching results are collected and plotted in figures 9, 10, 11 for the commands MyRIO, Forward, Barckward respectively. We see that indeed, the minimal DTW cost corresponds to the same word for all commands. For the decision step, it is so important to choose an appropriate threshold. Based on this we have chosen 18000 as a matching threshold, and for the selectivity (difference with other costs) a value of 3000 is chosen.

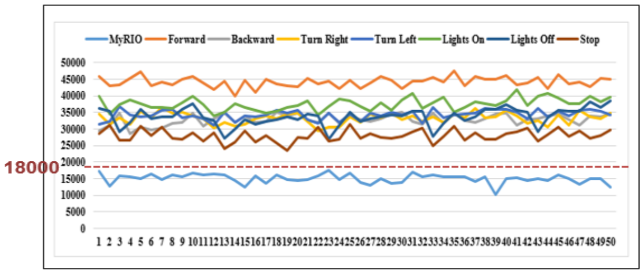

Figure 9 Distance between myRIO word and the other works

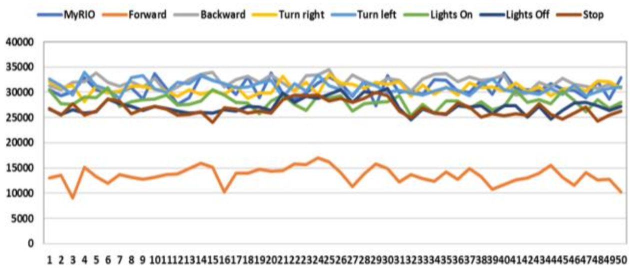

Figure 10 Distance between Forward word and the other works

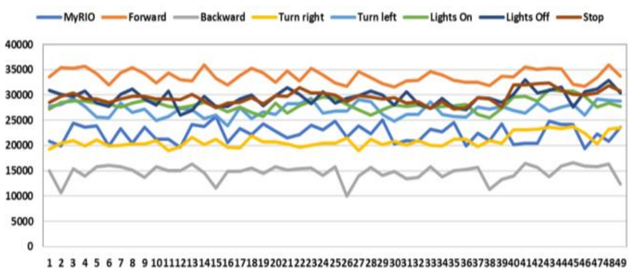

Figure 11 Distance between Backward word and the other works

3.3. Tests for recognition accuracy

To test the accuracy of the developed application, we have carried out an experimental test for the one speaker. 200 trials were carried out for each command, which makes a total of 1600 tests. The results are shown in table 1. The recognition accuracy (RA) is computed as follows:

$$\text{RA}=\left[\frac{\text{\# of times word recognized}}{\text{\# of total trials for that word}}\times 100\right]\% \quad (13)$$

We see from the table that with the chosen thresholds, an excellent recognition accuracy for the same speaker was achieved which is 99.75%. For three different users, the RA for each command is shown in table 1. It can be clearly seen, that the accuracy is reduced, this is natural as the algorithm is customized for a given speaker.

Table 3 Recognition accuracy for a single and many speakers

Word (command)	RA for one speaker	RA for 3 different speakers
myRIO	100%	93%
Forward	100%	99%
Backward	100%	99%
Turn right	99.5	100%
Turn left	99%	100%
Lights on	100%	100%
Lights off	100%	98%
Stop	99.5%	92%
Average (RA)	99.75	96.375%

3.4. Tests for the control logic

In order to test the whole operation of the ASR based voice control system, a sequence of commands are issued randomly and the behavior of the system is recorded online. A summary for the behavior with random commands set is shown in the following table. Notice that the system was initial sleep. So it will not execute any command until the wake up word "MyRIO" is recognized.

Table 4 The response of the system for some commands

Command	Previous word	Matched word	Executed word
myRIO	No command	myRIO	Wake up
Turn left	myRIO	Turn left	Turn left
Lights on	Turn left	Lights on	Lights on
Stop	Lights on	Stop	Stop
Forward	Stop	Forward	No command

3.5. Discussion

The experimental results were excellent, the RA was high especially for one speaker. The developed application has a customized set of commands that can be easily updated if other commands need to be introduced. It can be easily adapted to a new user, during the system operation. In addition, the system has a very high RA for a single user, which is pretty attractive given that many voice-controlled systems are customized to undertake commands from only their user (smart car, domestic robots, etc.). For the limitations, the RA reduces when a new user is tested using a predefined dictionary, this can be mitigated using a hybrid dictionary that includes word templates from many users. The DTW method, yet efficient for single word small vocabulary recognition, it fails in front of HMM and ANN based techniques. In addition, when the set of words is increased, the dictionary size increases considerably, therefore it may pose a memory problem and may cause computational burden that will affect the real time performance and responsiveness of the system. This problem can be mitigated by introducing the shared variable concept, and create the dictionary (or dictionaries if there are many users), into a PC Main program that will be stored and executed into the host PC to reduce the burden on the RT.

Conclusion

In this work, an ASR voice control system was designed and implemented using the NI-myRIO-1900 and LabVIEW. The MFCC feature extraction and the DTW pattern matching methods were used. The proposed algorithm has achieved a 99.75% of accuracy for a single speaker, and an accuracy of 96.375% for different speakers. The developped voice control system has many attractive features, it has high RA, it is customizable to different commands and it can be embedded in any application. the following suggestions can be considered for future work: Take advantage of the audio output capability of the NI MyRIO-1900 module to implement a speech to text alongside the voice control system, to make it capable to respond vocally to the user. This is interesting as modern voice assistants. The implementation of a continuous speech system using the ANNs, GMM or HMM for the development of a real-time application with the FPGA support in the NI MyRIO-1900 embedded module will be quite challenging.

References

[1]. T. R. Hill, Audio and speech processing with Matlab, CRC Press, 2019.

[2]. I. V. McLoughlin, Speech and Audio Processing, Cambridge University Press, 2016.

[3]. A. Mousa, "MareText Independent Speaker Identification," International Journal on Electrical Engineering and Informatics – Vol. 3, No. 1, pp. 100-108, 2011.

[4]. W. R. Kusuma, and P. B. Guhyapati., "Simulation of voice recognition system for controlling robotic applications," Journal of Theoretical and Applied Information Technology, vol. 39, no. 2, pp. 188-196, 2012.

[5]. S. Pleshkov, Z. Zahariev, and A. Bekiarski, "Development of speech recognition algorithm and LabVIEW model for voice command control of mobile robot motion," In: International Conference on High Technology for Sustainable Development, Sofia, Bulgaria, 2018.

[6]. K. Darabkh, A. Khalifeh, I. Jafar, A. Bathech and S. Sabah, "A yet efficient communication system with hearing-impaired people based on isolated words for Arabic language", IAENG International Journal of Computer Science, Vol. 40, No. 3, pp. 183-192 , 2013.

[7]. M. A. Al-Alaoui, L. Al-Kanj, J. Azar, and E. Yaacoub, "Speech recognition using artificial neural networks and hidden markov models," IEEE Multidisiplinary Engineering Education Magazine, Vol. 3, N. 3, 2008.

[8]. S. Sunny, P. S. David, and K. P. Jacob, "Recognition of speech signals: An experimental comparaison of linear predictive coding and discrete wavelet transforms", International Journal of Engineering Science, Vol. 4, No. 04, pp. 1594-1601, 2012.

[9]. D. H. Galatang, and Suyauto, "Syllable-based indosesian automatic speech recognition", International Journal on Electrical Engineering and Informatics, Vol. 12, No. 4, pp. 720-728, 2020.

[10]. G. Harshita, and G. Divya, "LPC and LPCC method of feature extraction in speech recognition system," In: 6th International Conference-Cloud System and Big Data Engineering (IEEE), pp. 498-502, 2016.

[11]. A. Lipeika, J. Lipeikien, and L. Telksnys, "Development of isolated word speech recognition system," Informatica, Institute of Mathematics and Informatics, Vilnius. Vol. 13, No. 1, pp. 37-46, 2002.

[12]. Q. H. Nyugen, and T. D. Gao, "A novel method for recognizing vietnamese voice commands on smartphones with support vector machine and convolutional neural networks", Wireless Communications and Mobile Computing, Vol. 2020, Article ID 2312908, 9 pages, 2020.

[13]. H. M. Mohammed, S. Alkassab, H. R. Mohammed et al. "Speech recognition system with different methods of feature extraction," International Journal of Innovative Research in Computer and Communication Engineering, Vol. 6, No. 3, pp. 1-10, 2018.

[14]. K. P. Bharath, and K. Rajesh, "Multitaper based MFCC feature extraction for robust speacker recognition system", In: 2019 Innovations in Power and Advanced Computing Technologies (i-PACT). IEEE, 10.1109/i-PACT44901, 2019.

[15]. National Instruments Inc, "NI myRIO-1900, User Guide and Specifications", available online at: https://www.ni.com/pdf/manuals/376047.pdf. Accessed on 02/17/2021.