Proceeding of the Electrical Engineering Computer Science and Informatics
Vol. 7, October 2020 **ISSN: 2407-439X**

Speech Recognition Implementation using MFCC and DTW Algorithm for Home Automation

Abdulloh Salahul Haq
School of Electrical Engineering
Telkom University
Bandung, Indonesia
abdullohsh@student.telkomuniversity.ac.id

Muhammad Nasrun
School of Electrical Engineering
Telkom University
Bandung, Indonesia
nasrun@telkomuniversity.ac.id

Casi Setianingsih
School of Electrical Engineering
Telkom University
Bandung, Indonesia
setiacasie@telkomuniversity.ac.id

Muhammad Ary Murti
School of Electrical Engineering
Telkom University
Bandung, Indonesia
arymurti@telkomuniversity.ac.id

***Abstract*—The use of speech recognition as part of home automation, especially for smart homes, is an exciting thing that is still being developed. That is because of human needs for comfort, convenience, quality of life, and better safety. Speech recognition built in this study is used as a device to control smart home devices by identifying the commands spoken by users, especially in a state of clean speech. The command used is a predetermined consecutive word. For the extraction of voice commands, the MFCC algorithm is used to match spoken words with templates using the Dynamic Time Warping (DTW) algorithm. DTW algorithm can find the difference between 2-time series that have different lengths of time. The results of the accuracy of this system by using these algorithms were successfully carried out by 86.67%, with an average time required to identify the commands of 5.28 seconds.**

Keywords—home automation, speech recognition, MFCC, DTW

I. INTRODUCTION

Home automation is used daily to provide comfort, convenience, more quality of life, and even security for the owners [1]. Automated homes are types of houses that usually have IoT technology, whether it's internet networks, personal computers, smartphones, or other devices connected to the internet network so that some features in the home can be controlled anywhere and anytime [2]. Automation is a word that is often spelled in the electronic field that has brought many revolutions in today's technology [3]. The term Internet of Things (IoT) is a concept that has the purpose of making humans and physical objects in their environment a part of the internet that is connected [2]. Home automation itself is designed and developed into one or more controllers that can perform various basic tasks to home electronic devices connected to the internet, such as electrical switches, light sensors, temperature sensors, smoke detectors, etc [1].

Speech recognition or Automatic Speech Recognition (ASR) is one of the systems used in home automation because, with a unique human voice, a user can give commands to the system to do what they are told. Doing what you command accurately, and doing it fast is a different matter. Because doing what is commanded accurately requires a sound feature extraction process that is good and doing what is ordered quickly, then computation on feature extraction must be rapidly done accompanied by a fast matching process. Speech recognition must also be able to understand successive speeches or sayings between words that have a sufficient time lag so that the time interval can result in the identification of input commands. From several ASR applications that have been implemented, several algorithms are used such as LPC (Linear Predictive Codes), MFCC (Mel Frequency Cepstral Coefficients), PLP (Perceptual Linear Prediction) or PLP-RASTA (PLP-Relative Spectra) for feature extraction from acoustic signals, and algorithms such as DTW (Dynamic Time Warping), HMM (Hidden Markov Model), MLP (Multi-Layer Perceptron), SVM (Support Vector Machine), and DT (Decision Trees) for matching or matching processes with templates [4] – [6].

In this research, the design and implementation of the Speech Recognition system, which is used in home automation, is used by using acoustic signals (in this case in the form of human voice commands) as inputs and producing output in the format of word sequences in the system [7]. The feature extraction will be performed on the voice command using the MFCC algorithm and matching the voice command data on the template using the DTW algorithm. Using this voice command, the system will turn on or turn off electronic devices.

II. RESEARCH METHOD

A. Design of System

The following is an overview of the system. In Fig. 1, a user inputs by saying commands to the microphone, then the microphone will capture the command as an acoustic signal-processing devices for Speech Recognition process acoustic signals. After the input signal is processed, extraction of the features possessed by the signal will be carried out so that matching processes can be carried out with data that has been trained (trained data/template), the speech recognition system integrated into the microprocessor device in the form of raspberry pi, will send a command to the NodeMCU device by using one of the M2M communication methods, so that the NodeMCU, in this case, functions as a switch can turn on or turn off electronic devices. In this system, NodeMCU or ESP8266 is the receiver of the speech recognition system.

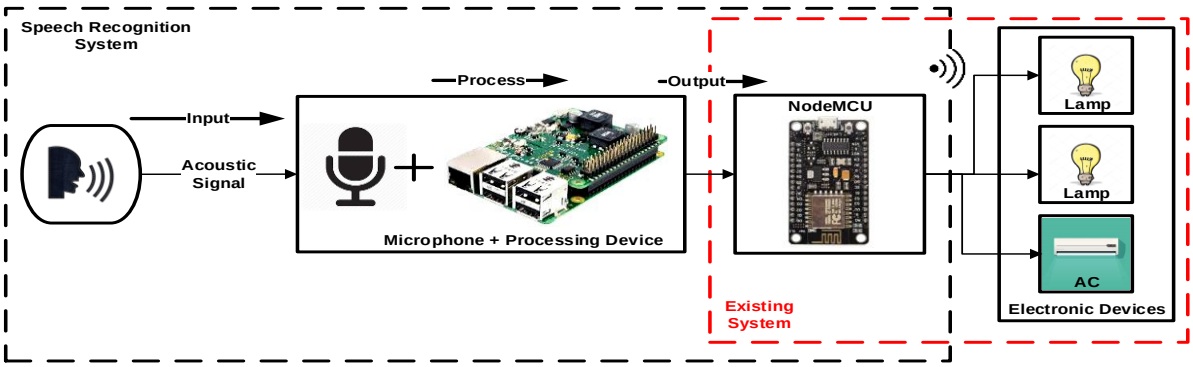


Fig. 1. Overview of the system.

B. *Speech Recognition*

Speech Recognition is a process of identifying sounds based on words spoken by converting an acoustic signal captured by an audio device. Speech recognition is also a system used to recognize word commands from human voices and then translate into data understood by computers—the following block diagram of how speech recognition works. The basic concept of this system's work is to receive acoustic signal input and produce a series of words [8].

The discussion and research and development regarding speech recognition or Automatic Speech Recognition (ASR) is still an exciting topic in its role in everyday life. Optimizing the quality and accuracy of ASR is still in the process, such as increasing speech recognition capabilities or reducing existing noise [4].

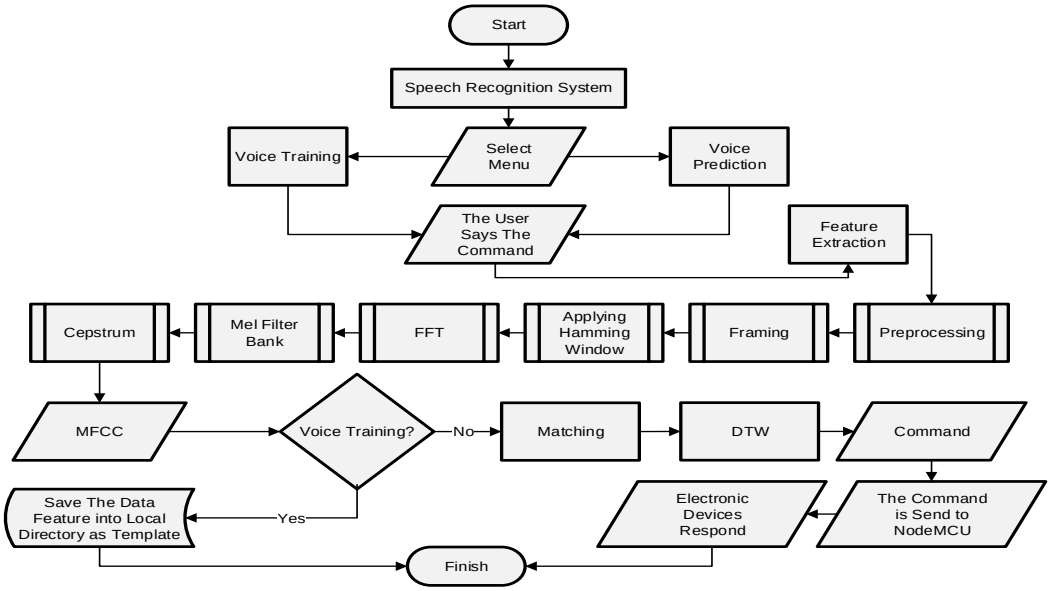


Fig. 2. Flowchart of the speech recognition system.

In Fig. 2, the system receives input from the user to select a menu on the system, which is the "Voice Training" or "Voice Prediction" menu. Then the user says the voice command, and the system records the command. The command that was successfully registered will be performed a feature extraction process using the MFCC algorithm. Then proceed to the processes owned by the MFCC algorithm; those are Preprocessing, Framing, Applying Hamming Window, FFT, Mel Filter Bank, Cepstrum, and MFCC as a process for extracting features that have sound. Furthermore, for the "Voice Training" menu, then the data will be stored in the as a template of the commands spoken by the user, and for the "Voice Prediction" menu, the matching process will be carried out on the data with the data in the database using the DTW algorithm. When the matching process is complete, the voice data (which is a command) will be sent to the NodeMCU device to be forwarded to the electronic device.

C. *Mel Frequency Cepstral Coefficients (MFCC) Algorithm*

Mel Frequency Cepstral Coefficients (MFCC) is a concept for extracting features possessed by an acoustic signal. Based on human hearing that cannot sense frequencies of more than 1KHz. Calculations carried out by MFCC are very dependent on the process of changing signals from analog to digital. MFCC performs calculations ranging from the length of the wave height, noise, and other things so that the words that are adequately spoken by the user are obtained [9].

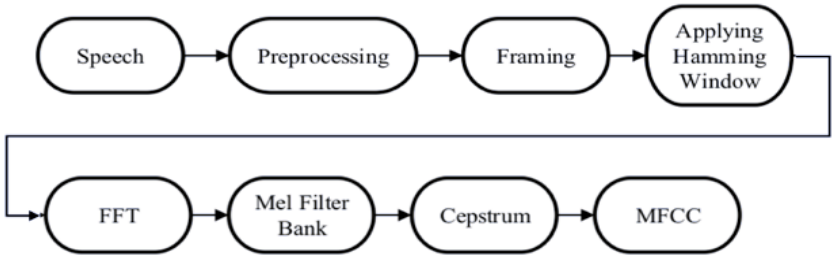


Fig. 3. MFCC flow diagram.

The first step in the speech recognition process is preprocessing, which is a process for filtering background noise and existing noise. Then the framing is done by changing the time unit t on the analog signal to n on the digital signal so that it has the function x(n), then the signal is divided into several frames for as long as the word is spoken. The Hamming window is applied to help reduce discontinuity at the ends of each frame, then Fast Fourier Transform (FFT) is performed to study the characteristics of the x(n) function in the frequency domain. Next is Mel Filter Bank, which is by converting frequencies to the mel scale to connect the perceived frequency, tone, pure tone, and actual frequency measured [10]. This work's scale of mel is to map the scale between linear frequencies of speech signals to a logarithmic scale for frequencies higher than 1 kHz [11].

D. Dynamic Time Warping (DTW) Algorithm

Dynamic Time Warping (DTW) is a concept to match data extraction of features performed by the MFCC on the input command (data testing) with the data in the template or training data. DTW will measure the distance and curvature of the two sound signals [9]. DTW is generally used to measure the similarity between two-time series signals, which can have variations in time and speed [12] even though the two signals have different signal lengths. The output of DTW is the distance value in scalar quantity [10]. This value represents how similar two sound signals are in time series.

DTW also has a function as a solution in aligning the time in successively spoken words [13]. This affects the correction of the utterance results to the final output because successive statements that do not have pauses or that have fewer pauses will result in errors in the identification of the speech [9].

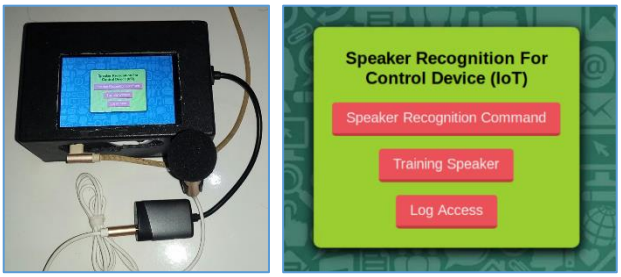
E. Machine to Machine (M2M) Communication

Machine to Machine Communication (M2M) is a form of communication between hardware (devices) using wireless signals to be able to connect and communicate without human assistance in real-time [14]. One of the methods in this communication is MQTT (Message Queuing Telemetry Transport), which is a protocol that works at the ISO TCP/IP layer, on the internet of things (IoT) system [15]. This protocol works with a server that sends messages to the client, and the client receives the message by minimizing data encoding and data decoding and only requires a small amount of memory to deliver the message [16]. In this system, three control signals from MQTT are used to deliver and receive data from the server to the client.

1. Connect: the control signal to make a connection to the server.
2. Subscribe: the control signal to get a message with a particular topic.
3. Publish: the control signal to send a message with a particular topic.

III. Result and Discussion

In this research, speech recognition devices built based on hardware requirements in Table I can be seen in Fig. 4 sections (a), and the user interface used on this device can be seen in other sections of Fig. 4.



(a) (b)

Fig. 4. (a) Speech recognition system device (b) Main view of the user interface.

In Fig. 4, section (b) is the prominent display of the user interface, which is used to choose between 3 menus, Recognition or prediction of voice commands, voice training, and log access from the user.

Tests conducted in this research are to identify voice input commands with voice data training that has been performed feature extraction so that if there is a match between voice input commands with voice training commands, the system will produce output in the form of commands and send them to the MQTT client to control devices that are contained in the smart home.

The training data used is the result of feature extraction from voice recordings from several commands that have been set for use in speech recognition and control systems to the NodeMCU device. Here are some commands that will be used in this system.

TABLE I. List of Command Voice Data and MQTT Message

Command	Content of The Command	Amount of Data	MQTT Message	Description
1	"Nyalakan perangkat satu"	5	ON1	To turn on the device 1
2	"Nyalakan perangkat dua"	5	ON2	To turn on the device 2
3	"Nyalakan perangkat tiga"	5	ON3	To turn on the device 3
4	"Matikan perangkat satu"	5	OFF1	To turn off the device 1
5	"Matikan perangkat dua"	5	OFF2	To turn off the device 2
6	"Matikan perangkat tiga"	5	OFF3	To turn off the device 3

In the Content of The Command in Table I is a command utterance in Indonesian which has a function as written in the Description, that has an output to send commands to control on/off three different electronic devices.

A. Voice Prediction

As for the "Sound Prediction" menu, the recording system used is streaming audio that works continuously listening to the sound that enters the microphone device using certain limits or rules on the system to record the audio streaming system's recorded sound. The rule is the recording process will be carried out when the detected sound intensity ≥ 35 and has a duration of 1.3 to 2 seconds only. If these rules are met, the voice command recording process will be carried out.

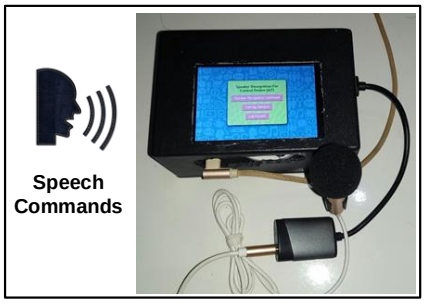


Fig. 5. Input the voice command into a speech recognition system device.

B. Preprocessing

At this stage of pre-emphasis, that is the process of filtering sound signals to be clearer when heading to the next process [17]-[19]. In the pre-emphasis stage, filtering is performed on the sound file signal using a filter coefficient of 0.097. The results of this process can be seen in Figure 6 section (b).

$$y(n) = x(n) - (x(n-1) \times \delta) \quad (1)$$

Where, x(n) = input signal value before the pre-emphasis process, y(n) = pre-emphasis output signal results, and δ = filter coefficient, by default the value is 0.095 or 0.097.

C. Framing and Applying Hamming Window

The next subprocess is framing. This process makes the sound signal into several sample N frames. The first frame consists of the first sample N [19]. The second frame consists of M samples after the first frame and overlaps with N-M samples. The third frame consists of 2M samples after the first frame (or M samples after the second frame) and overlaps with N-2M samples. This process continues until all sound signals (speech signals) are counted into one or more frames. The length of the commonly used overlap area ranges from 30% to 50% of each frame's size.

In this subprocess, windowing is performed on each frame formed in the framing subprocess. This windowing process serves to prevent aliasing or abnormal conditions in the signal [20]. The results of this process can be seen in (c) section of Fig. 6. The following is the formula for using the window function in the input signal.

$$x(n) = x_i(n) \times w(n) \quad (2)$$

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad (3)$$

Where, n = 0, 1, 2, … , N-1
x(n) = signal sample value
x_i (n) = sample value from frame to [i] signal
w(n) = window function, in the hamming window.

D. FFT (Fast Fourier Transform)

In this subprocess, each frame will be changed from the time domain to the frequency domain. This process will produce a frequency spectrum that is used to simplify computation and analysis [21].

$$D_k = \sum_{m=0}^{N_{m-1}} D_m \times e^{\frac{-j2\pi km}{N_m}}, \text{ Where k = 0, 1, 2, 3, …, } N_{m-1} \quad (4)$$

E. Mel Filter Bank

At this stage, the filtering of certain frequency band sizes in sound signals is carried out. The size used is called the Mel frequency scale, which is a scale that is linear at frequencies below 1kHz and is logarithmic at frequencies above 1kHz. The equation used is as follows [10], [11], [17].

$$\text{mel}(f) = 2595 \times \ln\left(1 + \frac{f}{700}\right) \quad (5)$$

Where, mel(f) = mel frequency scale
f = linear frequency.

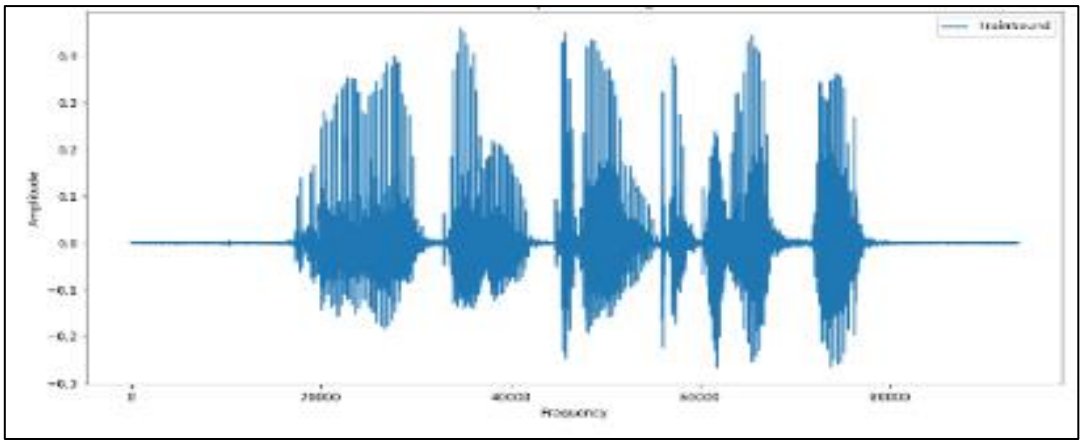
F. Cepstrum and MFCC (Mel Frequency Cepstral Coefficients)

The final process of MFCC is to change the log mel spectrum to the time domain. The result is called Mel-frequency cepstral coefficients [17]. That is a cepstral representation of the properties of the speech signal in a known frame. The equation used is as follows.

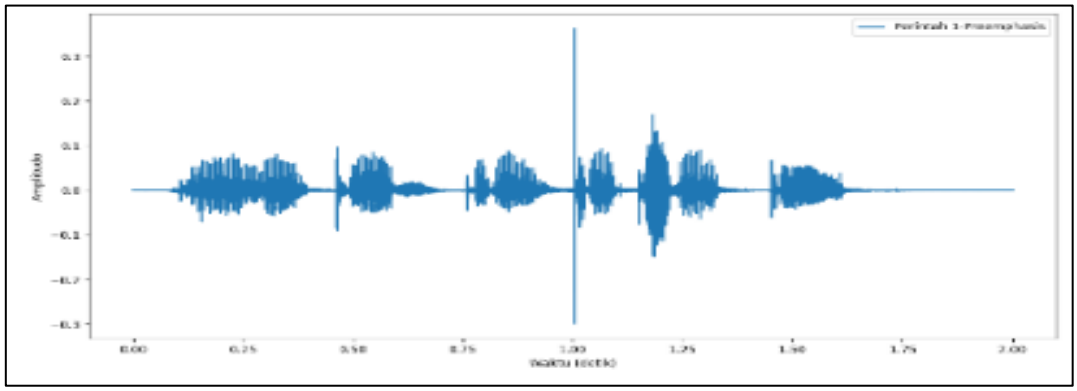
$$C_n = \sum_{k=1}^{K} (\log S_k) cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right] \quad (6)$$

Where, n = 1, 2, 3, …, K
C_n = coefficient of cepstrum mel frequency
S_k = Mel power coefficient.

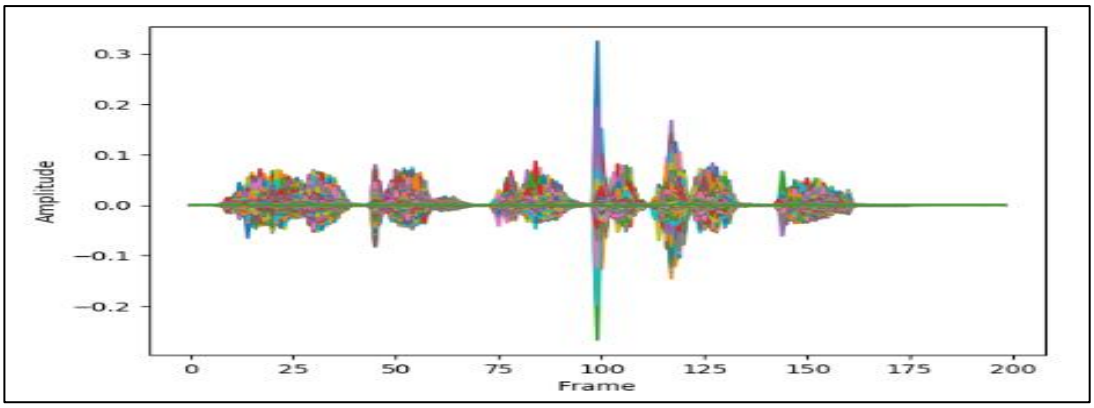
In a speech recognition system, usually, only the first 13 cepstrum coefficients are used [9], [12]. The extraction features are traditionally extracted in multiple time windows [22].



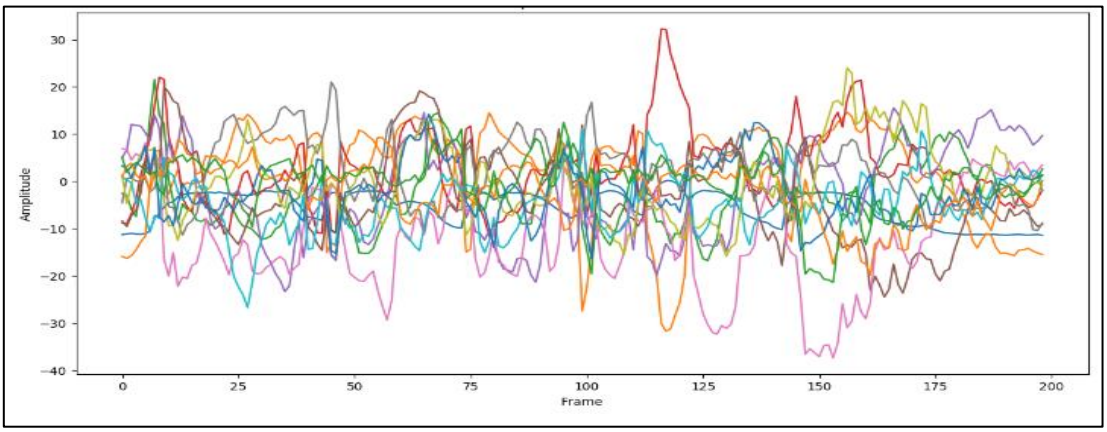
(a)



(b)



(c)



(d)

```
    6.50328857e-02   8.72959951e-01   3.87341988e-01  -5.30685536e-01
   -2.97278196e-01]
 [ -1.72707421e+00  -1.54031985e+00   2.20016365e-01  -9.91131021e-01
    9.83201222e-01  -5.31244268e-01   1.22764289e+00  -1.88782162e+00
   -5.42002094e-02   1.14301197e+00   3.52758649e-01  -6.29110116e-01
    1.46438411e-01]
 [ -1.70703695e+00  -1.58071413e+00   3.06902835e-01  -7.80428696e-01
    1.06278137e+00  -6.59284288e-01   1.36918010e+00  -1.91185110e+00
   -7.83855116e-02   1.05820970e+00   4.72298332e-01  -3.23945683e-01
    1.81237097e-01]
 [ -1.75647200e+00  -1.60767066e+00   4.88493751e-01  -5.73720645e-01
    1.17646663e+00  -6.17335883e-01   1.47463978e+00  -1.67302572e+00
   -1.37142584e-01   4.54635978e-01   5.05385759e-01   3.16751805e-01
    5.21013063e-01]]

('Number of Frame : ', 199)
```

(e)

Fig. 6. (a) Waveform from command 1 "Nyalakan perangkat Satu" (b) Command 1 after pre-emphasis was applied (c) Framing and applying Hammingwindow (d) Cepstrum in the form of spectrum (e) Cepstrum in the form of an array of 13 coefficient values with 199 multiple frames.

G. DTW (Dynamic Time Warping) and Matching

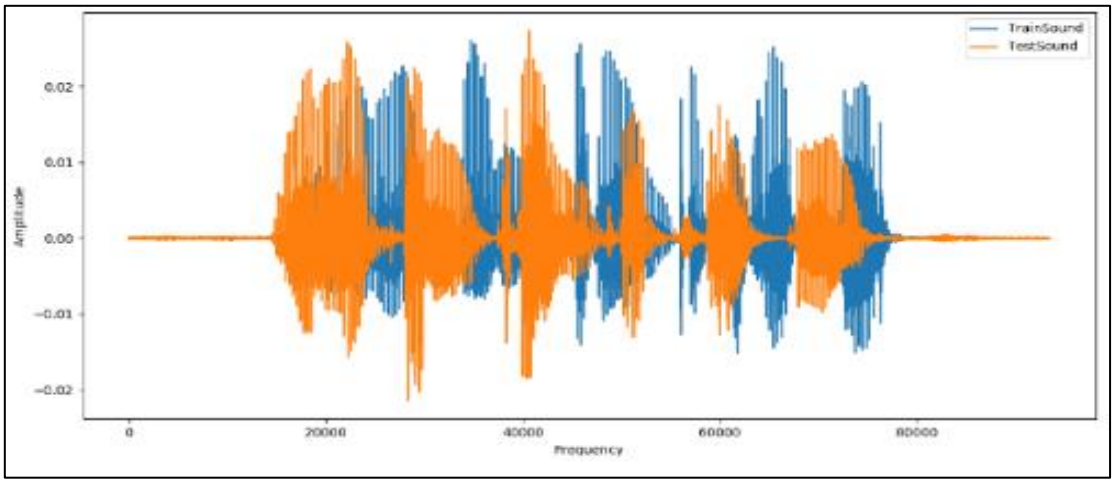


Fig. 7. Command 1 as input (TestSound: Orange) and template (TrainSound: Blue).

After the feature extraction process is performed on the two voice commands 1 in Fig.7 above, DTW performs the matching process by finding a similarity value called the distance between the signal characteristics. The equation used is as follows [9], [10], [23], [24].

$$\text{Dist}(x,y) = \text{Dist}(x,y) + \min\begin{cases} Dist(x-1,y) \\ Dist(x,y-1) \\ Dist(x-1,y-1) \end{cases} \quad (7)$$

$$\text{Dist}(d) = \sum_{k=0}^{k=L} Dist(d_{k_x}, d_{k_y}) \quad (8)$$

Where, L= 1,2, …, N
Dist(x,y) = path between time series x and y
Dist(d) = distance value between time series x and y
L = the length of the path formed from Dist(x,y).

H. Testing of Testing Data Identification with Training Data

In Table II, testing will be done to identify input commands by matching testing data with training data (templates)—testing data, which as input command is matched by distance similarity with training data in the system.

TABLE II. IDENTIFICATION OF INPUT COMMANDS

Command	Number of Tests Performed	Number of Successful Tests
1	5	4
2	5	4
3	5	5
4	5	4
5	5	5
6	5	4
Sum	30	26

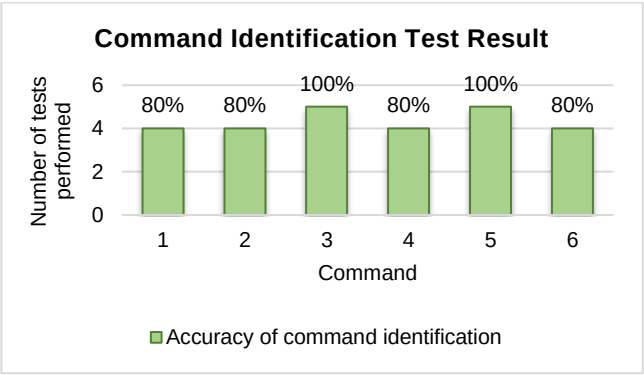


Fig. 8. Command identification test result.

The number of tests is five times for each command because there are five different speaker sources. Figure 8 displayed the results of testing the identification of input commands several six various commands. Tests with the highest success were found in the 3rd and 5th orders identified as many as five times out of a total of 5 tests.

Feature extraction with the MFCC method has been used in processing voice signals in previous studies, but using the HMM classification algorithm, and the results obtained are also good as in this study [25].

I. Testing of Response Time of Matching Testing with Data Testing

In this test, the calculation of how long it takes for testing data (input command) to be matched with training data is calculated to produce the expected output.

TABLE III. RESPONSE TIME TEST RESULT

Command	Response Time (second)					Average Time
	1	2	3	4	5	
1	5.45	5.35	5.76	5.21	5.21	5.69
2	5.71	5.44	5.43	5.15	5.68	5.48
3	5.95	5.08	5.44	5.00	5.82	5.45
4	5.75	5.50	5.86	5.40	6.39	5.78
5	5.64	5.29	5.33	5.03	5.66	5.39
6	5.56	5.11	5.16	5.06	5.54	5.28
Average Time of Overall Testing						5.51

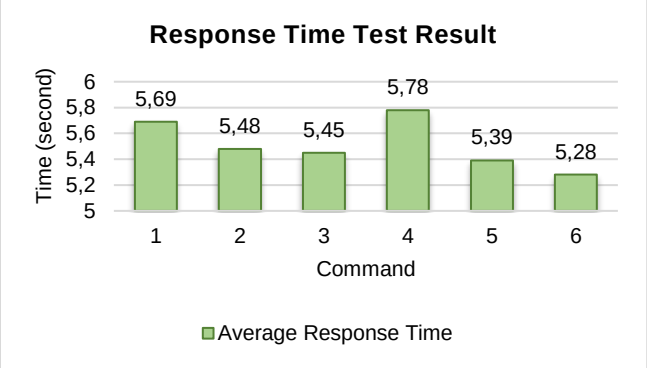


Fig. 9. Response time test result.

From Fig. 9, it can be concluded that the average response time of the best input command, contained in the 6th input command with an average time of 5.28 seconds. The average time needed for the whole matching process is 5.51 seconds.

J. Testing of Training Data Identification Testing with Test Data Based on Sound Intensity

The following test results identify training data with testing data based on the sound intensity of the identification process. In Fig. 10, identification testing based on sound intensity is carried out on three categories, namely in a situation <35 dB, 45-59 dB, and 60-67 dB. In the <35 dB category, the test was not successful because the noise level was at that level; in other words, the input command voice spoke at the noise level and made the identification process unsuccessful.

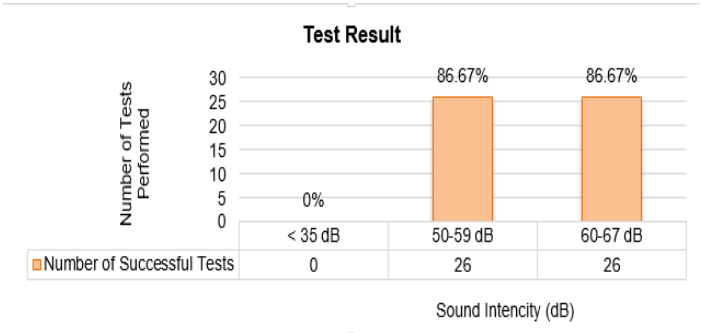


Fig. 10. The test result of command identification based on sound intensity.

IV. CONCLUSION

Speech Recognition using MFCC and DTW algorithms in this research resulted in an 86.67% accuracy rate, 30 tests. Response Time in the best input command, there is a 6th input command with an average time of 5.28 seconds. The average response time for the whole matching process is 5.51 seconds. This identification test in the speech recognition process was successfully carried out at the input sound intensity at> 45 dB.

REFERENCES

[1] Asadullah M, Raza A. *An overview of home automation systems*. 2016 2nd International Conference on Robotics and Artificial Intelligence (ICRAI). Rawalpindi. 2016; (December 2016): 27–31.

[2] Wibowo FW, Hidayat F. A low-cost home automation system based-on the internet of things. *Journal of Telecommunication, Electronic and Computer Engineering*. 2017; 9(2–4): 155–9.

[3] Baig I, Muzamil C, Dalvi S, Campus KT. Home Automation Using Arduino Wifi Module Esp8266. [Panvel]: Anjuman-I-Islam's Kalsekar Technical Campus; 2016.

[4] Kurniadhani B, Hadiyoso S, Aulia S, Magdalena R. FPGA-based implementation of speech recognition for automobile robot control using MFCC algorithm. *TELKOMNIKA*. 2019; 17(2): 125–132.

[5] Rossi M, Benatti S, Farella E, Benini L. *Hybrid EMG classifier based on HMM and SVM for hand gesture recognition in prosthetics*. Proceedings of the IEEE International Conference on Industrial Technology. Sevilla. 2015: 1700–1705.

[6] Vimala C, Radha V. Isolated speech recognition system for Tamil language using statistical pattern matching and machine learning techniques. *Journal of Engineering Science and Technology*. 2015; 10(5): 617–632.

[7] Jerfia GN, Vaishnavi J, Karthick P. CENTRALIZED SMART HOME AUTOMATION THROUGH VOICE RECOGNITION. *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)*. 2015; 13(1): 340–343.

[8] Jurafsky D, Martin JH. Speech and Language Processing 18 BT - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3rd ed. Upper Saddle River: Prentice Hall. 2009: 438.

[9] Mohan BJ, Ramesh Babu N. *Speech recognition using MFCC and DTW*. 2014 International Conference on Advances in Electrical Engineering, ICAEE 2014. Vellore. 2014: 1–4.

[10] Awad A, Omar H, Ahmed Y, Farghaly Y. Speech Recognition System Using MFCC and DTW. 2016; (December 2016): 4.

[11] Ittichaichareon C, Suksri S, Yingthawornsuk T. *Speech recognition using MFCC*. PSRC - Planet Sci Res Cent Proceeding. Pattaya. 2012; (July 2012): 135–138.

[12] Muda L, Begam M, Elamvazuthi I. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *JOURNAL OF COMPUTING*. 2010; 2(3): 138–43.

[13] Dhingra SD, Nijhawan G, Pandit P. ISOLATED SPEECH RECOGNITION USING MFCC AND DTW. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*. 2013; 2(8): 4085–4092.

[14] Elgazzar MH. *Perspectives on M2M protocols A comparative study between different M2M protocols*. 2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS). Cairo. 2015: 501–505.

[15] J. Bellido-Outeirino F, M. Flores-Arias J, J. Palacios-Garcia E, Pallares-Lopez V, Matabuena-Gomez-Limon D. *M2M Home Data Interoperable Management System Based on MQTT*. 2017 IEEE 7th International Conference on Consumer Electronics - Berlin (ICCE-Berlin). Berlin. 2017: 200–202.

[16] Bryce R, Shaw T, Srivastava G. *MQTT-G : A Publish / Subscribe Protocol with Geolocation*. 2018 41st International Conference on Telecommunications and Signal Processing (TSP). Athens. 2018: 1–4.

[17] Muhammad HZ, Nasrun M, Setianingsih C, Murti MA. *Speech Recognition for English to Indonesian Translator Using Hidden Markov Model*. International Conference on Signals and Systems. Bali. 2018: 255–260.

[18] Jing Z, Min Z. *Speech Recognition System Based Improved DTW Algorithm*. International Conference on Computer, Mechatronics, Control and Electronic Engineering. Changchun . 2010; 6: 320–323.

[19] Manurung DB, Dirgantoro B, Setianingsih C. *Speaker Recognition For Digital Forensic Audio Analysis Using Learning Vector Quantization Method*. 2018 IEEE International Conference on Internet of Things and Intelligence System (IOTAIS). Bali. 2018: 221–226.

[20] Wan C, Liu L. *Research of Speech Emotion Recognition Based on Embedded System*. The 5th International Conference on Computer Science & Education. Hefei. 2010: 1129–1133.

[21] Dixit A, Vidwans A, Sharma P. *Improved MFCC and LPC algorithm for bundelkhandi isolated digit speech recognition*. International Conference on Electrical, Electronics, and Optimization Techniques, ICEEOT 2016. Chennai. 2016: 3755–3759.

[22] Bernal-Ruiz C, Garcia-Tapias FE, Martin-del-Brio B, Bono-Nuez A, Medrano-Marques NJ. *Microcontroller Implementation of a Voice Command Recognition System for Human-Machine Interface in Embedded Systems*. 2005 IEEE Conference on Emerging Technologies and Factory Automation. Catania. 2005; 1(January): 587–591.

[23] Li J, Zheng LM, Yang L, Tian LJ, Wu P, Zhu H. *Improved dynamic time warping algorithm the research and application of query by humming*. Proceedings - 2010 6th International Conference on Natural Computation, ICNC 2010. Yantai. 2010; 7: 3349–3353.

[24] Pandey D, Singh KK. Implementation of DTW algorithm for voice recognition using VHDL. Proceedings of the International Conference on Inventive Systems and Control, ICISC 2017. Coimbatore. 2017: 1–4.

[25] Alifani, F., Purboyo, T.W. and Setianingsih, C., 2019, August. Implementation of Voice Recognition in Disaster Victim Detection Using Hidden Markov Model (HMM) Method. In 2019 International Seminar on Intelligent Technology and Its Applications (ISITIA) (pp. 445-450). IEEE.