IOP Conf. Series: Materials Science and Engineering **288** (2018) 012042 doi:10.1088/1757-899X/288/1/012042

The Implementation of Speech Recognition using Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machine (SVM) method based on Python to Control Robot Arm

D Anggraeni[1,2*], W S M Sanjaya[1,2], M Y S Nurasyidiek[1,2] and M Munawwaroh[1,2]

[1]Department of Physics, Faculty of Science and Technology, Universitas Islam Negeri Sunan Gunung Djati Bandung, Indonesia
[2]Bolabot Techno Robotic Institute, CV. Sanjaya Star Group, Bandung, Indonesia

[*]tsugumikaoru@gmail.com

Abstract. In this paper describe an implementation of speech recognition to pick and place an object using Robot Arm. To get the feature extraction of speech signal used Mel-Frequency Cepstrum Coefficients (MFCC) method and to learn the database of speech recognition used Support Vector Machine (SVM) method, the algorithm based on Python 2.7. The data learning which used to SVM process are 12 features, then the system tested using trained and not trained data show the best agreement to identifying the speech recognition. The speech recognition system has been implemented for control the 5 DoF Robot Arm based Arduino microcontroller to doing task pick and place the object.

1. Introduction

Speech control or usually called as Speech Recognition is the method to controlling something by human voices/speech. This method usually used for robotics system to help disability people or other aim. To develop speech recognition needed a method to identify speech signal, they are; feature extraction and machine learning.

Features extraction of speech used to know the behavior of speech signal are Mel-Frequency Cepstrum Coefficient (MFCC) [1] [2] [3] and Linear Predictive Coding (LPC) [4] [5] method. Artificial Intelligence method to learn and classify the speech, such as; Artificial Neural Networks [1] [5] [6], Fuzzy Logic [7], Support Vector Machine [1], Adaptive Neuro Fuzzy Inference System (ANFIS) [8], Hidden Markov [4], and other soft computing. Speech Recognition can be implemented to any field, for example; control Social Robot [9] [10], industrial [11], control smart home [12], control mobile robot [13], control wheel chair [14] [15], biomatrix [16], control arm robot [17], and other.

In this study will be describe a signal voice processing by using Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machine (SVM) method based on Python 2.7. Finally, the system will be implemented to control 5 Degree of Freedom (DoF) Robot Arm for pick and place an object based on Arduino microcontroller.

The paper is organized as follows. In section 2, described the theoretical background of MFCC and SVM on details. In section 3, describe method and system design. In section 4, described a hardware design of arm robot. In section 5, described application of speech recognition in detail. Finally, in Section 6 the concluding remarks are given.

2. Theoretical background

2.1 Feature extraction using Mel Frequency Cepstrum Coefficient (MFCC) method

Mel Frequency Cepstrum Coefficient (MFCC) is a method of feature extraction of voice signals. Feature extraction is the process of determining a value or vector that can be used as an object or an individual identity. MFCC is the most used method in various areas of voice processing field, because it is considered quite good in representing signal [12].

Feature is the coefficient of cepstral, the coefficient of cepstral used still considering the perception of the human hearing system. The workings of MFCC are based on the different frequencies that can be captured by the human ear so as to represent the sound signals as humans represent them. MFCC process block diagram can be seen in Figure 1.

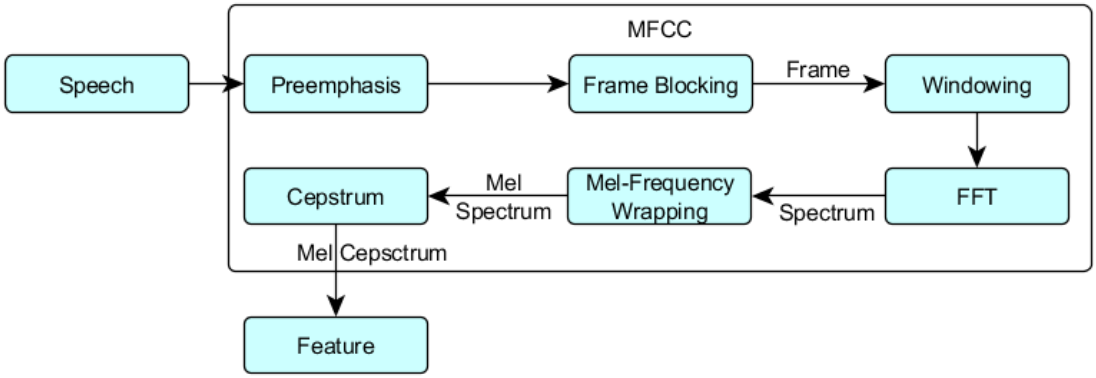


Figure 1. LPC Process

2.1.1 Preemphasis. In the process of speech signal pre-emphasis filter is required after the sampling process. The purpose of this filtering is to obtain a smoother spectral form of speech signal frequency. In other words, this filtering process is done to reduce noise during sound capture. Where the spectral shape is relatively high value for low areas and tends to fall sharply to the area of frequency above 2000 Hz [18]. The pre-emphasis filter is based on the input / output relationship in the time domain expressed in the following equation:

$$y(n) = x(n) - ax(n-1) \quad (1)$$

From Equation 1, a is a pre-emphasis filter constant, it is usually $0.9 < a < 1.0$.

2.1.2 Frame blocking. In this process, the sound signal is segmented into multiple overlapped frames, so there is not a single deletion of signals. This process will continue until all signals have entered into one or more frames as illustrated. Voice analysis was done by short-time analysis. The $x[n]$ long voice signal is divided into a number of frames. One frame has N voice data sample. Between one frame with another frame overlapping each other a number of M samples of voice data. The value of M is not more than N that is $2xM$.

2.1.3 Windowing. Windowing is a process for analyzing long sound signals by taking a sufficiently representative section. Windowing is a Finite Impulse Response (FIR) digital filter approach. This process removes the aliasing signal due to the discontinuity of the signal

pieces. Discontinuities occur due to the frame blocking process. If we define the window as $w(n)$, $0 \leq n \leq N - 1$, where N is the number of samples in each frame, the result of windowing is a signal:

$$y_1(n) = x_1(n)w(n), 0 \leq n \leq N - 1 \quad (2)$$

From Equation 2 $y(n)$ is the result signal of the convolution between the input signal and the window function and $x(n)$ represents the signal to be convolved by the window function. Where $w(n)$ usually uses window Hamming which has the form:

$$w(n) = 0.54 - 0.46.cos\left(\frac{2\pi n}{N-1}\right), 0 \leq n \leq N - 1 \quad (3)$$

2.1.4 Fast Fourier Transform (FFT). A function with limited period can be expressed in Fourier series. Fourier transform is used to convert a time series of bounded time domain signals into a frequency spectrum. The frame that has undergone the windowing process is converted into a frequency spectrum. FFT is a fast algorithm of Discrete Fourier Transform (DFT) which is useful for converting every frame to N samples from time domain into frequency domain. FFT reduces the repeatable multiplication contained in the DFT.

$$X_n = \sum_{k=0}^{N-1} x_k \, e^{-2\pi jkn/N} \quad (4)$$

Equation 4 show n = 0, 1, 2, …, N-1 and j = *sqrt*-1. $X[n]$ is the n-frequency pattern generated from the Fourier transform, W_k is the signal of a frame. The result of this stage is usually called Spectrum or periodogram.

2.1.5 Mel-Frequency wrapping. The perception of the human ear against the sound frequency does not follow the linear scale. The actual frequency scale uses units of Hz. The scale that works on the human ear is called the frequency Mel scale. The scale of Mel-Frequency is a low frequency that is linear under 1000 Hz and a logarithmic high frequency above 1000 Hz [19]. The following equation shows the relation of the Mel scale to the frequency in Hz.

$$F_{mel} = \begin{cases} 2595 * [log]10\left(1 + \frac{F_{HZ}}{700}\right), & F_{HZ} > 1000 \\ F_{HZ}, & F_{HZ} < 1000 \end{cases} \quad (5)$$

Where F_{mel} is the Mel scale and f is the frequency in Hz shown on Equation 5. One approach to the frequency spectrum in the Mel scale with the working function of the human ear as a filter is by Filter Bank. If the $F[N]$ spectrum is the input of this process, then the output is the $M[N]$ spectrum that is the $F[N]$ modified spectrum that contains Power Output of these filters. The spectrum coefficient of Mel is expressed by K, and is specially determined to be 20.

In Mel-frequency wrapping, the resulting FFT signal is grouped into this triangular filter file. The purpose of the grouping here is that each FFT value is multiplied against the corresponding filter gain and the result is summed. Then each group contains a certain amount of signal energy weight as expressed as $m_1 \ldots m_p$. The process wrapping to the signal in the frequency domain is done using the Equation 6.

$$X_i = log_{10}\left(\sum_{k=0}^{N-1} |x(k)| \, H_i(k)\right) \quad (6)$$

Where i = 1, 2, 3, …, M (M is the number of triangle filters) and $H_i(k)$ is the value of the i-triangle filter for the acoustic frequency of k.

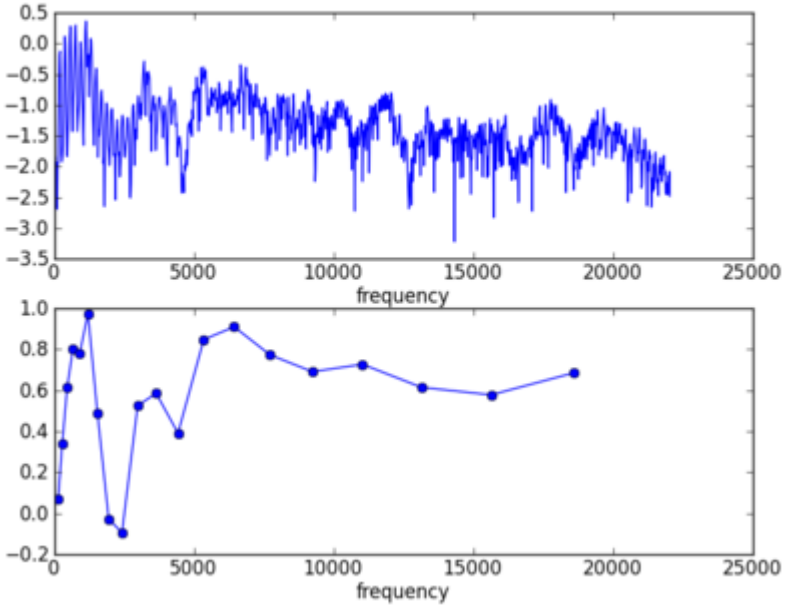


Figure 2. The Original Amplitude Spectrum and the Mel Bank Filter

2.1.6 Cepstrum. Humans listen to voice information based on time domain signals. At this stage Mel-spectrum will be converted into time domain by using Discrete Cosine Transform (DCT). The result is called Mel-frequency cepstrum coefficient (MFCC). Here are the equations used in cosine transformations:

$$c_j = \sum_{j=1}^{K} X_j \cos\left(j(i-1)/2\,\frac{\pi}{K}\right) \quad (7)$$

Equation 7 show C_j is the MFCC coefficient, X_j is the power spectrum of Mel frequency, j = 1, 2, 3, …, K (K is the number of desired coefficients) and M is the number of filters.

2.2 Machine learning using Support Vector Machine (SVM) method

Support Vector Machine (SVM) introduced first by Boser et al is a popular kernel based discriminative classification algorithm. The concept of SVM can be explained simply as a search for the best hyperplane that serves as a separator of two classes in the input space. SVM have been used for various machine learning, such as; object recognition, speech recognition, handwritten character recognition speaker recognition and language recognition. SVM is a binary classification algorithm, and is comprised of sums of kernel function $k(x_i, x_j)$. [20]

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K\left(x_i, x_j + d\right) \quad (8)$$

From Equation 8, $\sum_{i=1}^{N} \alpha_i t_i = 0$, $\alpha_i > 0$, and t_i represent of the ideal outputs either +1 or -1 depends of the class which have sample data. $f(x)$ value compare with the threshold to decides the output class of certain test sample. Multi-class data problem used a one-vs-all approach adapted usually to achieve classification. The SVM train by the Gaussian RBF kernel have the data point x_i and x_j get from Equation 9.

$$K\left(x_i, x_j\right) = exp\left(\gamma \left\| x_i, x_j \right\|\right)^2 \quad (9)$$

After multiple iterations on the train and test data, the optimal hyper-parameters γ and regularization constant C select for the SVM.

3. Methods

The main tools and component used in this research are: Robot Arm, Microphone, Personal Computer, Arduino microcontroller, connections, and others. Algorithm written in Arduino IDLE and Python 2.7. Figure 3 shown generally the process of Speech Recognition to pick and place an object using Robot Arm based on Python 2.7 shown as Figure 3.

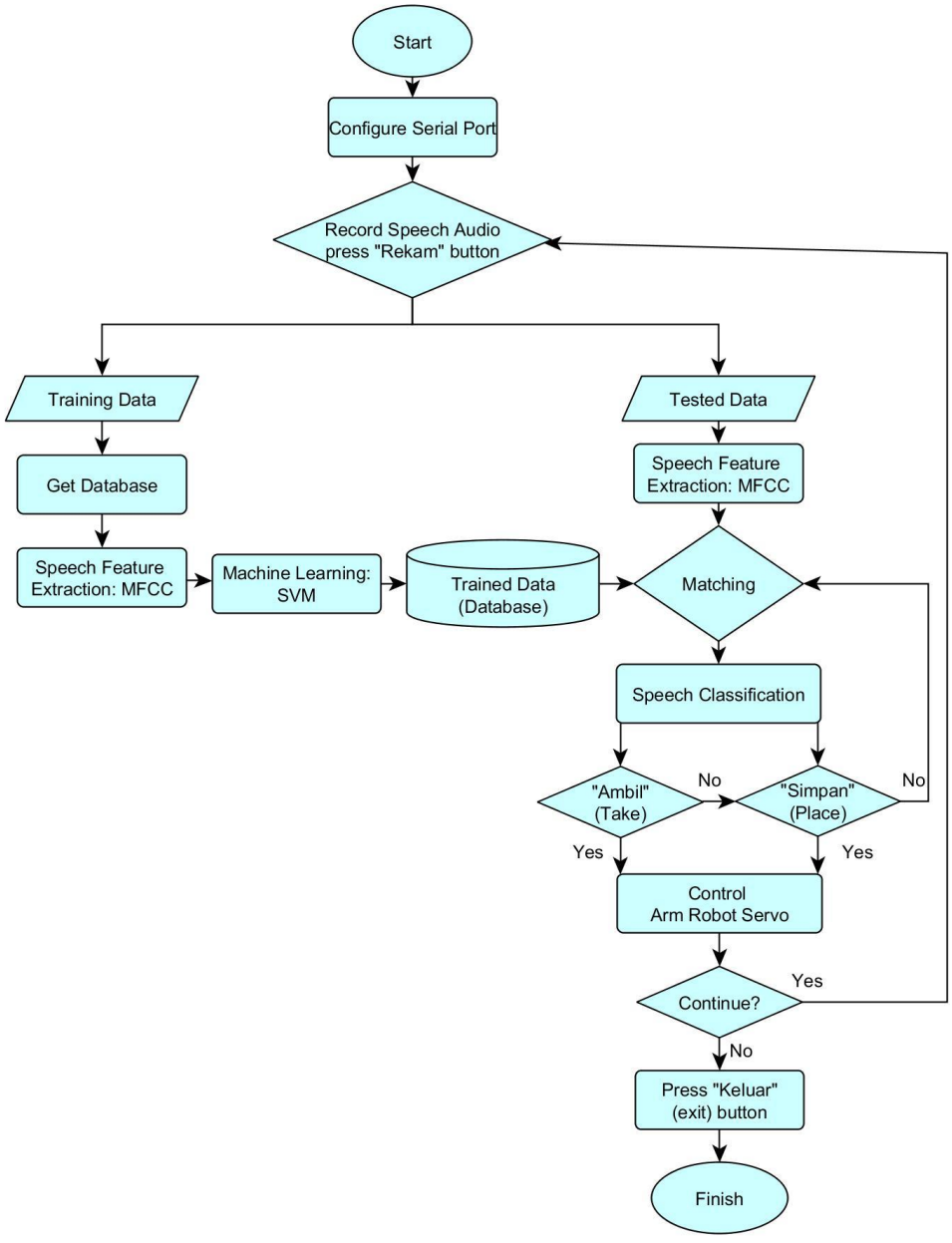


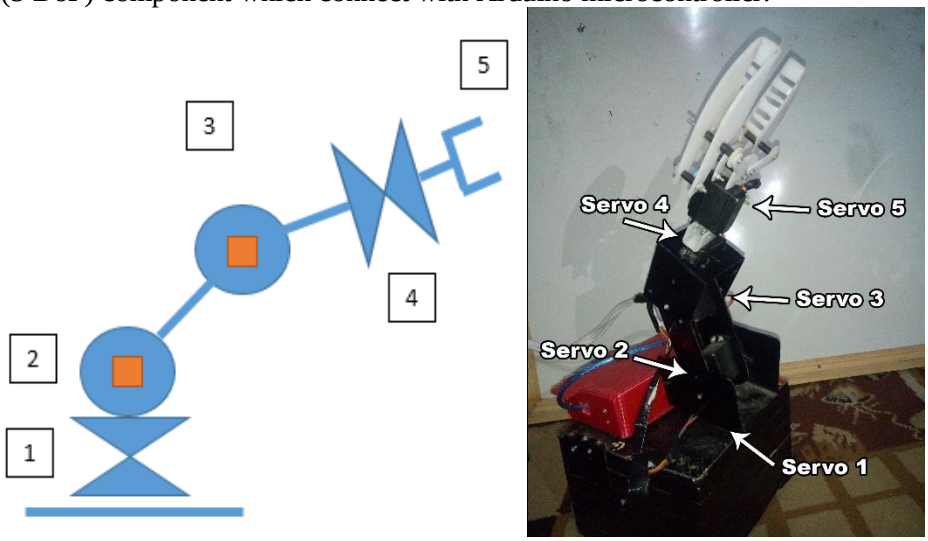
Figure 3. General Research Scheme

From the Figure 3 explain when the system starts and going to record the speech, the process divided to 2 processes: First process is make a training data, consist by features extraction using MFCC and using SVM Method to classifying the speech "pick" (Ambil) and "Place" (Simpan) in Bahasa. Second, is the testing process, such as MFCC features extraction, then matching with Trained Data. The matching data be processed to obtain speech classification. While the

classification process Robot Arm will be move to pick or place an object as our command. All processes work in real time based on Python 2.7 and Arduino microcontroller.

4. Hardware design

Figure 4 is the design and realization of Robot Arm which used in this research, consist by five motor servo (5 DoF) component which connect with Arduino microcontroller.



(a) (b)

Figure 4. Arm robot, (a) Schematic, (b) Realization

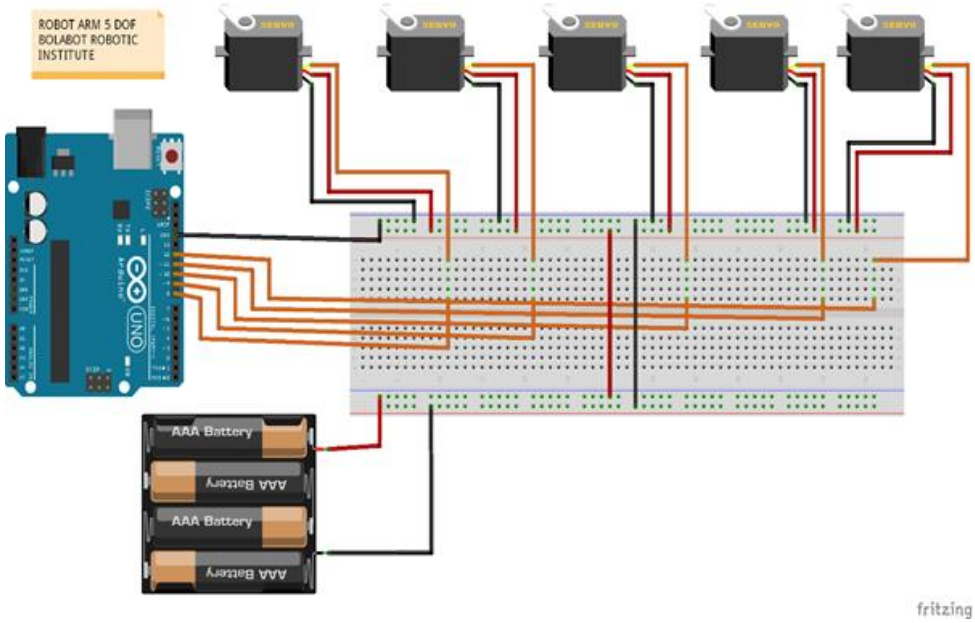


Figure 5. Arm Robot schematic circuit

The circuit schematic of Robot Arm shown as Figure 5, each servo has a supply using 5 volt and 100 mA of battery in order get a better result. Each servo "ground" must be connected with ground on Arduino microcontroller. Robot Arm servos divided into some function; Servo1 as base and rotate horizontal, connect to pin 8. Servo2 work as shoulder and rotate vertical, connect to pin 9. Servo3 work as elbow and rotate vertical, connect to pin 10. Servo4 work

as wrist and rotate horizontal, connect to pin 11. Servo5 work as gripper to place an object, connect to pin 12.

5. Results and discussion

5.1 Features extraction database using MFCC

In this section, to get the robot system which can understand with human speech command, the first step is building the extraction feature database of speech. The speech recognition which used as command to control Robot Arm, they are; Ambil (pick) and Simpan (place) in Bahasa. In this paper, database made from 12 feature extractions and each command made by 10 times of iteration. Table 1 is example of feature extraction of speech recognition database.

Table 1. The Sample of Feature Extraction and Target Data to Build Database of Speech Recognition

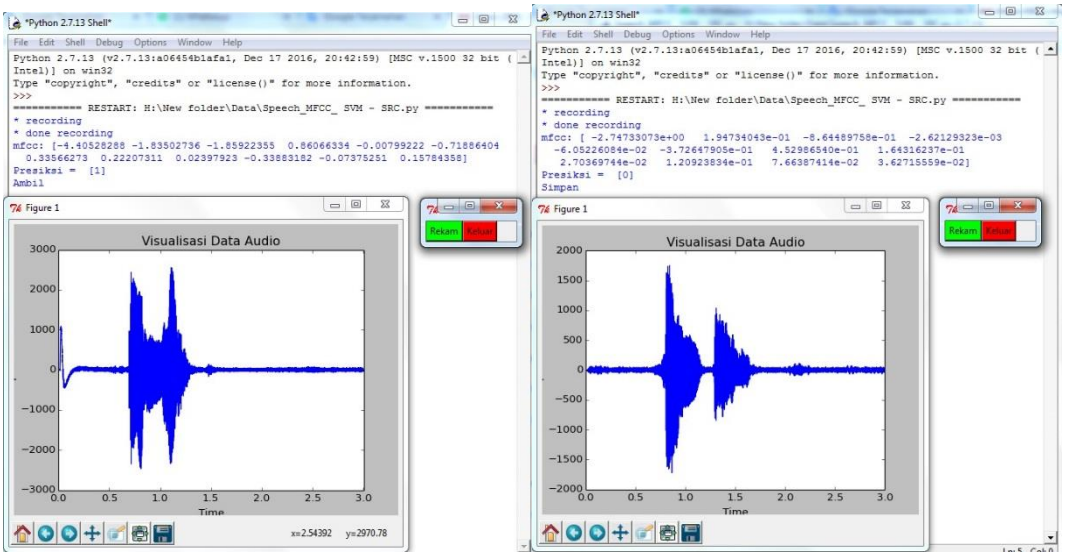
Feature1	Feature2	Feature3	Feature4	Feature5	Feature6	Feature7
-1.3894402	0.49349444	-0.8645897	-0.0418137	-0.0426207	-0.8438808	0.0983683
-2.4718204	0.9484124	-0.6900357	0.0824731	-0.0194620	-0.6640179	0.3672264
-2.5978917	0.7038279	-0.5864564	0.1729673	-0.1777287	-0.7418902	0.4003721
-3.7635666	-0.2808884	-0.3167043	0.8628483	0.3083095	-0.4068732	0.5752230
-2.8758395	0.5090323	-0.5636287	0.3342016	-0.1612223	-0.7586718	0.4330926
-4.9979384	-1.1852745	-1.2478011	0.5314934	0.1793190	-0.6129170	0.4109092
-5.2485744	-1.6981142	-1.6551700	0.2409050	0.0185828	-0.7366620	0.3793570
-4.6518404	-1.0972844	-1.4618936	0.3150198	0.1832025	-0.4717329	0.3949657
-4.9024164	-1.8260630	-1.8674469	0.4471173	-0.0050306	-0.5600959	0.4048072
-4.9956731	-1.4771963	-1.3060416	0.5286050	0.0519628	-0.4727540	0.5400230

Feature8	Feature9	Feature10	Feature11	Feature12	Target
0.4379968	0.2711105	0.0985333	-0.1980512	-0.1092834	0
0.3034041	0.1565578	0.0550747	-0.1512219	-0.0060576	0
0.3323181	0.1809759	0.1534970	0.0210253	0.0445491	0
0.3535670	0.0413849	0.0997217	0.1076920	0.0139399	0
0.4201280	0.2729261	0.1251865	0.0307905	0.0634154	0
0.2684356	0.0229020	0.1040800	0.1286904	-0.0324963	1
0.2412670	-0.0779835	0.0012042	0.1021270	-0.0454572	1
0.0948746	-0.1263564	-0.0301276	0.0369491	0.0481106	1
0.0756162	0.0193277	-0.0112913	0.0877277	0.0804827	1
0.2210660	0.0368926	0.0290753	0.0290403	-0.0018322	1

Table 1 show that the database consists by 12 feature extractions and the target value. The feature extraction as the identity of each speech recognition. While the target value will become an input for SVM method as database to control the Robot Arm. The target "0" is the value for command "Simpan" (place) the object, while target "1" is the value for command "Ambil" (pick) the object. The collected database classifies by SVM method, then the database called the Trained Data.

5.2 Speech recognition system test

Before going to the system test, interface based on Python 2.7 build to get the user friendly when operate the Speech Recognition system shown on Figure 6. The interface consists by menu (Rekam/Record and Keluar/Exit) to operate the program, shell windows to monitoring the result of speech recognition and graphic windows to display the waveform result of the speech recognition.



(a)Waveform for "Ambil" Command (b)Waveform for "Simpan" Command

Figure 6. Interface of Speech Recognition System

After build the speech database become Trained Data, furthermore the trained data is tested by trained respondent for data clarification. The test results shown on Table 2 at Trained Respondent section have the accuracy average rate of speech recognition by trained respondents (in database) is 80%. While, the respondents not trained (outside database) data produces an accuracy rate of 70%.

Table 2. The Trained Data Tested to Trained Respondent and Not-Trained Respondent

Examination	Command	Value	Trained Respondent Result	Not-Trained Respondent Result
1	Ambil	1	1	1
2	Simpan	0	0	0
3	Ambil	1	1	1
4	Simpan	0	1	1
5	Ambil	1	1	0
6	Simpan	0	0	1
7	Ambil	1	1	1
8	Simpan	0	0	0
9	Ambil	1	1	1
10	Simpan	0	1	0

5.3 Speech recognition implementation to robot arm

Identifying and categorizing the trained data of speech recognition are successfully, then the system implemented to 5 DoF Robot Arm to doing task Pick (Ambil) and Place (Simpan) the object. When the testing the speech recognition to control Robot Arm works well top pick and place the object shown on Figure 7.

(a)"Ambil" (pick) an Object (b)"Simpan" (place) an Object

Figure 7. Robot Arm Doing Task

6. Conclusion

This study has been presented to develop Robot Arm which controlled by speech recognition to doing pick and place an object. The speech recognition system based on Python 2.7 using MFCC and SVM method work successfully suitable the speech command. The results obtained speech recognition have a high average accuracy rate of speech recognition, which is 80% of the respondents trained data and 70% of the respondents not trained data. The system implements to 5 DoF Robot Arm based on Arduino microcontroller works effective to pick and place an object. The future works will focus on combination of speech recognition to Social Robot for Human-Robot Interaction.

Acknowledgements

The authors would like gratefully acknowledgment the financial support from LP2M UIN Sunan Gunung Djati Bandung.

References

[1] Sawakare P A, Deshmukh R and Shrishrimal P 2015 *International Journal of Scientific & Engineering Research* **6** 1693-1698
[2] Setiawan A, Hidayatno A and Isnanto R 2011 Aplikasi Pengenalan Ucapan dengan Ekstraksi Mel-Frequency Cepstrum Coefficients (MFCC) Melalui Jaringan Syaraf Tiruan (JST) Learning Vector Quantization (LVQ) untuk Mengoperasikan Kursor Komputer Tech. Rep. 3
[3] Fredj I B and Ouni K 2013 Optimization of Features Parameters for HMM Phoneme Recognition of TIMIT Corpus *International Conference on Control, Engineering & Information Technology* **2** 90-94
[4] Thiang and Wanto 2010 *Seminar Nasional Teknologi Informasi*
[5] Thiang and Wijoyo S 2011 Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot *International Conference on Information and Electronics Engineering*
[6] Das B P and Parekh R 2012 *International Journal of Modern Engineering Research* **2** 854-858

[7] Fredj I B and Ouni K 2013 *Science Journal of Circuits, Systems and Signal Processing* **2** 1-5
[8] Sanjaya W S M and Anggraeni D 2016 *Wahana Fisika* **1** 152-165
[9] Breazeal C 2003 *Advanced Robotics* **17** 97-113
[10] Mubin O, Henderson J and Bartneck C 2014 You Just Do Not Understand Me! Speech Recognition in Human Robot Interaction *International Symposium on Robot and Human Interactive Communication* (IEEE)
[11] Rambabu D, Raju R N and B V 2011 *International Journal of Computational* **3** 92-98
[12] Sanjaya W S M and Salleh Z 2014 *Al-HAZEN Jurnal of Physics* **1**
[13] Abdullahi Z H, Muhammad N A, Kazaure J S and Amuda F A 2015 *International Journal of Computer Science and Electronics Engineering* **3** 11-16
[14] Kumar A, Singh P, Kumar A and Pawar S K 2014 *International Journal of Emerging Technology and Advanced Engineering* **4** 391-393
[15] Tiwari K P and Dewangan K 2015 *International Journal of Science and Research* 10-11
[16] Wardana I N K and Harsemadi I G 2014 *Jurnal Sistem dan Informatika* **9** 29-39
[17] Sanjaya W S M, Anggraeni D and Santika I P 2017 Speech Recognition using Linear Predictive Coding (LPC) and Adaptive Neuro-Fuzzy (ANFIS) to Control 5 DoF Arm Robot *ICCSE* (Bandung: IOP Conference)
[18] Dave N 2013 *Int. J. Adv. Res. Eng. Technol.* **1** 1-5
[19] Mustofa A 2007 *J. Tek. Elektro* **7** 88-96
[20] Ali H, Jianwei A and Iqbal K 2015 *International Journal of Computer Applications* **118** 1-5