Speech Control of Robotic Hand Augmented with 3D Animation Using Neural Network

*Rifky Ismail[1,2], *Mochammad Ariyanto[1,2], Wahyu Caesarendra[1,2]
[1]Center for Biomechanics, Biomaterial, Biomechatronics and Biosignal Processing (CBIOM3S)
Diponegoro University, Semarang, Indonesia
*email: r_ismail@undip.ac.id ; ari_janto5@undip.ac.id

Ismoyo Haryanto[2], Hadianto K. Dewoto[2], Paryanto[2]
[2]Department of Mechanical Engineering
Diponegoro University
Semarang, Indonesia

***Abstract*— In this paper, speech control of robotic hand augmented with 3D animation system has been proposed and presented. Artificial neural network is employed for speech recognition method with tansig and softmax transfer function in hidden layer and output layer. Stream processing method is incorporated for processing the input signal in real time. Thirteen features in frequency domain and time domain that are commonly used in the EMG analysis are utilized in this system. To reduce the influence of noise, voice in noisy environment in the room is recorded as training data set. From the experimental results in offline speech recognition, ANN can recognize the voice command with very high accuracy. In the online real time speech recognition incorporating stream processing, the recognition accuracy decreased about 10%. The proposed speech control of robotic hand augmented with 3D animation is reliable enough with noisy environment.**

Keywords*— Speech control; artificial neural network; stream processing; robotic hand; 3D animation

I. INTRODUCTION

Recently, the numbers of studies in the field of robotic hand have been developed significantly. The robotic hand has been widely used to serve human both in normal/healthy people and people with disabilities such us prosthetic hand. More people can use it if the operation becomes easier and simpler. One of the simplest and easiest ways to operate the robotic hand is using speech. It is one of the most natural ways in communication with robot. Speech is one of the best modalities to interact with robot and to control the robot [1]. Speech control of robotic system has potential application in somewhere voice communication plays an important role in human robot interaction [2,3]. The user only need to talk with the robotic hand using his/her voice in order to the robotic hand can do some tasks.

The common methods for speech recognition are Hidden Markov models as in [4,5], support vector machines [6,7], dynamic time warping [8], and artificial neural network [1]. These methods have been used and tested by many researchers, and the results of this method have accurate enough to recognize the speech recognition. One of the challenges for speech recognition in speech control robot is to choose the accurate feature. Mel Frequency cepstral coefficients (MFCC) is one of the most widely used features in speech recognition. In this paper, we selected thirteen features which comprise of eight features in frequency domain and five features in time domain that commonly used in Electromyography (EMG) analysis.

In this study, the research develops a robotic hand augmented with 3D animation that can be controlled using speech interface. The proposed of speech control interface system is enhanced to reduce the influence of noise by using recorded voice in noisy room as training data. Five fingered low cost robotic hand is employed in this system incorporating 3D animation. Stream processing is used for real time online speech recognition system to drive the robotic hand and 3D animation simultaneously. This paper is organized as follows; the proposed methodology is presented in Section II. Section III will discuss developed hardware and software system. Section IV presents the experimental results. Conclusions and future works are given in Section V.

II. METHODS

A. Feature Calculation

In this study, we selected eight frequency and five time domain features. These features also have been widely used in EMG pattern recognition and analysis [9-11]. The features are selected because they give great accuracy in EMG pattern recognition. The time domain features are widely used due to their classification performance in low noise environments. In frequency domain features, the power spectral density (PSD) is widely used as major analysis [10]. The definition and the equation of selected features are given as follows:

- Mean frequency (MNF) is an average frequency which is computed as sum of product of the signal power spectrum and the frequency divided by total sum of the spectrum intensity [11]. It can be calculated as expressed in (1).

$$MNF = \frac{\sum_{j=1}^{M} f_i P_j}{\sum_{j=1}^{M} P_j} \quad (1)$$

Where f_j is spectrum frequency at frequency bin j, P_j is the signal power spectrum at frequency bin j, and M is length of the frequency.

- Median frequency (MDF) is a frequency at which the spectrum is divided into two areas with equal amplitude [10,11]. It can be calculated as follows

$$\sum_{j=1}^{MDF} P_j = \sum_{j=MDF}^{M} P_j = \frac{1}{2}\sum_{j=1}^{M} P_j \tag{2}$$

- Peak frequency (PKF) is a frequency when the maximum power happens. It can be expressed in (3) as follows

$$PKF = \max(P_j) \tag{3}$$

- Mean power (MNP) is an average signal in frequency domain. It can be defined as in (4)

$$MNP = \frac{\sum_{j=1}^{M} P_j}{M} \tag{4}$$

- Total power (TTP) is an aggregate of the EMG power spectrum. It is also called as zero spectral moment (SM0) [3]. As defined as follows

$$TTP = SM0 = \sum_{j=1}^{M} P_j \tag{5}$$

- The 1st, 2nd, and 3rd Spectral moments, the first three spectral moments are the most important spectral moments [10, 12]. The first three moments can calculated as in (6), (7), and (8).

$$SM1 = \sum_{j=1}^{M} P_j f_j \tag{6}$$

$$SM1 = \sum_{j=1}^{M} P_j f_j^{\,2} \tag{7}$$

$$SM1 = \sum_{j=1}^{M} P_j f_j^{\,3} \tag{8}$$

- Log detector (LOG) definition of the non-linear detector is changed to be based on logarithm and log detector (LOG) feature [10], which can be defined as

$$LOG = \frac{1}{N}\sum_{i=1}^{N} \log|x_i| \tag{9}$$

Where x_i denotes the voice signal in a segment i and N represent the length of input signal.

- Difference absolute standard deviation value (DASDV) is look like root mean squares (RMS) feature, in other words, it is a standard deviation value of the wavelength [13], and it can be calculated as in (10).

$$DASDV = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(x_{i+1} + x_i)^2} \tag{10}$$

- Mean (MN) is the average value of voice signal over the time segment. It can be computed using (11)

$$MN = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i} \tag{11}$$

- Waveform length (WL) can be defined as cumulative length of the signal over the time segment. It can be expressed as in (12)

$$WL = \sum_{i=1}^{N-1} |x_{i+1} - x_i| . \tag{12}$$

- Zero Crossing (ZC) is a number of times that amplitude values of the signal cross zero amplitude level [1].

$$ZC = \sum_{i=1}^{N-1} [sign(x_i \times x_{i+1}) \cap |x_i - x_{i+1}| \geq threshold];$$

$$sign(x) = \begin{cases} 1, if\ x \quad \geq threshold \\ 0, \qquad otherwise \end{cases} \tag{13}$$

B. *Artificial Neural Network*

In this study, the speech recognition of the 13 features mentioned before is classified using the artificial neural network (ANN) as shown in Fig. 1. The outputs of the speech recognition are noise, open, close, or hook. The recognition results are used for controlling the motion of robotic hand and 3D animation. Two layer feed forward networks with a tansig transfer function in hidden layer and a softmax transfer function in the output layer are utilized in this method.

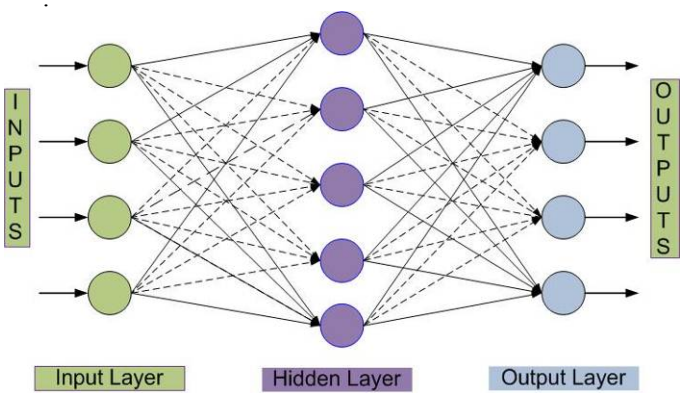


Fig. 1. ANN structure

The first output neuron in hidden layer can be expressed in (14)

$$a^1 = f^1(IWp + b^1) \tag{14}$$

where a^1 is output vector from input layer, p is n-length input vector, IW is input weight matrix, f^1 is transfer function of hidden layer, and b^1 is the bias vector of hidden layer.

The first output neuron in the output layer as can be expressed in (15)

$$a^2 = f^2(LW(f^1(IWp + b^1)) + b^2) \quad (15)$$

where a^2 is output vector from output layer, LW is output layer weight matrix, f^2 is transfer function of the output layer, and b^2 is the bias vector of the output layer.

The Levenberg-Marquardt training algorithm is utilized in neural network. It was designed to approach second-order training speed without having to compute the Hessian matrix. As typical training feedforward neural network, the performance function of this training algorithm has the form of a sum of squares, and the Hessian matrix can be approximated using equation (18) and the gradient can be computed as in (19).

$$H = J^T J \quad (18)$$

$$g = J^T e \quad (19)$$

where J is the Jacobian matrix that contains first derivatives of the neural network errors with respect to the weights and biases, and e is a vector of network errors.

The Levenberg-Marquardt training algorithm uses equation (20) to approximate the Hessain matrix

$$x_{k+1} = x_k - \left[J^T J + \mu I\right]^{-1} J^T e \quad (20)$$

where the scalar μ is zero, using the approximate Hessian matrix. When μ is large, this becomes gradient descent with a small step size.

The ANN for recognition purpose use Mean Square Error (MSE). The MSE measures the magnitude of the estimated errors as shown in (21). Better model will show the smaller values of MSE.

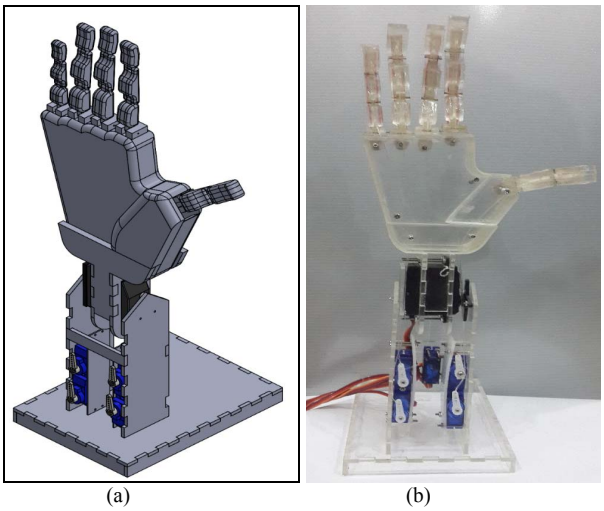
$$mse_{error} = \frac{\sum (y_1 - y_2)^2}{m} \quad (21)$$

where y_1 is the real output in recognition, y_2 is the output from ANN recognition, and m is the total number of samples in speech recognition. The used ANN has 20 neurons in hidden layer and 4 neurons in output layer.

III. HARDWARE AND SOFTWARE SYSTEM

A. Hardware System

In this paper, the research try to develop speech control of robotic hand augmented with 3D animation using ANN. The five fingered low cost robotic hand which has six degree of freedom was used. The proposed 3D (computer aided design) CAD model and the assembled robotic hand are presented in Fig. 2. The 3D model is designed and developed in SolidWorks computer aided design (CAD) software because it is easy to use and operate. The assembled robotic hand is made from acrylic. It has 6 actuators, one medium servo for controlling the wrist motion and five micro servos for controlling the fingers independently. It can be powered using 5 Volts battery with maximum current of 2 Amperes.



(a) (b)

Fig. 2. Robotic hand: (a) 3D model in CAD, (b) Assembled robotic hand.

Fig. 3 shows the overall hardware systems. It comprises of unidirectional microphone, host computer, supporting computer, Arduino Uno Microcontroller, Arduino MEGA microcontroller, and the robotic hand. Unidirectional microphone is selected to acquire the voice command. Host computer works for feature calculations and speech recognition using ANN developed in MATLAB/Simulink environment. Arduino Uno transmits the speech recognition result to Arduino using Arduino IO toolbox that can be downloaded in MathWorks website. The Arduino MEGA is employed to transmit data from Arduino Uno to supporting computer and to drive the robotic hand. The program that is run on Arduino MEGA is built using Simulink Support Package for Arduino. Supporting computer runs 3D Animation developed in SimMechanics environment. The robotic hand and 3D animation of robotic hand can move simultaneously in online speech recognition method using stream processing.

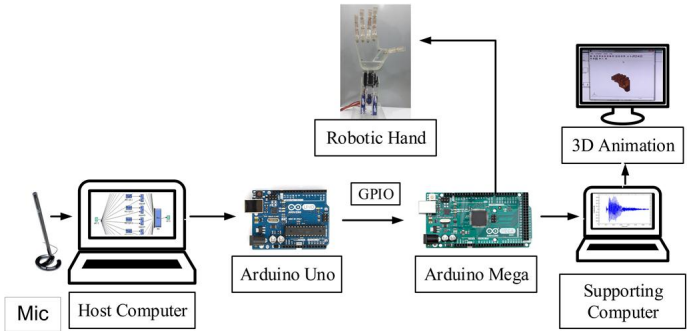


Fig. 3. Integrated hardware system.

B. *Software System*

In this section, the operation of speech control for robotic hand and its 3D animation will be presented. The summarized of the operation system can be seen in Fig. 4. The speech command is read and acquired using unidirectional microphone. The acquired voice signal is then processed using stream processing technique. The next step, the processed signal is calculated in 13 frequency and time domain features. The features go to the ANN to obtain the speech recognition results. When the speech recognized, the system move the robotic hand and 3D animation simultaneously.

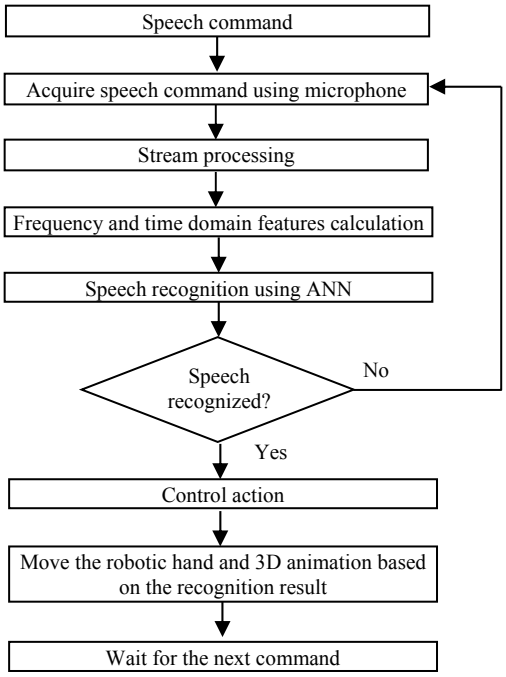


Fig. 4. Speech recognition command flow chart

The ANN for speech recognition in this system is based on training result. The resulted neural network from training is then generated in Simulink block diagram for online speech recognition based on stream processing method. The process of acquiring audio, stream processing, ANN recognition, and control action is developed using block diagram in Simulink environment as shown in Fig. 5.

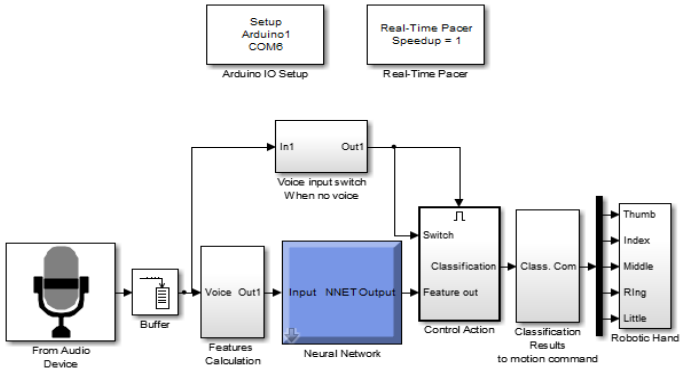


Fig. 5. Speech recognition using neural network in host computer.

Stream processing is most real time signal processing applications for handling large amounts of data. In this system, stream processing is employed for speech recognition. The stream processing is based on frame based sample and it is not by sample based. It can reduce the momory required, reduce the computation burden, give fast computation, and run in real time application. The stream processing parameters that are utilized in this system are summarized in Table I.

TABLE I. STREAM PROCESSING PARAMTERS

Paramter	***Values***
sample rate	22050 Hz
queue duration	1 second
frame size	512 sample.
ouput buffer size	11025 per channel
buffer overlap	512

For 3D animation of robotic hand, 3D model of robotic hand system is developed in SolidWorks CAD software. The 3D CAD model is exported into SimMechanics block diagram using SimMechanics Link for 3D animation of the robotic hand. SimMechanics Link is a CAD plug-in for exporting CAD assemblies from CAD software. The SimMechanics Link generates an XML file detailing the structure and properties of CAD assembly and 3-D geometry files for visualizing the various CAD parts. It can be freely downloaded on the MathWorks website. The result of the imported block diagram of robotic hand from SolidWorks to SimMechanics is shown in Fig. 6. The block diagram in Fig. 6 is read the classification results from host computer using Arduino MEGA. The integration of robotic hand and 3D Animation driven by EMG sensor has been successfully developed in [14].

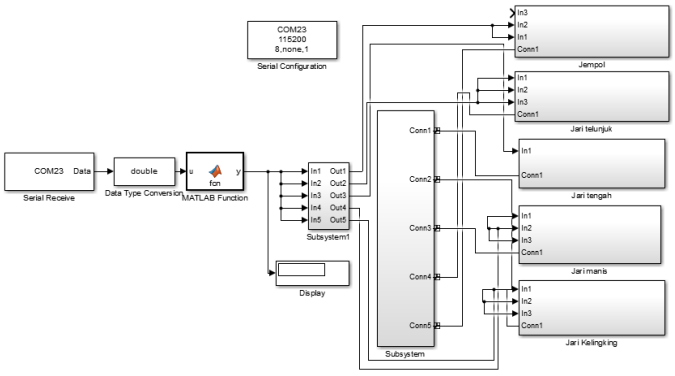


Fig. 6. 3D Animation SimMechanics in supporting computer

IV. EXPERIMENTS AND RESULTS

In the experiment for controlling the robotic hand, user/person talks to the robotic hand augmented with 3D animation using ‘open’, ‘close’, and ‘hook’ voice command. The voice comands then go to the speech recognition system to drive the robotic hand and 3D animation as shown in Fig. 7. In this section,the experiment is divided into two parts, offline and online speech reconition using stream processing.

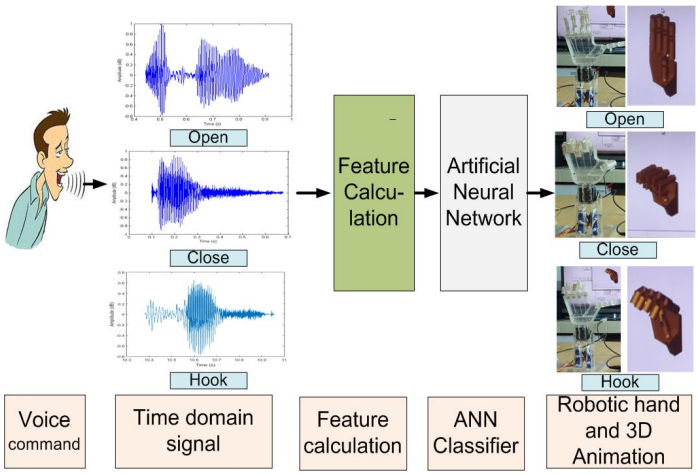


Fig. 7. Online classification test for motion controlof robotic hand augmented with 3D Animation SimMechanics.

ANN for speech recognition requires a training data set that is utilized to learn and to obtain the values of the optimized weights and biases. The training of ANN in this recognition uses Levenberg-Marquardt training algorithm and the perfomance utilizes mean square error (MSE). The input of ANN is 8 features in frequency domain and 5 features in time domain as presented in section II. The networks use 20 neurons in hidden layer. To reduce the influence of noise, voice in noisy environment such as the sound of people talking each other in the room is recorded as training data set. Each of voices from noise, open, close, and hook uses 322 recorded data. Each of 322 recorded data is divided into training, validation, and testing. The data input in ANN for training, validation, and, testing purposes are divided randomly.

The ANN outputs for offline pattern recognition results are presented in the two following table. Table II presents the confusion matrix during training. The highest accuracy for traing is open. The overall performance of ANN during training is 99.1 % and overall error is 0.9 %. This means there is very few miss classification during training process.

TABLE II. CONFUSION MATRIX DURING TRAINING

Command	True Classification			
	Noise	***Open***	***Close***	***Hook***
Noise	207	0	0	0
Open	4	235	1	0
Close	0	0	229	0
Hook	3	0	0	223
Number samples	214	100	230	223
Accuracy (%)	96.7	100	99.6	100
Overall accuracy (%)	99.1			

Table III summarizes the confusion matrix result all of training, validation and testing together. The highest accuracy is open. The overall error is 1.9% and the overall performance is 98.1 %. Based on the Table 3, it can be concluded that the ANN can recognize the voice command with very high accuracy and minimal error in the offline recognition.

TABLE III. OVERALL CONFUSION MATRIX

Command	True Classification			
	Noise	***Open***	***Close***	***Hook***
Noise	311	0	0	0
Open	6	320	2	4
Close	0	0	316	1
Hook	5	2	4	317
Number samples	322	322	322	322
Accuracy (%)	96.6	99.4	98.1	98.4
Overall accuracy (%)	98.1			

The resulted neural network from training is generated in the block diagram that can run in Simulink environment as shown in Fig. 5. Stream processing method for real time online speech recognition is employed in this system using the paramaters as shown in Table I. To test the reliability with the noise, the speech control of robotic hand augmented with 3D animation is tested in the room without noise and with noise. The data for training, validation, and testing is recorded from 1 person. In testing the online speech recognition, two persons drive the robotic hand and 3D animation using their voices.

The experiment of online speech control using stream processing conducted by the first person is shown in Fig. 8. The experiments are conducted several times by two persons in quiet room and noisy room. The performance of robotic hand augmented with 3D animation for online real time speech recognition can be seen online at https://www.youtube.com/watch?v=U-HaC3dzWc0.

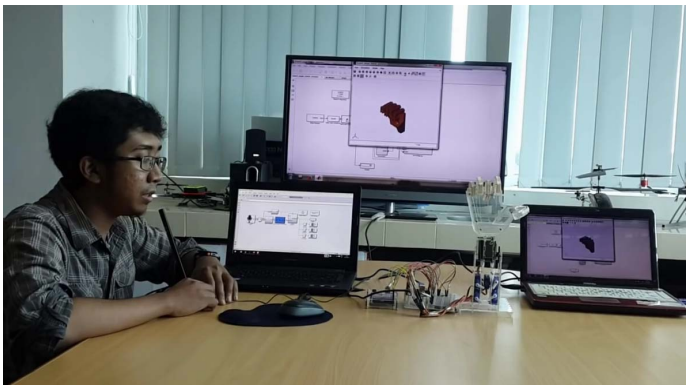


Fig. 8. Online classification test for motion control of robotic hand augmented with 3D Animation SimMechanics.

The experimental results of online recognition in quite and noisy environment are presented in Table IV and Table V. The open voice has the highest accuracy off all both in quite and noisy environment. The hook voice has the lowest accuracy off all both in quite and noisy environment. Hook voice has the lowest accuracy because it is influenced by noise which approaches the features of hook voice. The overall accuracy in quite environment of first person and second person is 90 % and 86.67 % respectively. Meanwhile, overall accuracy in noisy environment of first person and second person becomes 83.33 % and 76.67% respectively. Based on the real time results, the longer the elapsed time the greater the delay. The time delay occurs when the elapsed time runs after about ten minutes.

TABLE IV. EXPERIMENTAL RESULTS OF ONLINE CLASSIFICATION IN QUITE ENVIRONTMENT

Command	First person			Second person		
	Test times	Correct times	% correct	Test times	Correct times	% correct
Open	10	10	100	10	10	100
Close	10	9	90	10	8	80
Hook	10	8	80	10	8	80

TABLE V. EXPERIMENTAL RESULTS OF ONLINE CLASSIFICATION IN NOISY ENVIRONTMENT

Command	First person			Second person		
	Test times	Correct times	% correct	Test times	Correct times	% correct
Open	10	10	100	10	10	100
Close	10	9	80	10	7	80
Hook	10	6	60	10	6	60

V. CONLUSSIONS AND FUTURE WORKS

In this paper, speech control of robotic hand augmented with 3D animation is developed using stream processing in MATLAB/Simulink environment. The system is used 13 features in frequency and time domain that are commonly used in the EMG analysis. Based on the experimental results in offline speech recognition, ANN can perform very well to recognize the voice command with very high accuracy. In the online real time speech recognition using stream processing, the performance of accuracy decreased slightly. The proposed speech control of robotic hand and 3D animation is reliable enough with noisy environment.

On the future works, the features and the stream processing parameters will be optimized to increase the performance accuracy especially in the online real time speech recognition. The voice command vocabulary number as data input will be increased for complex robotic hand tasks and multi-modal robot control. The hardware system will be redesigned in order to it can be implemented on the people with disabilities.

REFERENCES

[1] N. Joshi, A. Kumar, P. Chakraborty and R. Kala, "Speech controlled robotics using Artificial Neural Network," 2015 Third International Conference on Image Information Processing (ICIIP), Waknaghat, 2015, pp. 526-530.

[2] Bojan Kulji, Simon János and Szakáll Tibor, “Mobile robot controlled by voice”, International Symposium on Intelligent Systems and Informatics,189-192, 2007.

[3] Peter X. Liu, A. D. C. Chan, R. Chen, K. Wang, Y. Zhu, “Voice Based Robot Control”, International Conference on Information Acquisition,543547, 2005.

[4] S. Dwivedi, A. Dutta, A. Mukarjee and P. Kulkarni, "Development of a speech interface for control of a biped robot," Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on, 2004, pp. 601-605.

[5] R. Phoophuangpairoj, "Using multiple HMM recognizers and the maximum accuracy method to improve voice-controlled robots," Intelligent Signal Processing and Communications Systems (ISPACS), 2011 International Symposium on, Chiang Mai, 2011, pp. 1-6.

[6] Meng-Ju Han, J. ing-Huai Hsu, Kai-Tai Song and Fuh-Yu Chang, "A new information fusion method for SVM-based robotic audio-visual emotion recognition," 2007 IEEE International Conference on Systems, Man and Cybernetics, Montreal, Que., 2007, pp. 2656-2661.

[7] J. Manikandan and B. Venkataramani, "Hardware implementation of voice operated robot using Support Vector Machine classifier," 2012 Fourth International Conference on Advanced Computing (ICoAC), Chennai, 2012, pp. 1-6.

[8] Xiaoling Lv, Minglu Zhang and Hui Li, "Robot control based on voice command," 2008 IEEE International Conference on Automation and Logistics, Qingdao, 2008, pp. 2490-2494.

[9] M. Ariyanto et al., "Finger movement pattern recognition method using artificial neural network based on electromyography (EMG) sensor," 2015 International Conference on Automation, Cognitive Science, Optics, Micro Electro-Mechanical System, and Information Technology (ICACOMIT), Bandung, 2015, pp. 12-17.

[10] Angkoon Phinyomark, Pornchai Phukpattaranont, Chusak Limsakul, “Feature reduction and selection for EMG signal classification”, Expert Systems with Applications, Volume 39, Issue 8, 15 June 2012, Pages 7420-7431.

[11] Oskoei, M. A., & Hu, H,. “Support vector machine based classification scheme for myoelectric control applied to upper limb”. IEEE Transactions on Biomedical Engineering, 2008, 55(8), 1956–1965.

[12] Sijiang Du and M. Vuskovic, "Temporal vs. spectral approach to feature extraction from prehensile EMG signals," Information Reuse and Integration, 2004. IRI 2004. Proceedings of the 2004 IEEE International Conference on, 2004, pp. 344-350.

[13] Kim, K. S., Choi, H. H., Moon, C. S., & Mun, C. W.” Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions,” Current Applied Physics, vol.11(3), pp. 740–745, November 2011.

[14] R. Ismail, M. Ariyanto, W. Caesarendra, A. Nurmiranto, “Development of Robotic Hand Integrated with SimMechanics 3D Animation”, 2016 International Seminar on Intelligent Technology and Its Application (ISITIA), Lombok, 2016, pp.633-638.