Development of Speech Control for Robotic Hand Using Neural Network and Stream Processing Method

Ismoyo Haryanto, Mochammad Ariyanto, Wahyu Caesarendra and Hadianto K. Dewoto

***Abstract*— The purpose of this paper is to develop speech control for robotic hand using voice input. The voice signal acquired by microphone is processed in real time using stream processing for handling large amounts of data. The processed signal is recognized using Neural Network with *tansig* and *softmax* transfer function in hidden layer and output layer. The Network consists of 20 neurons in hidden layer. Eight frequency domain and five time domain features are employed in speech recognition system. The recognition results from ANN are sent to Arduino Uno to drive the robotic hand motion. Based on the experiment results, ANN can recognize the voice command with 95.9% in offline recognition and 90 %. in online real time recognition. The proposed of speech control system was also tested in noisy environment. The overall accuracy of speech control decreases 10 % in noisy environment. Speech control for robotic hand using stream processing method has been successfully developed.**

***Index Terms*—Robotic hand, Speech, Neural Network, Feature calculation, Stream processing.**

I. Introduction

ARTIFICIAL neural network have been widely used in pattern recognition, nonlinear regression, and control system due to its performance result. In this study, the research will develop speech control of robotic hand using stream processing method based on artificial neural network (ANN). The neural network has been successfully implemented in speech recognition as studied in literatures [1,2]. ANN can recognize the speech recognition with good performance. The other common classification methods are Hidden Markov models as in [3], support vector machines [4], and dynamic time warping [5].

Feature extraction is one of the most important steps in speech recognition system. It extracts useful features from the raw data which can help the classifier to make decisions. The common features that are commonly used in speech recognition are Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding Coefficients (LPCC). In this paper, thirteen features that have been widely used due to its performance in Electromyography (EMG) analysis are employed in this speech control of robotic hand.

Manuscript received October 22, 2016.

I. Haryanto, M. Ariyanto and W. Caesarendra are with the Mechanical Engineering Department, Diponegoro University, Jl. Prof. Sudarto, SH – Tembalang, Semarang 50275, Indonesia (e-mail: ismoyo_h@undip.ac.id; ari_janto5@undip.ac.id ; w.casearendra@gmail.com).

Hadianto K. Dewoto was with the Mechanical Engineering Department, Diponegoro University, Jl. Prof. Sudarto, SH – Tembalang, Semarang 50275, Indonesia (e-mail: hadykun@gmail.com).

This paper will focus on developing a speech control of robotic hand using stream processing method. Stream processing is employed to handle large data that comes from acquired speech command. One of common challenges in speech recognition is the noise. In order to reduce the effect of noise, the sound from noisy room is recorded and added in data input for training data in neural network. The robustness of proposed speech system will be tested in quiet and noisy room.

II. Feature Calculations

In this paper, thirteen features which comprises of eight frequency domain and five time domain features is selected and utilized in this study. These features have been found in many literatures giving the maximum classification performance in EMG analysis [6-8]. The frequency domain features are MNF, MDF, PKF, MNP, TTP, SM1, SM2, SM3. The time domain features are LOG, DASDV, MN, WL, ZC. The features can be summarized in Table 1. In frequency domain, 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 bin. In time domain, x_i denotes the voice signal in a segment i and N represent the length of input signal Final Stage.

III. Neural Network

Artificial Neural Network (ANN) is utilized in speech recognition system to recognize speech command from user. The general structure of ANN can be depicted in Figure 1. The structure of feed-forward ANN model comprises of three-layers of nodes. ANN is widely used to perform complex tasks such as control systems, pattern recognition, forecasting, identification, speech, and computer vision.

The first output neuron in hidden layer can calculated using equation (1) and the first output neuron in the output layer is defined in (2).

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

$$a^2 = f^2(LW(f^1(IWp + b^1)) + b^2) \tag{2}$$

TABLE I
UTILIZED FEATURES

Features	Description	Formula
MNF	Average frequency as sum of product of the signal power spectrum and the frequency divided by total sum of the spectrum intensity [9]	$MNF = \frac{\sum_{j=1}^{M} f_i P_j}{\sum_{j=1}^{M} P_j}$
MDF	Frequency at which the spectrum is divided into two regions with equal amplitude [6,9]	$\sum_{j=1}^{MDF} P_j = \frac{1}{2}\sum_{j=1}^{M} P_j$
PKF	Frequency when the maximum power occurs.	$PKF = \max(P_j)$
MNP	Average signal power spectrum	$MNP = \frac{\sum_{j=1}^{M} P_j}{M}$
TTP	Aggregate of the EMG power spectrum	$TTP = \sum_{j=1}^{M} P_j$
SM1	1st Spectral moment	$SM1 = \sum_{j=1}^{M} P_j f_j$
SM2	2nd Spectral moment	$SM2 = \sum_{j=1}^{M} P_j f_j^2$
SM3	3rd Spectral moment	$SM3 = \sum_{j=1}^{M} P_j f_j^3$
LOG	Non-linear detector is changed to be based on logarithm and log detector (LOG) feature [6]	$LOG = \frac{1}{N}\sum_{i=1}^{N} \log\lvert x_i\rvert$
DASDV	It is a standard deviation value of the wavelength [10]	$\sqrt{\frac{\sum_{i=1}^{N-1}(x_{i+1}+x_i)^2}{N-1}}$
MN	The average value of voice signal over the time segment	$MN = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i}$
WL	Cumulative length of the signal over the time segment	$WL = \sum_{i=1}^{N-1}\lvert x_{i+1} - x_i\rvert$
ZC	Number of times that amplitude values of the signal cross zero amplitude level [1]	$ZC = \sum_{i=1}^{N-1}[sign(x_i \times x_{i+1}) \cap \lvert x_i - x_{i+1}\rvert \geq threshold]$;

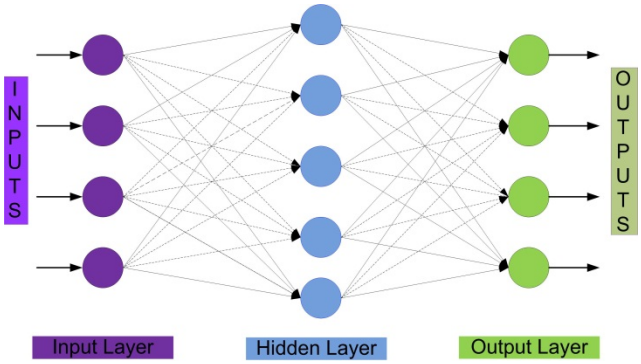


Fig. 1. Neural network structure.

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. In the output layer, a^2 denotes 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.

Levenberg-Marquardt training algorithm is selected and employed in this speech recognition. It was designed to approach second-order training speed without having to compute the Hessian matrix. As typical training feed forward networks, the performance function of this training algorithm has the form of a sum of squares, and the Hessian matrix can be approximated using equation (3) and the gradient can be calculated as in (4).

$$H = J^T J \tag{3}$$

$$g = J^T e \tag{4}$$

For measuring the resulted error from recognition, Mean Square Error (MSE) is utilized in this ANN. For better accuracy of ANN, the value of RMS becomes small. The MSE computes the magnitude of the errors as shown in (5).

$$MSE = \frac{\sum (y_1 - y_2)^2}{m} \tag{5}$$

Where y_1 denotes the real output from recognition, y_2 is the output from ANN, and m is the total number of samples in speech recognition.

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. The inputs are thirteen features and the outputs of the speech recognition are noise, open, close, or hook as shown in Figure 2. The recognition results are used for controlling the motion of robotic hand. The used ANN has 20 neurons in hidden layer and 4 neurons in output layer. The ANN for speech recognition is developed in MATLAB/Simulink environment.

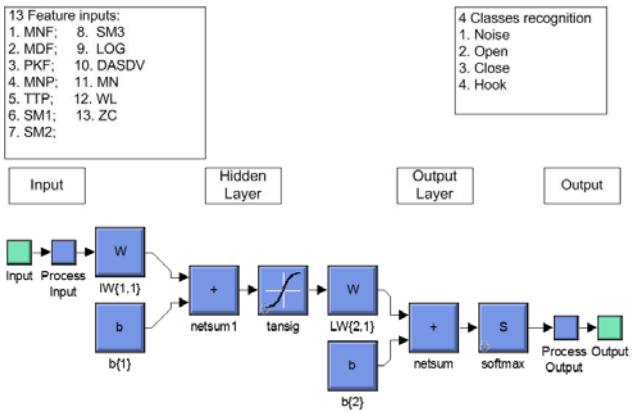
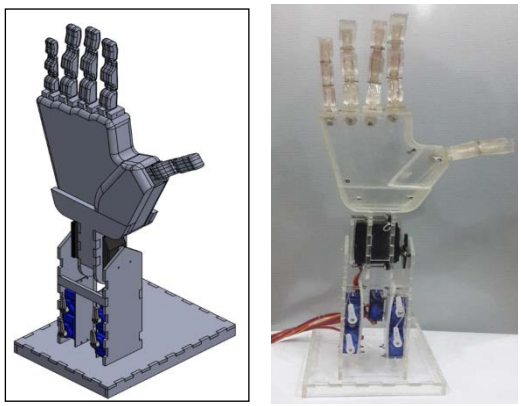


Fig. 2. Neural network in speech control

IV. HARDWARE SYSTEM

Low cost 6 degree of freedom (DOF) anthropomorphic robotic hand is built in this speech recognition system. It has five micro servos on each fingers and 1 medium servo in wrist. The robot has 15 joints in fingers and made from acrylic. The flexion motion of finger is driven by micro servos using guitar string and the extension motion of finger is driven by torsional spring that is attached on each joint in each fingers. The proposed mechanism is robust and simple enough. The total weight of the robotic hand system, including mechanical material and electronics is 450 grams. The size is approximately about 180 mm x 80 mm, and the width of the palm is 25 mm. This size is about the average size of a human hand in Indonesia. The size of each finger is summarized in Table 2. The robotic hand can be powered with the current and voltage about 2 A and 5 V. The 3D design of robotic hand is developed in Solid Works software because of easy to use. The 3D computer aided design (CAD) model and the assembled one can be seen in Figure 3.



(a) (b)

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

The hardware system of speech control is comprises of unidirectional microphone, computer, Arduino Uni, and the robotic hand as shown in Figure 4. Unidirectional micriphone acquires the voice command from user. The voice signal is converted in analog voltage signal that can be measured in host computer. The acquired signal is processed using stream processing method that runs in computer and recognized using ANN in real time. Thirteen features is executed and calculated in real time using stream processing to handle large data. The recognition results is send to Arduino Uno microcontroller via serial USB to drive the micro servos. Arduino Uno drives the robotic hand actuator using digital input output (DIO) pins.

TABLE II
FINGER DIMENSIONS OF THE ROBOTIC HAND

Fingers	Proximal (mm)	Medial (mm)	Distal (mm)
Thumb	38	-	35
Index	29	30	28
Middle	29	30	28
Ring	29	30	28
Little	23	24	23

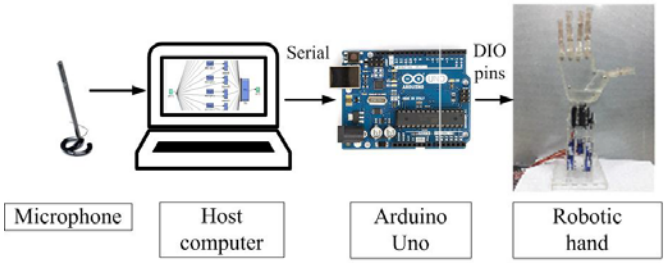


Fig. 4. Hardware integration

V. SOFTWARE SYSTEM

The block diagram for controlling the robotic hand using voice command is developed in MATLAB/simulink environment. The main block diagram is depicted by Figure 5. Voice is acquired using From Audio Device block from Digital Signal Processing (DSP) Blockset. The signal is processed using stream processing method to handle large data and memory efficeincy. Selected parameters of stream processing is summarized in Table 3. The processed signal is calculated to 13 features and send them to the ANN block. The ANN block is generated from training resultsin MATLAB environment. The resulted network from training is then built in Simulink for online speech recognition based on stream processing method. The recognized voice command is sent to the Arduino Uno microcontroller using Arduino IO block that can be downloaded freely in MathWorks website.

The operation of speech control for robotic hand is presented in Figure 6. The operation of speech control system runs in real time. The operation block diagram as shown in Figure 5 is embedded in host computer. The communication of host computer and Arduino Uno microcontroller uses serial commnunication with 115200 bautrate. The bautrate is fast enaugh for transmitting the recognized voice from host computer to the microcontroller.

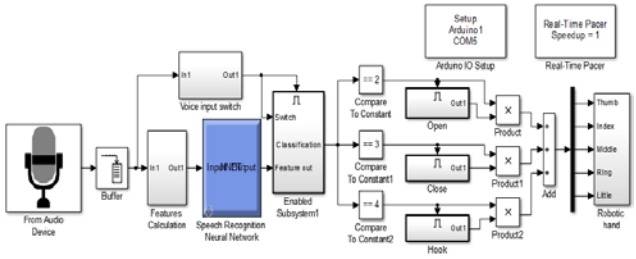


Fig. 1. Speech recognition block diagram in host computer

TABLE III
PARAMETERS OF STREAM PROCESSING

Paramter	Values
Sample rate	22050 Hz
Queue duration	1 second
Frame size	512 sample
Ouput buffer size	11025 per channel
Buffer overlap	512

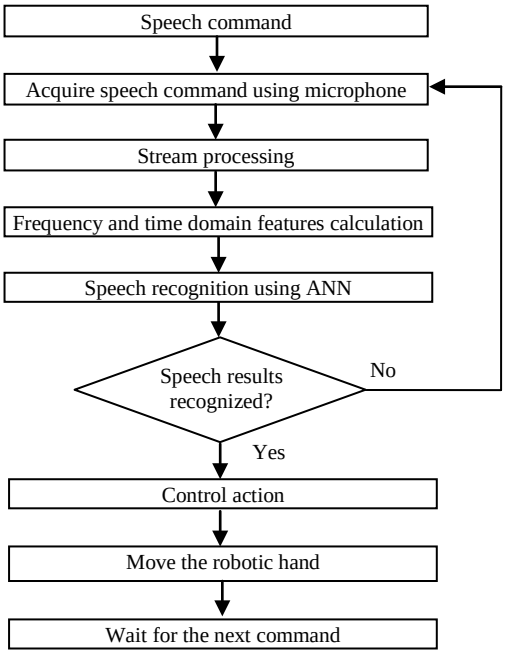


Fig. 6. Speech recognition command flow chart

VI. EXPERIMENT RESULTS

In the experiments, a person talks to the robotic hand with 'open', 'close', and 'hook' voice command. The voice commands then go to the ANN speech recognition system as can be seen in Figure 7. The robotic hand moves the fingers based on the recognized voice commands.The speech control system wil be tested in two condition, offline speech recognition and online speech recognition using stream processing method.

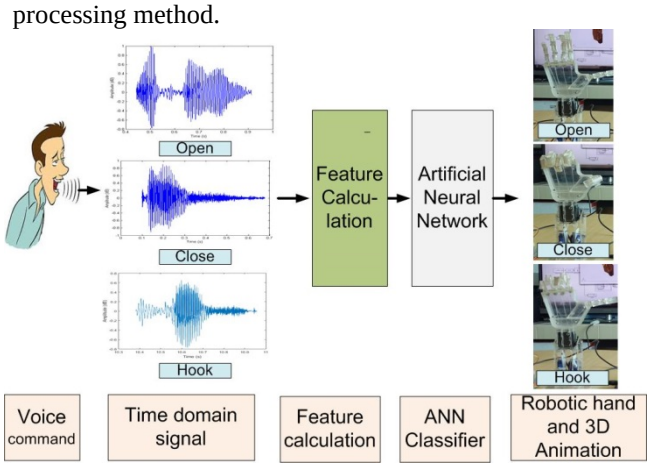


Fig. 7. Speech recognition system

In order to reduce the noise effect, voice command 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"is recorded 322 times. Each of recorded voices is divided into training, validation, and testing. The data input in ANN for training, validation, and, testing purposes are divided randomly. Figure8 shows at about 136 epochs the MSE of training, test, and validation get stabilized and the MSE is very low. This means that the neural network gives very high accuracy of speech recognition. The value of MSE is 0.0150 the resulted the end of the neural network training. This architecture is selected as final network for speech recognition of robotic hand.

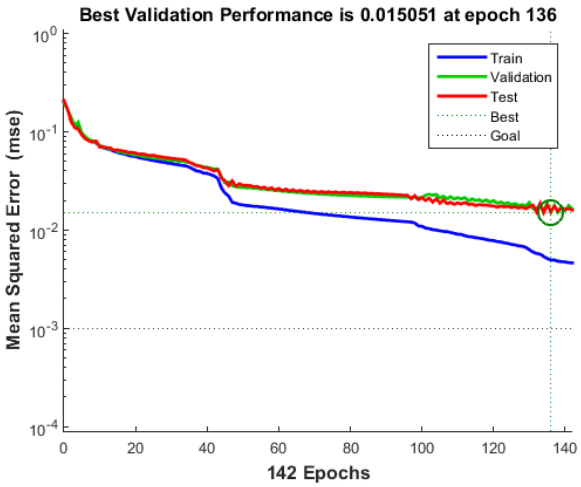


Fig. 8. Neural network performance during training

A. *Offline Recognition*

In the offline recognition, confusion matrix is widely used to present the recognition results. Each rows of matrix represents the cases in predicted class. Table 4 to table 7 presents confusion matrix for training, validation, testing, and all of the data combined together. There are 1288 data samples divided into 3 subsets, 902 for training, 193 for validation, and 193 for training. Table 4 displays confusion matrix results during training.The ANN performs very good result for open (100% recognition). The overall accuracy is 99.1 % and overall error is 0.9 %.This means that the networks can classify the speech command with minimal error during training. Table 5 and Table 6 show the confusion results during validation and testing. The highest recognition accuracy is open for validation and hook for testing. The overall performance in both validation and testing is 95.9 %. It indicates that the network performs good results during validation and testing.

For the overall confusion matrix, 1288 data input is taken. Table 7 presents the overall confusion matrix result. From the result, open is the highest accuracy of all speech recognition. The confusion shows that the overall performance is 98.1 % and the overall error is 1.9 %. Based on the table, it can be concluded that the ANN can perform speech recognition task with very high accuracy in offline speech recognition.

TABLE IV
CONFUSION MATRIX RESULTS 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
Accuracy (%)	96.7	100	99.6	100
Overall accuracy (%)			99.1	

TABLE V
CONFUSION MATRIX RESULTS DURING VALIDATION

Command	True Classification			
	Noise	Open	Close	Hook
Noise	54	0	0	0
Open	1	51	0	3
Close	0	0	40	1
Hook	1	1	1	40
Accuracy (%)	96.4	98.1	97.6	95.9
Overall accuracy (%)			95.9	

TABLE VI
CONFUSION MATRIX RESULTS DURING TESTING

Command	True Classification			
	Noise	Open	Close	Hook
Noise	50	0	0	0
Open	1	34	1	1
Close	0	0	47	0
Hook	1	1	3	54
Accuracy (%)	96.2	97.1	92.2	98.2
Overall accuracy (%)			95.9	

TABLE VII
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
Accuracy (%)	96.6	99.4	98.1	98.4
Overall accuracy (%)			98.1	

B. Real Time Online Recognition

In this section, online real time speech recognition test is conducted using stream processing method using stream processing parameter as presented in Table 3. To test the robustness of speech control, the system will be tested in noisy environment by same person. Table 8 shows that the speech recognition system can perform with high accuracy in quite environment. From the Table 9, the overall performance of speech recognition system to drive the robotic hand only reduces 10 %. The overall performance of online real time speech system is 90 % in quite environment and 80 % in noisy environment.

TABLE VIII
ONLINE RECOGNITION IN QUITE ENVIRONMENT

Speech	Test times	Correct times	% correct
Open	10	10	100
Close	10	9	90
Hook	10	8	80

TABLE IX
ONLINE RECOGNITION IN NOISY ENVIRONMENT

Speech	Test times	Correct times	% correct
Open	10	10	100
Close	10	9	80
Hook	10	6	60

VII. CONCLUSIONS

Speech control of robotic hand using stream processing method has been successfully built in this paper. The speech recognition system has high accuracy in both offline and online recognition. The system can recognize the voice command with 95.9 % and 90% accuracy in offline and online real time recognition. The overall accuracy of online speech recognition decreases 10 % in noisy environment.For future study, the stream processing parameters will be optimized to improve the accuracy in online speech recogniton. The vocabulary of data input also will be increased to speech recognition system.

REFERENCES

[1] Gülin Dede, Murat Hüsnü Sazlı, "Speech recognition with artificial neural networks", Digital Signal Processing, vol. 20(3), pp. 763-768, 2010.

[2] 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, pp. 526-530, 2015.

[3] 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, ROMAN 2004, 13th IEEE International Workshop on, pp. 601-605, 2004.

[4] 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, 2012.

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

[6] Angkoon Phinyomark, Pornchai Phukpattaranont, Chusak Limsakul, "Feature reduction and selection for EMG signal classification", Expert Systems with Applications, vol. 39(8), 15 June 2012, pp. 7420-7431, 2012.

[7] 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, pp. 12-17, 2015.

[8] Angkoon Phinyomark, Chusak Limsakul, Pornchai P. "A Novel Feature Extraction for Robust EMG Pattern Recognition", Journal of Computing, vol. 1, pp. 71-80, 2009.

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

[10] K.S. Kim, H.H. Choi, C.S. Moon and C.W. Mun, ” 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.

Ismoyo Haryanto received B.Sc. and M.Sc. degrees from Institute of Technology Bandung, Bandung, Indonesia, in 1993 and 1997 respectively, and the Dr.-Ing. from the University of Technology Munich, Germany, in 2016. He is currently a Lecturer at Diponegoro University, Semarang, Indonesia. His research interests include soft computing, multidisciplinary optimization and structural dynamics.

Mochammad Ariyanto received B.Sc. and M.Sc. degrees from Department of Mechanical Engineering, Diponegoro University, Indonesia, in 2010 and 2013 respectively. Currently, he is a Lecturer at Diponegoro University, Semarang, Indonesia. His research interests are robotics, biomechatronics, and intelligent control.

Wahyu Caesarendra received B.Sc. degree from Department of Mechanical Engineering, Diponegoro University, Indonesia in 2010, and M. Eng. from Pukyong National University Korea. He got PhD degree in University of Wollongong Australia in 2015. Currently, he is a Lecturer at Diponegoro University, Semarang, Indonesia. His research interests are signal processing, pattern recognition, and mechatronics.

Hadianto K. Dewoto received his B.Sc. degree in Mechanical Engineering Diponegoro University in 2016. His research include speech control and stream processing.