2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)
July 6-11, 2014, Beijing, China

Vowel Recognition System of Lipsynchrobot in Lips Gesture Using Neural Network

Indra Adji Sulistijono, *IEEE Member*, Haikal Hakim Baiqunni, Zaqiatud Darojah and Didik Setyo Purnomo

***Abstract*— In this research, we propose a system of vowel recognition on the shape of the lips using a neural network backpropagation. For recognizing the shape of the vowels in the lips by image capturing through the webcam, and then processed through image processing which includes edge detection, filtering, mouth feature extraction, integral projection and vowels recognition on the lips using a back propagation neural network. The vowel recognition results obtained at the lips of the testing process neural network based on the training data and test data. The proposed method works well. The final results obtained from this research is a system that able to detect the lips with good features, with the success of recognizing vowels using 250 training data is 70%. The output data from the vowel recognition on the lips is used as an input to Lipsynchrobot that followed the gesture of human lips.**

I. Introduction

The need for an automatic lip-reading system is ever increasing. In fact, today, extraction and reliable analysis of facial movements make up an important part in many multimedia systems such as videoconference, low communication systems, lip-reading systems. In addition, visual information is imperative among people with special needs. We can imagine, for example, a dependent person ordering a machine with an easy lip movement or by a simple syllable pronunciation. Moreover, people with hearing problems compensate for their special needs by lip-reading as well as listening to the person with whome they are talking.

Robotic systems with artificial emotions and facial expressions have emerged since the late 1990s Tamagotchi from Bandai, though the software maker agent for virtual environments, regarded as the first to make this system [1], [2]. This study has been made Lipsynchrobot which is the development of the project FMX has ever made before. The word ”Lipsynchrobot” is short for ”Lips Synchronize Robot” which has the meaning of ”Lips Movement Imitators Robot”, in which the head of the robot can imitate human lip movement. In mimicking human lip movement needed a vowel recognition methods on human lips. In vowel recognition system on the lips using a webcam capture image data and processed through the OpenCV (Open Source Computer Vision). Image processing is used to perform pattern recognition vowels on human lips, which in terms of the pattern recognition using Artificial Neural Network (ANN).

In this context, many works in the literature, from the oldest [1] until the most recent ones [3], [4], [7]-[9] have proved that movements of the mouth can be used as one of the speech recognition channels. Recognizing the content of speech based on observing the speakers lip movements is called lip-reading. It requires converting the mouth movements to a reliable mathematical index for possible visual recognition.

Humanoid robot has a high potential of being a future form of computer that acts and supports our daily activities in the shared infrastructures with the human beings. In spite of such expectation, humanoid robots at the present substantially lack in mobility. They can neither cope with sudden contacts with the unknown, nor respond to unexpected decisions due to emergency such as stop or avoid. A technical reason lies in the fact that no control algorithm has been developed or implemented in a responsive form that allows the above-stated high-mobility.

Word recognition methods have been developed, so that even unspecified voices uttered by unspecified person can be recognized in high recognition rate. However, in noisy environments, noises cause serious recognition problems and recognition rates become low as a result. But word recognition is practically expected to be utilized in noisy environments. Considering these points, supplemental use of visual information of lip shape movements is expected to improve the performances. In order to use lip shape movements for word recognition, a fast and accurate extraction method of lip shapes from face images is required.

Area extraction technique is able to applied to lip shape extractions. Area extraction is one application of area extraction techniques and is developed by several methods such as spatial filtering methods. We have already worked on real time facial expression simulator [12],[13]. Therefore, we proposed vowel on recognition system of lipsynchrobot in lips gesture using neural network backpropagation. First, image capturing through the webcam, and then processed through image processing which includes edge detection, filtering, mouth feature extraction, integral projection and vowels recognition on the lips using a back propagation neural network.

This paper is organized as follows, section II explains how the feature is extracted from the humanoid robot, image

Indra Adji Sulistijono is with the Mechatronics Engineering Division, Electronics Engineering Polytechnic Institute of Surabaya, Politeknik Elektronika Negeri Surabaya (PENS), Kampus PENS, Jalan Raya ITS Sukolilo, Surabaya 60111, Indonesia. (Tel: +62-31-594-7280; Fax: +62-31-594-6114; Email: indra@pens.ac.id).

Haikal Hakim Baiqunni is Employee at PT Indonesia Epson Industry, Cikarang, Bekasi, Indonesia. (Email: haikaru.noriyuki@gmail.com).

Zaqiatud Darojah and Didik Setyo Purnomo are with the Department of Mechanical and Energy Engineering, Politeknik Elektronika Negeri Surabaya (PENS), Kampus PENS, Jalan Raya ITS Sukolilo, Surabaya 60111, Indonesia. (Email: {zaqiah,didiksp}@pens.ac.id).

This work was supported by UPPM-PENS, and EEPIS Robotics Research Center (ER2C).

978-1-4799-2072-3/14/$31.00 ©2014 IEEE

processing, Gabor filter, local binary pattern and integral projection. Section III shows lip vowel recognition. Section IV talk about several experimental results and section V concludes the paper.

II. Feature Extraction

A. Humanoid Robot

Android is a robot that mimics the appearance of the human form or emulate part of the human body. This enables a more humane interaction with humans. In general, Android has a body, two arms, two legs, and head. Although some are made from one part of human body. Some android also has a full face with eyes and mouth.

According to Jeff Prucher [3] Android is a robot or synthetic organism designed to look and act like a human being, especially with a body that has flesh-like resemblance. Android is a humanoid robot which was built to resemble the humans, as well as the robot "Doldori" made in Korea country that are able to express human-like facial expressions [3]. Facial expression is a method of emotional expression by humans to recognize the emotional robot [12]. This research will discuss about the Lipsynchrobot that can interact in a way mimicking the movement of the lips form vowels. The robot is controlled via a wireless serial communication using bluetooth. Lipsynchrobot has a camera sensor to capture the movement of the lips vowels with the help of image processing and Artificial Neural Network (ANN). Lipsynchrobot has 5 basic shapes of vowels A, I, U, E and O.

B. Image Processing

Image Processing is a system which the input is in the form of images (pictures) and the result is also the form of the image (pictures) [6]. Image processing here has a purpose that is improve the quality of the image to be easily interpreted by a human or computer and pattern recognition, the grouping of symbolic and numeric data (including images) automatically by the computer to an object in the object can be recognized and interpreted.

C. Gabor Filter

To get important information from an image texture needed a feature extraction process. Techniques in extracting characteristic using Gabor wavelet function which is used to extract the characteristics of the normalized image. 2D Gabor filter is obtained by modulating a 2D sine wave at a certain frequency and orientation of the Gaussian envelope [10]. 2D Gabor filter function minimizes traits that are not important in the spatial and frequency area. 2D Gabor basic function is defined as:

$$g(x,y) = exp\left(\frac{x'^2 + y'^2}{2\sigma^2}\right) cos\left(2\pi\frac{x'}{\lambda} + \psi\right) \quad (1)$$

$$x' = xcos\theta + ysin\theta \quad (2)$$

$$y' = -xsin\theta + ycos\theta \quad (3)$$

where λ = sine wavelength, θ = Gabor function orientation, y = offset phase and σ = deviation standard of Gaussian envelope.

D. Local Binary Pattern

LBP (Local Binary Pattern) was first introduced in 1996 by Ojala to describe texture in grayscale mode. LBP operator based on a 3x3 neighbourhood representing the local texture around the center pixel [4]. LBP is defined as the ratio of the binary pixel value on central pixel image with 8 surrounding pixel values . For example on a 3x3 sized image, binary values at the center of the image compared with around value. Around value would be 1, if the value of the central pixel is smaller, and will be 0 if the binary center value is larger. After that, arrange 8 binary values clockwise or vice versa and change the 8-bit binary into decimal value to replace the value of the center pixel [11]. In Figure 1. an image processing scheme using LBP (Local Binary Pattern).

LBP can be formulated into the equation:

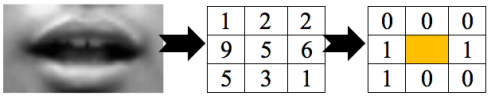


Fig. 1. Local binary pattern scheme.

LBP can be formulated into the equation:

$$LBP(x_c, y_c) = \sum_{p=0}^{p-1} 2^P s(i_p - i_c) \quad (4)$$

where (x_c,y_c) = pixel center with intensify i_c.

And the function $s(x)$ is defined as follows:

$$s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (5)$$

E. Integral Projection

Integral projection is a method used to find the location of the object area. This method can be used to detect the boundary of different image regions so as to find the location of the features of an image. The integral is working by summing the values of pixels from each row and each column. Integral Projection is defined by:

$$h(j) = \sum_{i=1}^{Nrow} x(i,j) \quad (6)$$

$$h(i) = \sum_{i=1}^{Ncolumn} x(i,j) \quad (7)$$

where $h(j)$ = number of pixels X each row, $h(i)$ = number of pixels Y each column.

F. Mouth Feature Extraction

To extract mouth features performed image processing to get the pixel data values in an image, especially in the mouth. To get the data pixels in the mouth do cropping only in mouth area.

After getting the location of the lips on the face, that location is marked as a work area extraction of Region of Interest (ROI). To perform feature extraction on the lips using

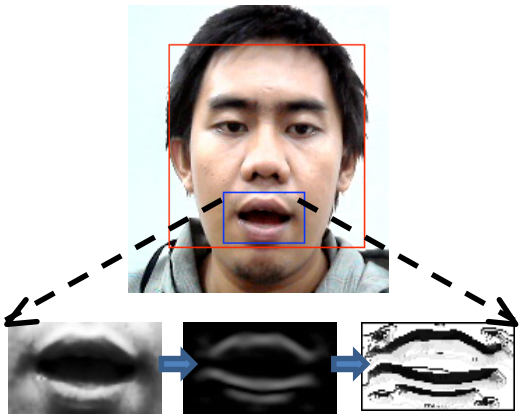


Fig. 2. Mouth feature extraction.

Gabor filters techniques. Edge technique of Sobel and Cann is difficult to use because the color of the edge on the lips i vague and approaching the value of skin color. Constraint in the extraction the edge of the lips is there is noise aroun the lips, it needs to be removed first.

After getting mouth shape feature extraction using Gabor filters, the next step to find the form mouth texture using Local Binary Pattern (LBP) techniques. Feature values of mouth pixels from texture searcher using the LBP taken by integral projections. The size of piece of the mouth image (x & y) it has dimensions of 90 x 50. The amount of data which generated from the integral projection towards the image mouth produced 140 data feature of the mouth.

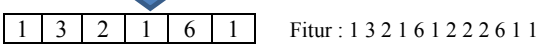
III. Lips Vowel Recognition

Neural Network Backpropagation is neural network with multilayer topology with one input layer (layer X), one or more hidden layers (layer Z) and one output layer (layer Y). Each layer has neurons (units) are modelled with a circle. Among the neurons in one layer with neurons on the next layer connections associated with the model that has the weights, w and v. Hidden layer (Z) may have a bias, which has a weight equal to one [5]. The algorithm used is back propagation algorithm with feed forward architecture.

The purpose of using the neural network back propagation in this system due to the advantages possessed by the network, that is the ability to learn patterns (learning), adaptability, and is useful for situations where there is a lot of data that can not easily be modeled with mathematical equations.

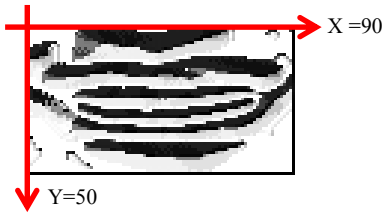
Structure of the neural network that will be carried out using a multilayer neural network architecture. The system architecture consists of 140 nodes on layer input, 20 nodes in the hidden layer, and 5 nodes in the output layer which is a binary combination of 5 vocal patterns to be recognized. The fifth combination is:

- 10000 for Vocal A
- 01000 for Vocal I
- 00100 for Vocal U
- 00010 for Vocal E



Gambar 3.22. Pencarian Nilai Fitur Gambar Menggunakan Integral Proyeksi

Pada sistem ini, gambar mulut yang dintegral proyeksi memiliki dimensi sebesar 90 x 50 (Gambar 3.23). Maka hasil dari integral proyeksi gambar mulut memiliki data sebanyak 140 nilai fitur piksel. 140 nilai tersebut didapat dari hasil penyusunan nilai dari piksel X dengan piksel Y. Tiap piksel citra atau gambar yang diproses memiliki nilai piksel, atau dengan kata lain dalam satu gambar semua piksel memiliki nilai cluster dengan intensitas cluster yang berbeda.



Gambar 3.23. Dimensi ROI Mulut

Fig. 3. Mouth region of interest dimension.

Berikut ini adalah potongan program untuk mengambil nilai fitur piksel horizontal (X) dan vertikal (Y) gambar area mulut.

```
IplImage *img3 = cvLoadImage("LBP.bmp");
file_test = fopen("file_test.txt", "w");
CvScalar pixel;
int i,j;
int jelmp;
f=0;
```

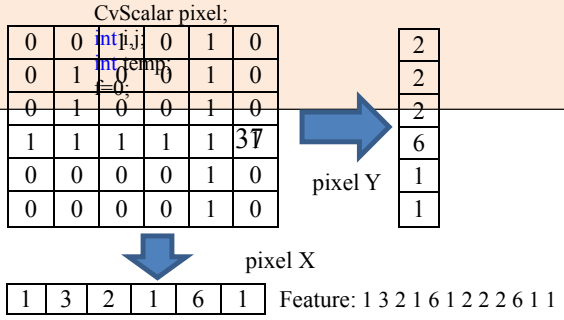


Fig. 4. The sample of integral projection.

System modelling of vowel recognition forms on the lips can be seen in Figure 5. Training on back propagation is divided into three stages, that is: feed forward stage on input training pattern, backpropagation of the associated error and adjustment of the weights.

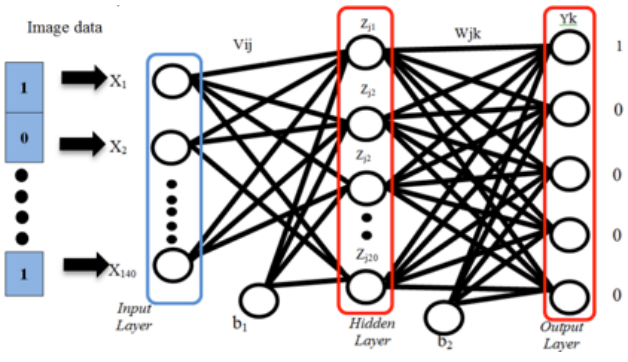


Fig. 5. The proposed neural network architecture in this system.

In Figure 5 can be seen in the input layer of neural network is the value obtained from the integral projection of the data as much as 140. From the 140 data is distributed one by one into the input layer, which amounted to 140 nodes. From input value as much as 140 data with a target output as much as 5 layer node the network will begin to learn the pattern

to meet the desired target output combination that has been specified previously.

In this system, artificial neural networks are trained using data training in the form of images. The number of image that trained as much as 250 images of faces which have different mouth shapes in the pronunciation of the vowels A, I, U, E, O. Image data that will be trained will be sought in advance through the mouth features previously (detection of facial features). Names of the people are made in the training data can be seen in Table I.

TABLE I

NAMES OF THE PEOPLE WHO MADE FOR THE TRAINING DATA.

No	Name
1	Haikal
2	Yani
3	Luthfia
4	Ari
5	Arisandi
6	Tri
7	Rizky
8	Odi
9	Sucahyo
10	Alfan

IV. EXPERIMENTAL RESULTS

First, we showed the experimental result from the feature extraction for area of the face and mouth location as shown in Table II.

The data then processed through image processing then finally using integral projection, as input for neural network backpropagation. After training, it will generate a new weights and bias values which used in testing the Neural Network. For testing the Neural Network using a feed forward architecture. In this test performed experiments in which everyone does does varying pronunciation of vowels. The experiments were performed 2 times to the people whose data are stored on the training data and are not stored as training data in neural network. Pronunciation of vowels form A, I, U, E, O can be seen in Table I and Table III with different people and pronunciation.

In Figure 6 is a representation of Table III. Can be seen in the percentage of lip patterns of Ari and Haikal be recognized well. Meanwhile, at the percentage of Alfan there is a value of 5.16% on the I vowel patterns, however, the percentage is failed because the percentage value dominant towards O vowels.

From the results of the Table IV it can be seen the success of the system to recognize the pattern of the lips vowels are not present in the training data is more dominant against the introduction of vowels I. Meanwhile the A and U vowel systems can hardly recognize the overall vocal form.

In Figure 7 it can be seen that the percentage of Edi lips patterns be could recognized well. Meanwhile the percentage of the Ovin's vowels overall shape is almost cannot be detected properly. To test the accuracy of the system based

TABLE II

THE RESULT OF SUCCESSFUL IN MOUTH VOWEL RECOGNITION.

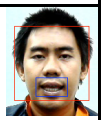
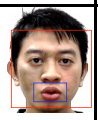
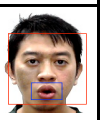
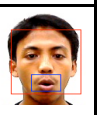
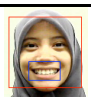
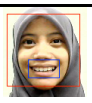
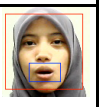
NAME	A	I	U	E	O
HAIKAL					
YANI					
ARSAN					
ALFAN					
FIA					
VINA					

TABLE III

SYSTEM TEST RESULT ON WHICH THE DATA WERE TRAINED.

Name	A(%)	I(%)	U(%)	E(%)	O(%)
Haikal	100	95,40	99,99	99,99	100
Yani	0	99,98	99,99	99,97	99,99
Ari	99,98	99,98	99,99	99,99	99,99
Rizky	100	95,85	0	55,55	99,99
Alfan	99,98	5,16	99,92	99,99	0

on the percentage of success in recognizing the pattern shape of the mouth can be seen in Table V.

Based on Table V can be calculated are not successful the systems in the vowel recognition in general, where the number of experiment in 5 people with 5 vowels pattern in mouth.

Σ Experiment = 10 x 5 = 50
Σ Successful= 35
Σ Failed= 15

So it can be calculated percentage of success and failure as a whole is as follows:

% Success= (35/50) x 100%=70%
% Failed =(1550) x 100%=30%

As for the percentage of the individual - each mouth shape
- Vokal A :

% Success = (610) x100% = 60%
% Failed = (410) x100% = 40%

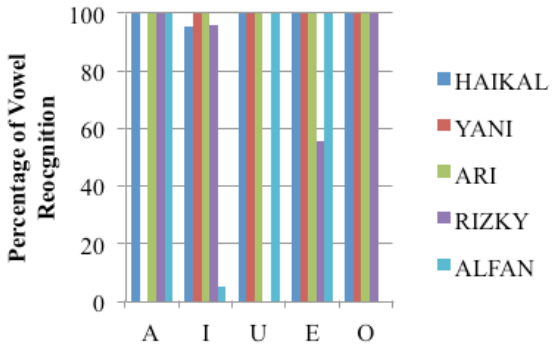


Fig. 6. Percentage chart vowel recognition on experiment 1. The data were trained.

TABLE IV

SYSTEM RESULT ON DATA THAT ARE NOT TRAINED.

Name	A(%)	I(%)	U(%)	E(%)	O(%)
Adi	0	99	100	32	0
Edi	100	19	99	100	0
Vina	99	99	0	2	100
Ovin	0	0	0	6	99
Yudis	0	99	0	97	66

- Vokal I:

% Success = (810 x100% = 80%

% Failed = (210) x100% = 20%

- Vokal U:

% Success = (610) x100% = 60%

% Failed = (4/10) x100% = 20%

- Vokal E:

% Success = (810) x100% = 80%

% Failed = (210) x100% = 20%

- Vokal O:

% Success = (710) x100% = 70%

% Failed = (310) x100% = 30%

Based on these calculations shows that the detection of I and E vowels is better than the other vowel detection. From the Table V can be seen that the results of the failure experiment due not found the right features or determination of feature is incorrect. With the determination of fault features then the resulting value will be tested with existing data if the value is close to the value that has been trained the neural network will refer the decision to the data which have the closeness from the data test. If no value is close to the data so the decision of the recognition not recognized.

Results of the neural network output is sent via RS-232 serial PC to the Arduino. The output data from the vocal value A, I, U, E, O is sent sequentially in formation format. The formation data used is "K A I U E O", in which the character "K" is the data protocol header followed by data content A I U E O. This formation is used for uniformity for easy data processing in OpenGL simulator and Arduino. Seen in Figure 8 serial communication successfully performed using hyper terminal.

Feature values have been obtained and test them using

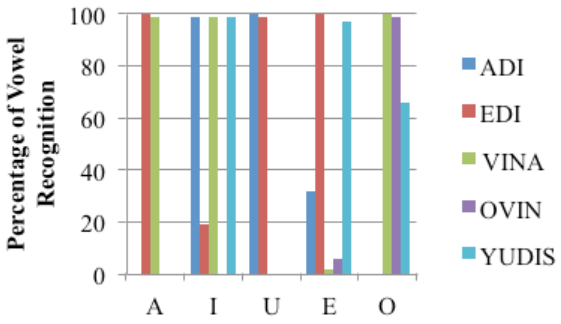


Fig. 7. Percentage chart vowel recognition on experiment 2. The data were not trained.

TABLE V

THE RESULT OF SUCCESSFUL IN MOUTH VOWEL RECOGNITION.

Name	A	I	U	E	O
Haikal	√	√	√	√	√
Yani	x	√	√	√	√
Ari	√	√	√	√	√
Rizky	√	√	x	√	√
Alfan	√	x	√	√	x
Adi	x	√	√	√	x
Edi	√	√	√	√	x
Vina	√	√	x	x	√
Ovin	x	x	x	x	√
Yudis	x	√	x	√	√

artificial neural network simulator is validated using OpenGL shown in Figure 9.

V. CONCLUSIONS

We proposed vowel recognition system of lipsynchrobot in lips gesture using neural network backpropagation. The proposed method works well. According to the test result program has been carried out, it can be concluded that, the successful of vowel recognition in the form of the mouth with the amount of training data is 250 data and testing to 10 people is 70% with pattern recognition of I and E vowels better than the other vowel detection.

The number of neurons in the hidden layer and the input number of slightly variables or too accurate model for the data being trained only produces good output data to be trained for it, but could not produce a good output for data that is not included in the training data.

For future works, from the output of OpenGL system, we'd like to implement into the EEPIS Lipsynchrobot.

REFERENCES

[1] T. Fukuda, M.-J. Jung, M. Nakashima, F. Arai, and Y. Hagesawa., "Facial Expressive Robotic Head System for Human-Robot Communication and Its Application in Home Environment", Proc. IEEE, Vol. 92, No. 11, 2004.

[2] Official Web site of Tamagotchi (in Japanese). Available: http://tamagotch.channel.or.jp, Accessed: tanggal 19 Juni 2012.

[3] Lee, Sung Hui, "A Linear Dynamic Affect-Expression Model: Facial Expressions According to Perceived Emotions in Mascot-Type Facial Robots", IEEE International Robot & Human Interactive Communication (RO-MAN-06), Korea, August 26-29, 2006.

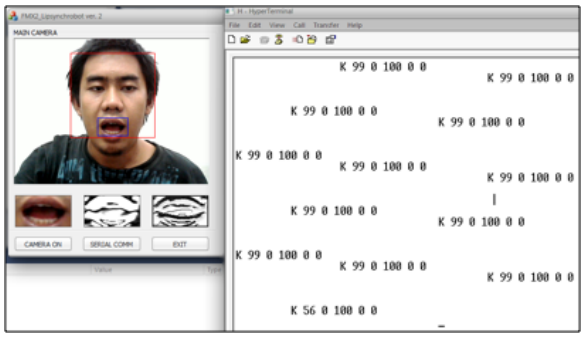


Fig. 8. Sending data format.

[4] Ahonen, T., Abdenour Hadid, and Matti Pietikäinen, "Face Description with Local Binary Pattern: Application to Face Recognition", vol. 28 no. 12, pp. 2037-2041, 2006.

[5] Fausett, Laurent., Fundamentals of Neural Networks: Architectures, Algorithms, and Applications, Prentice-Hall, Inc., New Jersey, 1994.

[6] Naoyuki Kubota and Indra Adji Sulistijono, Evolutionary Robot Vision for People Tracking Based on Local Clustering, 2008 World Automation Congress (WAC2008), Proceeding (CD ROM) of the 6th International Forum on Multimedia and Image Processing (IFMIP2008), Hawaii, USA, 28 September - 2 October, 2008.

[7] Salah Werda, Walid Mahdi and Abdelmajid Ben Hamadou, Lip Localization and Viseme Classification for Visual Speech Recognition, International Journal of Computing & Information Sciences Vol.5, No.1, April 2007, pp.62-75.

[8] Satoh T, Haisagi M, Konishi R, Analysis of Features for Efficient Japanese Vowel Recognition, IEICE Transaction Information and Systems, Vol. E90-D, No.11, 2007, pp.1889-1891.

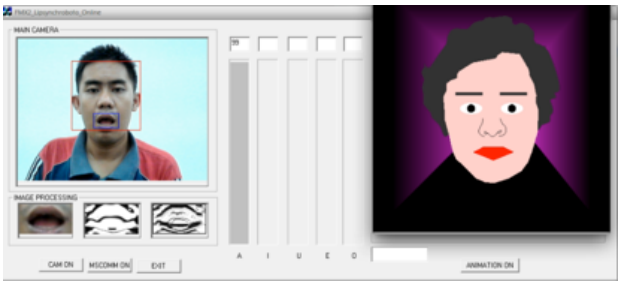
[9] Nakamura S, Kawamura T, Sugahara K, Vowel Recognition System by Lip-Reading Method Using Active Contour Models and its Hardware Realization, SICE-ICASE International Joint Conference, Busan, Korea, 2006, pp.1143-1146.

[10] Kurniawan, Dwi Ely., Face Recoginition Using Gabor Filter, Thesis, University of Diponegoro, Semarang, Indonesia, 2012.

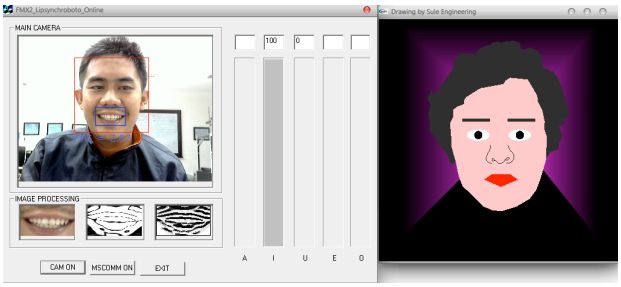
[11] Valerina, Fani, Comparison Local Binary Pattern and Fuzzy Local Binary Pattern for Image extraction on Medicine Plant, Thesis, Institute of Agriculture Bogor, 2012.

[12] Indra Adji Sulistijono, Zaqiatud Darojah, Abdurahman Dwijotomo, Dadet Pramadihanto and Naoyuki Kubota, Facial Expression Recognition Using Back Propagation (in English), Proceeding (CD ROM) of The 2010 International Symposium on Intelligent Systems (iFAN 2010), Tokyo, Japan, Sep 24-25, 2010, paperID:569.

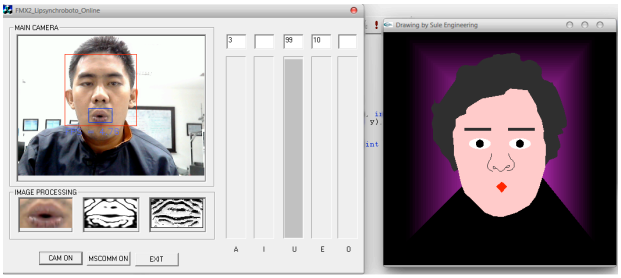
[13] Indra Adji Sulistijono, Dadet Pramadihanto, Abdurahman Dwijotomo, Rachman Ardyansyah and Naoyuki Kubota, Real Time Facial Expression Simulator Using OpenGL, Proceeding of the 2nd International Workshop on Advanced Computational Intelligence and Intelligent Informatics (IWACIII 2011), Suzhou-Jiangsu, China, Nov 19-23, 2011.



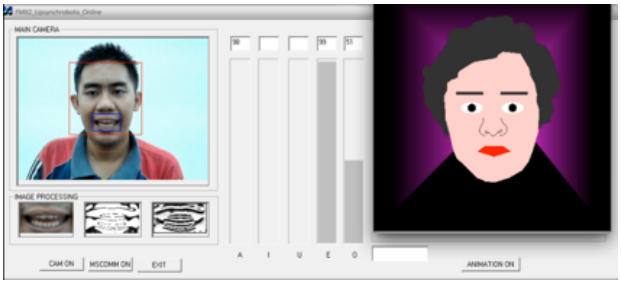
(a) "A" Vocal



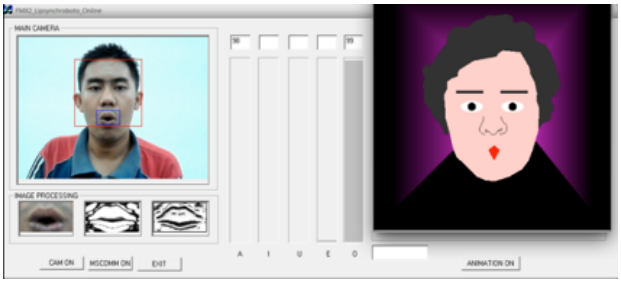
(b) "I" Vocal



(c) "U" Vocal



(d) "E" Vocal



(e) "O" Vocal

Fig. 9. Validation results of pattern recognition on simulator OpenGL: (a) "A" Vocal; (b) "I" Vocal; (c) "U" Vocal; (d) "E" Vocal; (e) "O" Vocal.