TELKOMNIKA, Vol.13, No.1, March 2015, pp. 173~180
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v13i1.1302

Lip Motion Pattern Recognition for Indonesian Syllable Pronunciation Utilizing Hidden Markov Model Method

Balza Achmad, Faridah, Laras Fadillah
Department of Engineering Physics, Faculty of Engineering, Universitas Gadjah Mada
Jalan Grafika 2, Yogyakarta Indonesia
e-mail: balzach@ugm.ac.id, faridah@ugm.ac.id

Abstract

A speech therapeutic tool has been developed to help Indonesian deaf kids learn how to pronounce words correctly. The applied technique utilized lip movement frames captured by a camera and inputted them in to a pattern recognition module which can differentiate between different vowel phonemes pronunciation in Indonesian language. In this paper, we used one dimensional Hidden Markov Model (HMM) method for pattern recognition module. The feature used for the training and test data were composed of six key-points of 20 sequential frames representing certain phonemes. Seventeen Indonesian phonemes were chosen from the words usually used by deaf kid special school teachers for speech therapy. The results showed that the recognition rates varied on different phonemes articulation, ie. 78% for bilabial/palatal phonemes and 63% for palatal only phonemes. The condition of the lips also had effect on the result, where female with red lips has 0.77 correlation coefficient, compare to 0.68 for pale lips and 0.38 for male with mustaches.

Keywords*: hidden markov model, lip motion, pattern recognition, syllable pronounciacion*

1. Introduction

Visual communication plays important role in noisy environment as well as for hearing impaired person, which audio communication is not possible. Many researchers have developing methods to overcome these problems; one of them is by lip-reading. Lip movements during prononciation syllables or words will form specific patterns. From these lip patterns, we can find out what was said by other people without hearing his/her voice.

Image processing and pattern recognition fileds have been growing very fast, allowing us to establish a system for automatic lip reading. Petajan [1] suggested that visual system will help speech recognition processes become more effective. Yau et al [2] introduced a technique of speech recognition combined with a visual speech model based on facial movement video. Ma et al [3] developed a Bayesian model for lip-reading patterns under moderate noise exposure. Meanwhile, along with the development of communication tools, Kim et al [4] developed a method of lip-reading in a real time fashion for smart phones. Lip-reading method applied to several languages was developed by Saitoh et al [5], while Shin et al [6] developed this system for Korean language.

One focus of research in lip reading is on the selection pattern recognition method. How et al [7] performed lip-reading on syllables /ba/, /da/, /fa/, /la/, /ma/ using Artificial Neural Network (ANN) on video, audio, the combination of both. Another widely used method is the method of Hidden Markov Model (HMM). Puviarasan et al [8] used HMM method to recognize 33 words in English by people with hearing impairments. Two features are used, namely discrete cosine transform (DCT) and discrete wavelet transform (DWT), with recognition rate of 91% and 97% respectively. Nursing [9] conducted research on lip-reading of three French syllables using HMM and our-point method for the feature extraction; namely: the point above, bottom, right, and left. The system can read correctly the syllables /ba/ 63.64% , /be/ 72.73% and /bou/ 81.82% .

Lip-reading system for Indonesian phoneme itself has been developed by Faridah et al [10] and applied as a speech therapeutic tool for deaf kids in Indonesia. The system used Neural Network for lip pattern recognition in pronuncing vowel phonemes: /a/, /i/, /u/, /e/, and /o/ in Indonesian language. However, the system was not able to provide satisfactory results.

In this paper, we use HMM method for lip pattern recognition in pronunciation of phonemes in Indonesian language. The phonemes to be recognized are representing three

Received October 11, 2014; Revised December 8, 2014; Accepted December 28, 2014

different consonant sound formation, namely bilabial (lips consonants), dental (dental consonants), and palate (hard palate consonant).

2. Research Method

2.1. The Data

The data used in this research is in the form of facial video of 25 speakers, comprising variety of speakers, namely female with red lips, female with pale lips, male with a mustache and pale lips, as shown in Figure 1. Each speaker pronounciated 17 phonemes composed of bilabial, palatal, dental as well as mixed consonants, as presented in Table 1. The videos were taken under 240 -270 lux lighting. Of the 25 video data, 15 were used as HMM training data and 10 were used for testing.

(a) **(b)** **(c)**

Figure 1. Examples of facial image of (a) female with red lip, (b) female with pale lip, and (c) male with mustache and pale lip

Tabel 1. Modeled phonemes

Bilabial	*Palatal*	*Dental*	Mixed
$^{(1)}$Ba- $^{(2)}$Bi-$^{(3)}$Be-$^{(4)}$Bo	$^{(10)}$Sa	$^{(13)}$La	$^{(15)}$Cak
$^{(5)}$Ma-$^{(6)}$Me	$^{(11)}$Ja	$^{(14)}$Ta	$^{(16)}$Dak
$^{(7)}$Pa-$^{(8)}$Pi-$^{(9)}$Pu	$^{(12)}$Ca		$^{(17)}$Tol

Note : The number in parentheses are the index of the fonem for recognition

2.2. Video Image Processing

The data from the video are processed to produce feature extraction using steps illustrated in Figure 2.

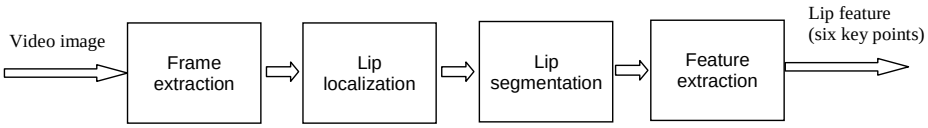


Figure 2. Video image processing block diagram

For the first step, 20 frames are extracted from the video during the prononciation of the phonemes (Figure 3). Each frame is then undergone a series of image processing to obtain lip feature for pattern recognition. The first image processing is performed to detect the location of the lip. The basic method used in the detection of lip location is Cascade Classifier method. Once the location of the lip is obtained, the image area around the lip is then cropped, hence we can focus on smaller area of image.

Figure 3. Image frames during pronounciation of phonemes

The cropped lip image is then segmented hence the lip can be separated from the surrounding facial skin. Lip segmentation algorithm is performed based on the difference in color composition between the lip and the skin [11]. Skin color is determined more on color composition compare to brightness. Color composition of the skin is remarkably constant even exposed by different illumination. Hulbert and Poggio [12] defines the value of the pseudo hue to illustrate this difference,

$$h(x,y) = \frac{R(x,y)}{R(x,y)+G(x,y)} \quad (1)$$

with R(x,y) and G(x,y) are the red and green component for each pixel in the image.

Snakes method, which was first developed by Kass et al [13], is then applied to the pseudo hue image to obtain the outer contour of the lip. In this paper, lip movement pattern are constructed by the lip shapes of each frames. The lip shape itself is represented by six key points on the lip contour, as shown in Figure 4.

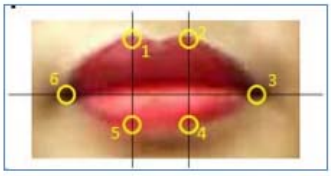


Figure 4. Six key points representing lip movement pattern

In order to find these six key points, the contour obtained by the Snake method is evaluated, in which six points are selected. Slope normalization sometime is necessary in the case of lips that are not upright, which may occure during the movement of the lip from frame to frame. An example of complete image processing for each frame is given in Figure 5.

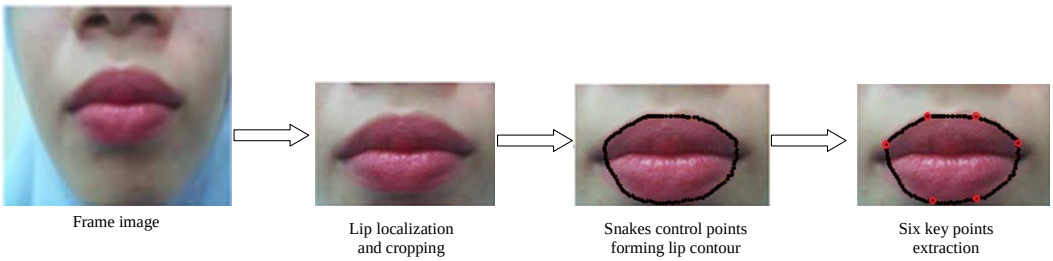


Figure 5. Complete image processing for each frame

2.3. Hidden Markov Model

Hidden Markov Model (HMM) is a statistical model of a system on which hidden parameters are determined from observable parametes. The observable parameters are used as inputs of the HMM in the form of a database. The architecture of the input database is shown in Figure 6.

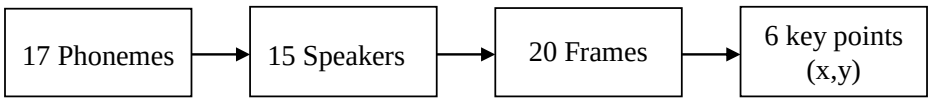


Figure 6. Database Architecture of the HMM

The HMM architecture for this study is given in Figure 7. In this study, we use One Dimentional Hidden Markov Model [14], in which the probability of transition from all state to observable parameter are the same.

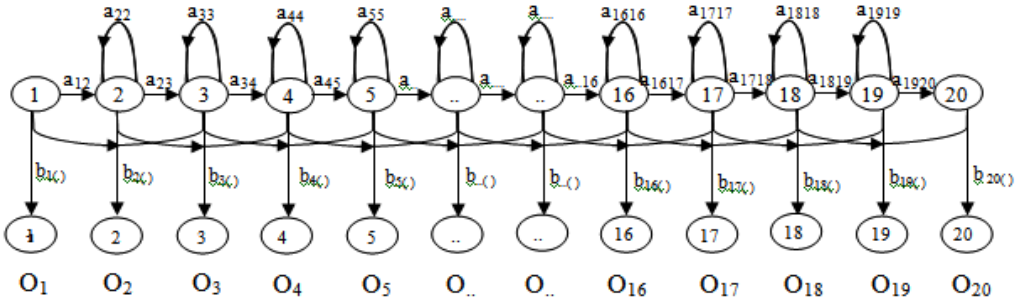


Figure 7. The Hidden Markov Model architecture in this study

The HMM modeling stages, commonly referred to as database construction, is shown Figure 8. The parameters in the database will be used as a benchmark in recognition process during testing phase.

There are three main stages in the construction of model database.

1. Labelling, is the process of making a label for each test data file consisting lip feature, which are composed of six key points of 20 frames representing certain phonemes.
2. Codebook formation, which is used to store input-output data pairs for training. In this paper, the input of six key points in 20 frames are represented by the Euclidean distance.
3. HMM model construction, which calculates the HMM parameter using 15 training data. This process will generate 17 models according to the number of trained phonemes.

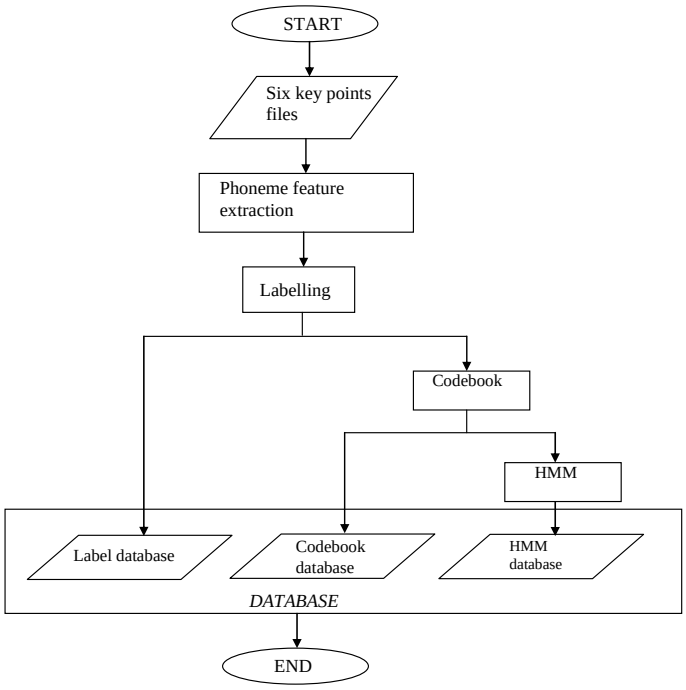


Figure 8. Flow chart of database construction

2.4. HMM Testing

In this stage, the constructed models are tested using 10 test data. The steps in this stage can be seen in Figure 9. The general idea of HMM testing is to find phoneme, represented by label, which has maximum log of probability when applied to the HMM model.

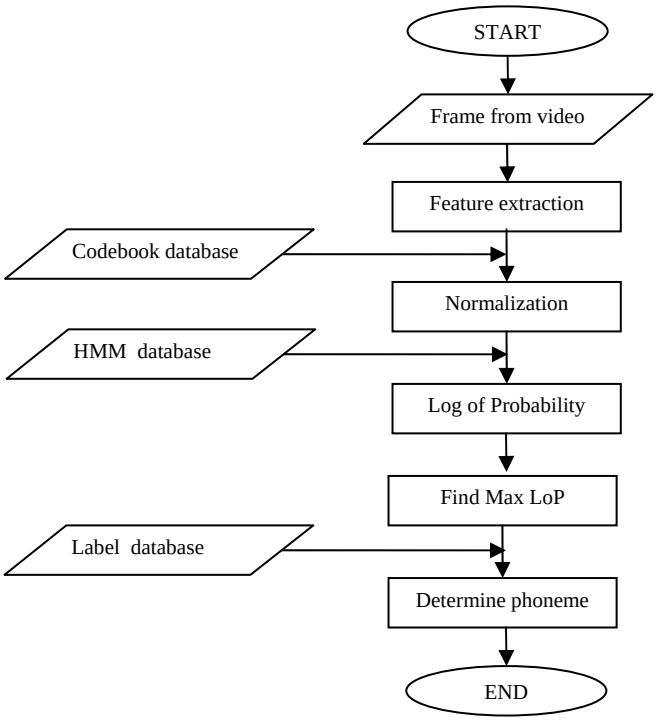
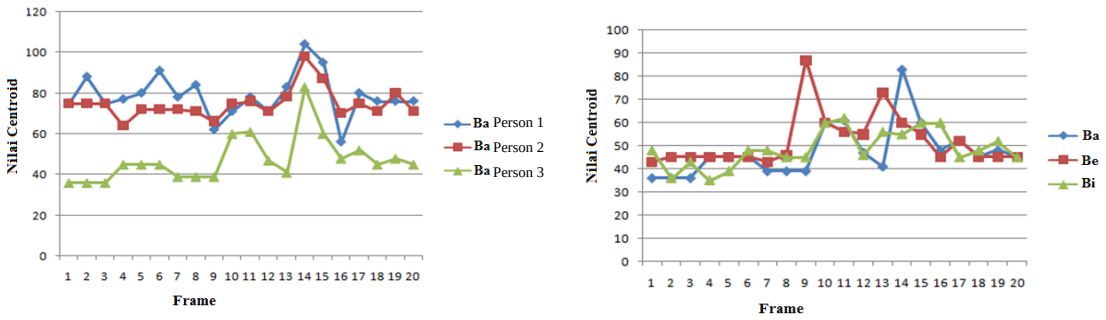


Figure 9. Flow chart of HMM Testing

The performance of the model has been carried out for video inputs consisting 17 Indonesian phonemes. The analysis will be performed to determine the similarity between the centroid for different people pronouncing the same phonemes. Another analysis will be performed to calculate successful rate in recognizing test data.

3. Results and Analysis

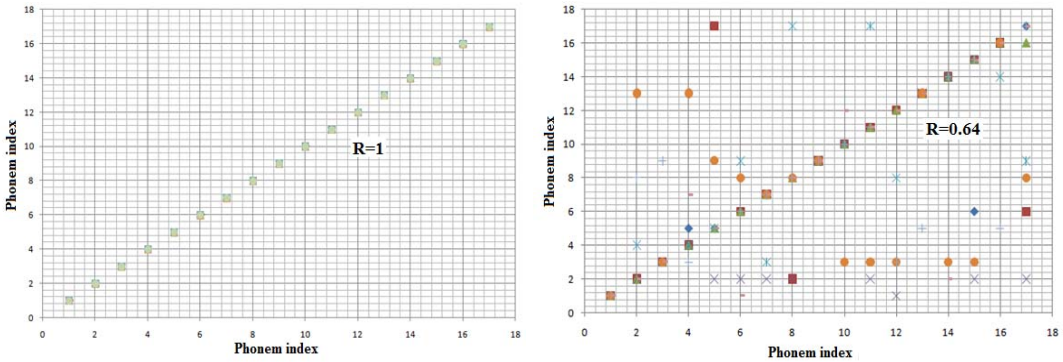
The test shows that there are similarities in the value of centroid formed by 20 frames when pronouncing the same phoneme by different speakers, as shown in Figure 10a. On the other hand, each person move their lips differently while pronouncing different phonemes, as shown in Figure 10b.



(a) (b)

Figure 10. Centroid pattern of phoneme (a) /ba/ by three different person (b) /ba/, /be/, dan /bi/ by the same person

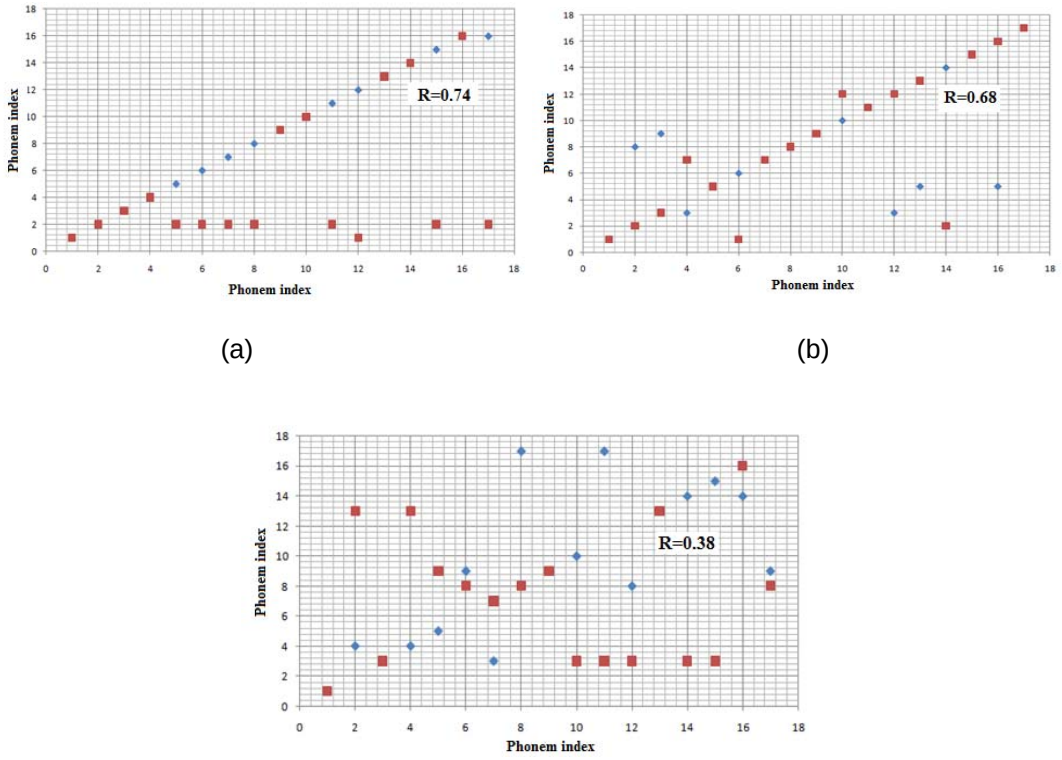
Training process done for 15 training data for each phoneme creates 17 HMM models. Each model has specific characteristics represented by transmission and emission matrices. Figure 11 shows test result of the models using 15 training data and 10 testing data. Testing using training data gives correlation coefficient R = 1, means that 100% training data can be recognized perfectly. Meanwhile, when applying the models using testing data, the correlation coefficient R = 0.64. The source of error can be the lip conditions, articulation, as well as external factors such as lighting and positional changes during video recording which was not controlled in this study.



(a) (b)

Figure 11. Test results of applying the HMM models to (a) traing data (b) test data

The relation between input and output of the HMM models by varying lip conditions can be seen in Figure 12. The correlation coefficent for red lips, pale lips, and pale lips with mustaches are 0.74, 0.68 and 0.38 respectivelly. Thus, the HMM models can recognize correct phonemes for female with red lips compare to other lips conditions.

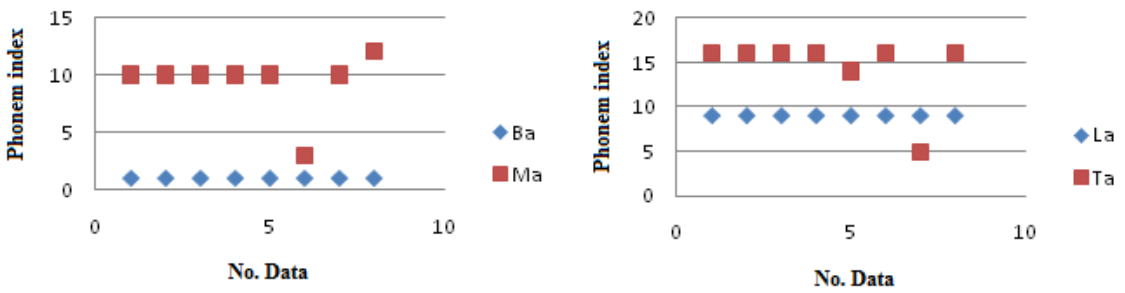


(a) (b)

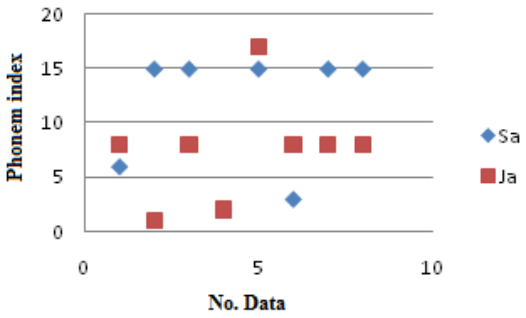
(c)

Figure 12. Test results for (a) red lips (b) pale lips (c) pale lips with mustaches

Articulation can also be one of the error sources in recognizing phonemes using HMM models. Figure 13 shows the recognition results of different phonemes articulation, while pronouncing bilabial, palatal and dental phonemes. The success rates for bilabial and dental phonemes are 78%, while for palatal is 63%. This is because when pronouncing different bilabial and dental phonemes, the lip movements are differently, while for palatal phonemes, all phonemes produce similar lip movement.



(a) (b)



(c)

Figure 13. Test results for (a) bilabial (b) dental (c) palatal phonemes

4. Conclusion

1. The developed HMM models has performance in term of correlation coefficient of R = 1 for training data and 0.64 for test data.
2. Phonemes pronounced by female with red lips can be recognized better with R = 0.77, compare to pale lips with R = 0.68 and pale lip with mustache R = 0.38.
3. Articulation in pronouncing phonemes also has effect in recognition rate, bilabial and palatal phonemes has 78% recognition rate, while palatal only gives 63% recognition rate.

References

[1] Petajan ED. *Automatic Lipreading to Enhance Speech Recognition*. IEEE Conference on Computer Vision and Pattern Recognition. San Fransisco. 1985: 40-47.
[2] Yau WC, Kumar DK, Arjunan SP. Visual Recognition of Speech Consonants using Facial Movement Features. *Integrated Computer-Aided Engineering*. 2007; 14(1): 49-61.
[3] Ma WJ, Zhou X, Ross LA, Foxe JJ, Parra LC. Lip Reading Aids Word Recognition Most in Moderate Noise: A Bayesian Explanation Using High-Dimensional Feature Space. *PLoS ONE*. 2009; 4(3): 1-14.
[4] Kim Y, Kang S, Jung S. *Design and Implementation of A Lip Reading System in Smart Phone Environment*. 10th IEEE International Conference on Information Reuse and Integration, Las Vegas. 2009: 101-104.
[5] Saitoh T, Morishita K, Konishi R. *Analysis of Efficient Lip Reading Method for Various Languages*. 19th International Conference on Pattern Recognition (ICPR), Tampa. 2008: 1-4.
[6] Shin J, Lee J, Kim D. Real-time Lip Reading System for Isolated Korean Word Recognition. *Pattern Recognition*. 2011; 44: 559-571.
[7] Bagai A, Gandhi H, Goyal R, Kohli M, Prasad TV. Lip Reading Using Neural Networks. *International Journal of Computer Science and Network Security*. 2009; 4: 108-111.
[8] Werda S, Mahdi W, Hamadou AB. Lip Localization and Viseme Classification for Visual Speech Recognition. *International Journal of Computing and Information Science*. 2007; 5(1): 62-75.
[9] Puviarasan N, Palanivel S. Lip Reading of Hearing Impaired Persons using HMM. *Expert Systems with Applications*. 2011; 38: 4477-4481.
[10] Faridah, Utami SS, Wibowo S, Wijaya E. Speech Therapy Instrument for Deaf People in Indonesia. *Media Teknik.* 2008; 30(2): 201-206.
[11] Eveno N, Caplier A, Coulon PY. *A New Color Transformation for Lips Segmentation*. IEEE Workshop on Multimedia Signal Processing (MMSP'01). Cannes. 2001: 3-8.
[12] Hulbert A, Poggio T. Synthesizing A Colour Algorithm from Examples. *Science.* 1998; 239: 482-485.
[13] Kass M, Witkin A, Terzopoulos D. Snakes: Active Contour Model. *Int. Journal Computer Vision*. 1988; 4: 321-331.
[14] Rabiner LR. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the IEEE. 1989; 77(2): 257-286.