Active Lip Localization Based on Lip Movements Recognition Using YCbCr and CIELa*b* Color Space

Thein Thein, Kalyar Myo San

Abstract— Automatic lip reading was used for various purposes for speech training to facilitate hearing and improve speech recognition. Nevertheless, the recognition of the movement of the lips actively passes through research topics with a lot of improvements. Therefore, in order to extract visual information, reliable movements of the lips are required. The major challenge is to localize the movements of the lips accurately because of the many possible movements of the lips and forms of the lips. The accuracy and reliability of speech recognition systems can be better by using visual information from lip movements, and the need for the lip reading system continues to develop for each language. In lip reading system, lip localization is the major step to read the lips for extracting visual information from the video input. This paper presents the lip localization method for Myanmar consonants recognition based on lip movements using the combination of YCbCr and CIELa*b* color space and Moore Neighbor contour tracing algorithm for localization. The proposed method shows how accurate localizing and lip tracking are useful for speech recognition. The experimental results show the automatic lip localizing the lip shape for Myanmar consonants using the only visual information from lip movements which is useful for visual speech of Myanmar languages.

Index Terms— CIELa*b* color space, Lip Localization, Lip Movement, Lip Reading, YCbCr color space, Moore Neighbor contour tracing algorithm.

1 Introduction

MANY researchers have demonstrated that useful information about the speech content can be obtained through the lip of speakers [1, 2]. Lip reading is a technique of understanding speech by visually interpreting the movements of lips. Lip localization plays a vital roles in lip reading system. If the localizing the lip boundary based on lip movements is not accurate, it directly affects the lip tracking, feature extraction of lip movements and recognition accuracy for speech recognition system. However, lip movement recognition is active undergoing research topics with a lot of improvements that recovers various difficulties faced in the research. Many researchers proposed various techniques to precisely identify the lip area and they presented various features to get significant recognition result.

The purpose of this study is to propose a visual teaching method for Myanmar Consonants Recognition for the hearing impaired by precisely localizing lip movement when the speaker produces the Consonants. Moreover, this studies aims to the extraction and localization. In this paper, we propose approaches to localize accurate lip boundary based on lip movements for Myanmar Consonants recognition. Myanmar consonants can be described in terms of four factors: (1) One syllable consonants, (2) Two syllable consonants, (3) Three syllable consonants, and (4) Four syllable consonant. Here we have tried to extract accurate lip movements for one syllable consonants (င (Nga) ၊ ည (Nya) ၊ မ (Ma) ၊ လ (La) ၊ ဝ (Wa) ၊ သ (Tha) ၊ ဟ (Ha) ၊ အ (Ah)) and two syllable consonants (က (Ka Gyi) ၊ ခ (Kha Gway) ၊ ဂ (Ga Nge) ၊ ဃ (Ga Gyi) ၊ စ (Sa Lone) ၊ ဆ (Sa Lain) ၊ ဇ (Za Gwe) ၊ ဒ (Da Dway) ၊ ဏ (Na Gyi) ၊ န (Na Nge) ၊ ပ (Pa Saug) ၊ ဘ (Ba Gone) ၊ ရ (Ya Gaug) ၊ ဠ (La Gyi)) of Myanmar language.

The proposed article is organized as follows: section 2 presents the related works, section 3 goes into the method for the proposed lip reading system. The experimental results will be mentioned in section 4, conclusion and future work presents in section 5. Finally, reference will be done.

2 Related Works

Many researchers proposed various techniques for lip localization. There are several techniques used to localize the lips. They use different types of inputs and provide a lip output with a different level of accuracy.Kji lwano et al. [3] are applied on side view of the face to localize the lips. Luca Lombardi et al., Namrata Dave, N. Otsu [5, 12, 13] are also applied on front view of the face from which lip is localized. Vicente P. Minotto et al. [4] implemented a novel color based approach for lip localization based visual feature extraction method which gave a good accuracy for their database.

Faridah et al. [14] proposed a robust lip tracking algorithms using localized color active contours and deformable models. They used a combined semi-ellipse as the initial evolving curve and compute the localized energies in color space to separate from the original lip image into lip and non-lip regions. And then, they presented dynamic radius selection of the local region with a 16-point deformable model to extract the lip. Meng Li and Yiu-ming Cheung [8] developed a robust still image lip localization algorithm designed as a visual front end of a practical AVASR system and presented Gabor filter based facial feature extraction for lip localization.

Research efforts were concentrated in the localization of the lip limit for segmentation of the lips. Many studies have shown that color information of particular region can be used to identify the skin or face in digital images. The main idea

behind this approach is to separate from other details in given image, so that segmentation of mouth can be done efficiently. T. Coianiz et al. [7] used the hue component of the HSV representation to highlight the red color which is assumed to be associated with the lips in the image. Later, Alan Wee-Chung Liewa et al. [9] used the HSV color space for lip detection. They used the hue signal to locate the position of the lips in mouth region. N. Eveno et al. [11] proposed a new color mixture and chromatic transformation for lip segmentation in 2001. In their approach, a new method for transformation of the RGB color space and a chromatic map was applied to separate out the lips and facial skin. They argued that their proposed approach is able to provide robust lip detection under variable lighting conditions. Stefan Badura and Michal Mokrys [10] proposed another method for mouth segmentation in 2003. In their approach, they used new transformation method to convert given color image into the CIE-Lab color space and CIE-LUV color space, and then they calculated a lip membership map using the fuzzy clustering algorithm. The ROI around the mouth can be identified from the face area after application of morphological filtering on given image.

3 The Propose Method For Lip Localization

The proposed lip reading system composed of three subsystems. The first one is lip localization system, which localizes the lips in the digital inputs, the next one is the feature extraction system which extracts features of lip movement suitable for visual speech recognition, and the final one is the classification system. In this work study will be carried out to localize upper and lower lip boundary which are useful for recognize lip movements.

Fig. 1 illustrated the detail processing stages for the proposed lip reading system. In this paper, we mainly focus on lip localization to extract accurate lip boundary based on lip movements. There are two steps to localize the lip for automatic speech recognition, namely lip segmentation, lip contour extraction and lip contour tracking to localize upper and lower lip boundary.

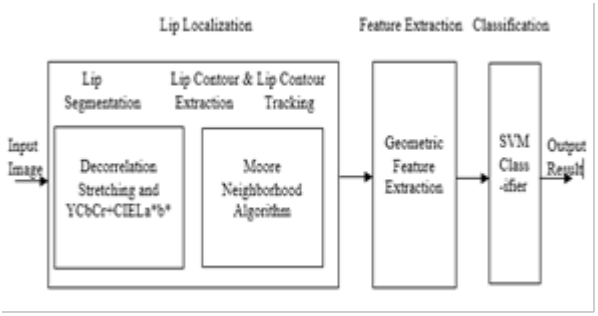

Fig. 1. Schematic diagram for the proposed lip reading system.

- *Thein Thein, University of Computer Studies, Mandalay (UCSM), Myanmar, theinthein.cmw@gmail.com*
- *Kalyar Myo San, University of Computer Studies, Mandalay (UCSM), Myanmar, kalyar.myosan@gmail.com*

Lip localization is the important step of lip reading system to detect and extract accurate lip boundary. There are three kinds of models for localizing the lips. The first one is low level or image based that uses mouth region of the image to localize the lips, features of lips and skin pixels are used. It is finding out the height and width of the lips, not the edges of the lips to locate lips. The second one is high level or model based that uses integrity constraints and pixel information to segment the lip. It is finding the corner of the lips to detect the accurate lips. The final one is the hybrid model which is using the parameters of both the models. In this paper, the hybrid model is used to localize the lip boundary.

3.1 Lip Segmetation

Lip segmentation aims to separate the lip from the background skin color. In Visual Speech Recognition (VSR), facial animation, automatic face recognition and lip reading system, the lip segmentation become an important issue. In these systems, face detection and lip region extraction is normally done by Region of Interest (ROI) detection procedure in each frame. Meng Li and Yiu-ming Cheung [8] proposed a new approach to automatic lip segmentation via a probability model in color space and morphological filter. They used hue and saturation value of each pixel within the lip segment to estimate the model parameter. P.Sujatha et al. [6] presented a new method for automatic lip detection using geometric projection method and adaptive thresholding. By using adaptive thresholding techniques, the performance of the lip tracking method is evaluated.

In this paper, lip segmentation method needs to be performed before the contour extraction process. The proposed system starts frame normalization by breaking the video image as a preprocessing stage. One frame of lip shape only changes a little compared with the neighboring frame in a given of lip motion sequence frames. And then, to segment lip region, template matching method is used to extract required lip region from the face area and we employed 5 × 5 medium filtering method is used to reduce the soft paper noise effect. Fig. 2(a) shows the original images and Fig. 2(b) shows the results image after extracting the lip region. Fig. 3 and Fig. 4 show the normalized image frames. Decorrelation Stretching color enhancing method, YCbCr color space and CIELa*b* color transformation method are based on differences in color composition between lip marked more on color composition compare to brightness, even on different people. Color compositions of skins are remarkably constant even when exposed by a lot of illumination.

(a)

(b)

Fig. 2. (a) Original images of four speakers, (b) Extracted lip region.

Fig. 3. (a) to (j) Sample number of selected frames for utterance of ಅ (Ah) (one syllable consonant).

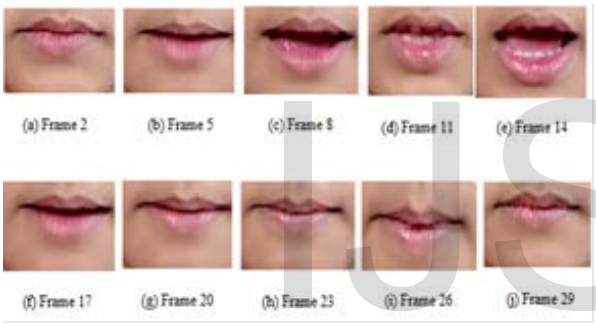

Fig. 4. (a) to (j) Sample number of selected frames for utterance of ಂ (Ga Nge) (two syllable consonant).

The image lip information is in RGB color space. To subtract the robust lip region, RGB color scheme of the image is not improper for immediate processing because it contains a lot of mixed information about lightness etc. Firstly, enhancing color image by using decorrelation stretching method with stretch limit in our experiments. Secondly, RGB color image is transform in Cr layer of YCbCr color space and then RGB color image is transform into CIEL*a*b* color space based on first layer L channel. Finally, color image is contrast by using histogram equalization to extract lip region exactly and accurately. Fig. 5(a) shows the results of enhanced images. Fig. 5(b), Fig. 5(c) and Fig. 6 are the results of color transformed images.

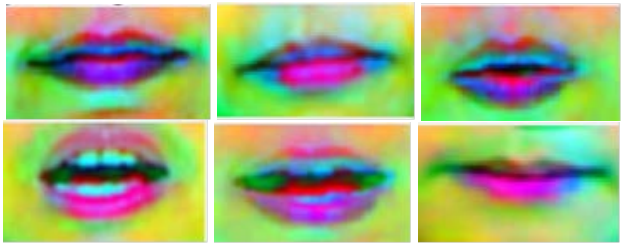

Fig. 5. (a) Results of color enhanced images on different lip shape of different speakers.

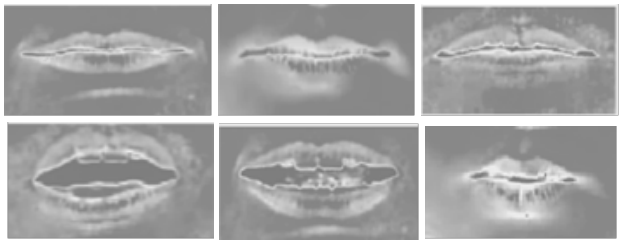

Fig. 5. (b) YCbCr color transformed images on different lip shape of different speakers.

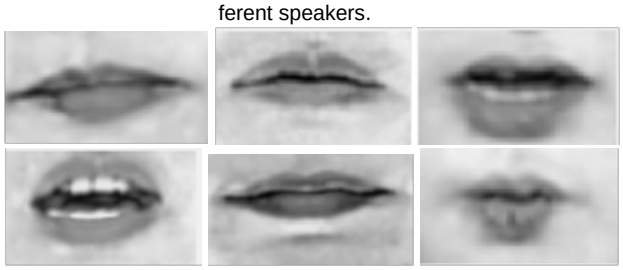

Fig. 5. (c) La*b* color transformed images on different lip shape of different speakers.

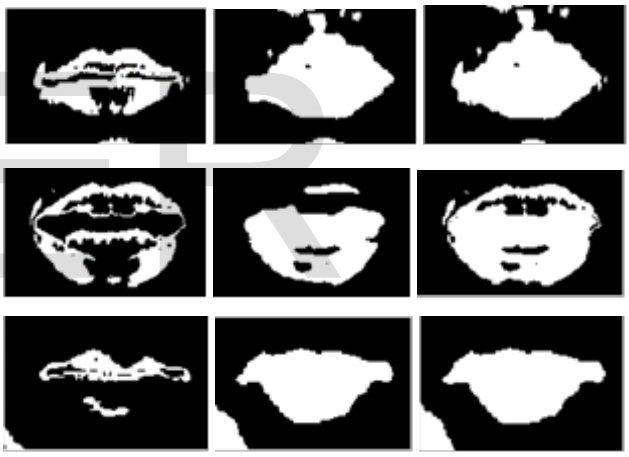

Fig. 6. The first column (a) shows color transformed images for Cr layer of YCbCr, the second column (b) shows color transformed images for L channel of Lab, and the third column (c) shows the combination of YCbCr (Cr layer) and Lab (L channel).

3.2 Lip Contour Extraction and Tracking

Lip contour extraction is important to the lip tracking system. Our proposed lip localization approach is to proceed first step by extracting upper and lower lips boundary and Region of Interest (ROI) using histogram equalization. The process of adjusting intensity values can be done automatically using histogram equalization. Histogram equalization involves transforming the intensity values so that the histogram of the output image approximately matches a specified histogram. For better lip localization, histogram equalization is used to color contrast. Therefore, the lip area appears much darker than the skin. Fig. 7 shows the results of equalized image.

In second step, contour extraction and tracing lip boundary on these ROI throughout the speech sequence and extracting of precise and pertinent visual features from the speaker's lip

region based on lip movement by using Moore Neighborhood Tracing Algorithm.

Speech is signal in time, for lips reading video sequences presents this signal, and the whole task is executed on series of consecutive images. Therefore it is necessary to design effective algorithms, because of large amount of data. Information about lips position in each image frame can enhance the effectiveness of lips reading [10]. Generally two basic approaches can be used for lip contour extraction and tracking:

- Tracing and localizing lips boundary in each frame.
- Tracking Upper and Lower lips regions over all frames.

Localizing and tracing for lip boundary process is difficult and not always efficient. Therefore in our model we propose lips tracing algorithm. There are four of the most common contour tracing algorithms, namely: the Square Tracing algorithm, Moore-Neighborhood Tracing Algorithm, Radial Sweep and Theo Pavlidis' Algorithm. The first two, are easy to implement and are therefore used frequently to trace the contour of a given pattern. In this paper, we used Moore Neighbor Algorithm (MNT) to extract and track the lip contour on the lip boundary. These algorithm consists of two phases: (1) lip contour extraction for the first lip frame, (2) lip tracking in the subsequent lip frames.

In this paper, we extract twelve coordinate points both in upper and lower lip boundary. These feature points are taken from the contour extraction step by Moore Neighborhood Tracing Algorithm which gives proper the twelve points feature. These points serve as initialization points for tracking. With this approach, we are able to estimate lip boundary in each frame of video sequence. First, we find a combined half-ellipse adjacent to the lip region, as the initial evolutionary curve for evolution, so that the lip image can be segmented in the lip region and non-lip regions. Second, we split the image resulted from lip ROI extraction stage into vertically six pieces to extract lip features. Fig. 8 shows the results of extracted lip contour.

Finally, after extracting the lip contour of the previous lip framework, to initialize the lip boundary and trace the lip contour of the current frame, the initial evolution curve embedded into the Moore's neighborhood tracking algorithm. Fig. 9 and Fig. 10 show an example of lip tracking results.

Fig. 7. Equalized images on different lip shape.

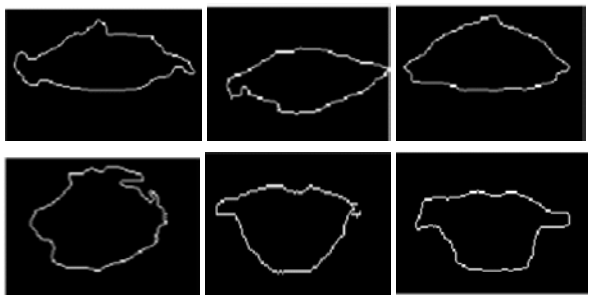

Fig. 8. Extracted lip contour.

Fig. 9. The negative lip tracking results for utterance of one syllable and two syllable consonants on only selected frame.

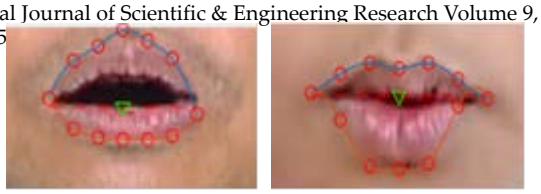

Fig. 10. The positive lip tracking results for utterance of one syllable and two syllable consonants on only selected frame.

4 Experimental Results

4.1 Database for Lip Reading System

The proposed lip contour extraction and lip movement tracking methods have been implemented and tested on own audiovisual lip images and videos of Myanmar consonants.

The own audiovisual database was recorded in three lighting system. Database consists of fourteen speakers, two persons of male speakers and twelve persons of female speakers. Both are white and black skin. Sony DVCam -DSR 300A professional video camera is used. Videos are recorded in mp4 format with 29 frames/second. Recording distance is constant. All video recorded are taken from frontal view. Each image frame has resolution of 720×480. Video was recorded for 8 consonants of one syllable consonants and 14 consonants of two syllable consonants. The images of two syllable and one syllable Myanmar consonants is illustrated in Table 3 and Table 4.

4.2 Results of Lip Localization

To obtain accurate and precise recognition results for lip reading system, considerable significant localization accuracy is required. The purpose of this paper is to detect, localize and track the lip movement to achieve automatic and reliable lip reading goals. So, we experimented the segmentation process with YCbCr, CIELa*b*, the combination of YCbCr and CIELa*b* color space model for localizing the lip. Table 1 and Table 2 show the comparison of localization accuracy rate on different methods for two and one syllable Myanmar consonants.

Fig. 11 and Fig. 12 show the evaluation charts of lip localization accuracy for one syllable and two syllable consonants on different methods. By comparing with different approaches, our proposed method has the high localization accuracy and minimum of error rate, according to the results in Table 1 and Table 1. These results demonstrated that when we used the combination of de-correlation stretching color enhancing method, YCbCr, CIELa*b* color space in segmentation process and Moore neighbor tracing algorithm is used to extract and track the lip contour, the localization accuracy rate increase. Therefore, the proposed lip localization techniques have achieved a more satisfactory result for lip reading system.

TABLE 1
Localization Accuracy Rate For Two Syllable Consonants

Two Syllable Consonants	Localization Accuracy Rate				Error Rate			
	CIELa*b*+ Moore Neighbor %	YCbCr+ Moore Neighbor %	CIELa*b*+ YCbCr + Moore Neighbor %	CIELa*b*+ YCbCr + Decorrelation Stretching+ Moore Neighbor (proposed method) %	CIELa*b*+ Moore Neighbor %	YCbCr+ Moore Neighbor %	CIELa*b*+ YCbCr + Moore Neighbor %	CIELa*b*+ YCbCr + Decorrelation Stretching+ Moore Neighbor (proposed method) %
ကျ (Ka Gyi)	88.72	96.38	97. 62	98.66	11.28	3.62	2.38	1.34
ခွ (Kha Gway)	72.45	94.82	96.49	96.91	27.55	5.18	3.51	3.09
ဂ (Ga Nge)	84.56	96.19	88.73	97.33	15.44	3.81	11.27	2.67
ဃျ (Ga Gyi)	84.71	90.96	95.49	96.48	15.29	9.04	4.51	3.52
စ (Sa Lone)	67.02	82.83	94.22	94.19	32.98	17.17	5.78	5.81
ဆ (Sa Lain)	85.31	95.72	96.13	96.02	14.69	4.28	3.87	3.98
ဇ (Za Gwe)	75.65	95.26	95.17	98.44	24.35	4.74	4.83	1.56
ဒ (Da Dway)	71.77	92.88	95.17	96.29	28.23	7.12	4.83	3.71
ဏ (Na Gyi)	85.28	94.69	96.84	96.71	14.72	5.31	3.16	3.29
န (Na Nge)	85.47	96.31	96.79	98.13	14.53	3.69	3.21	1.87
ပ (Pa Saug)	80.29	92.21	93.00	96.40	19.71	7.79	7.00	3.60
ဘ (Ba Gone)	52.85	67.71	87.88	89.88	47.15	32.29	12.12	10.12
ရ (Ya Gaug)	68.56	93.91	96.44	98.46	31.44	6.09	3.56	1.54
ဠ (La Gyi)	83.78	96.92	96.53	98.45	16.22	3.08	3.47	1.55

Total accuracy rate/ error rate	77.60	91.91	94.75	96.60	32.40	8.09	5.25	3.40

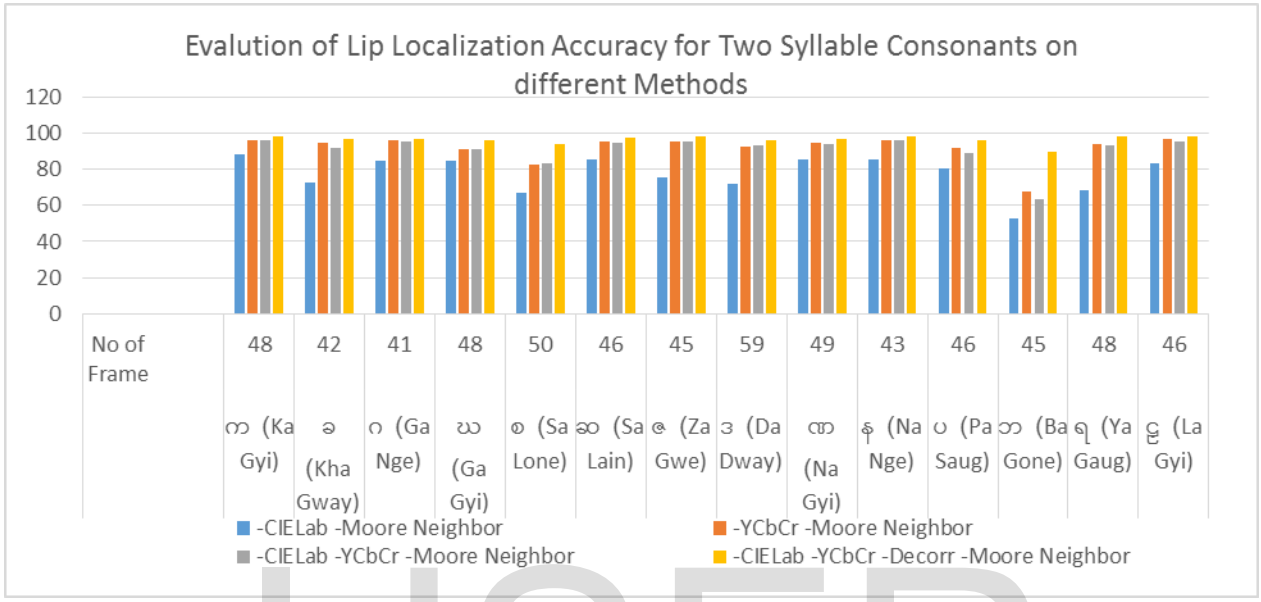

Fig. 11. Evaluation of Lip Localization Accuracy for Two Syllable Consonants on different Methods.

TABLE 2
LOCALIZATION ACCURACY RATE FOR ONE SYLLABLE CONSONANTS

One Syllable Conso-nants	Localization Accuracy Rate				Error Rate			
	CIELa*b*+ Moore Neigh-bor %	YCbCr+ Moore Neighbor %	CIELa*b*+ YCbCr + Moore Neighbor %	CIELa*b*+ YCbCr + Decorrelation Stretching+ Moore Neighbor (proposed method) %	CIELa*b*+ Moore Neighbor %	YCbCr+ Moore Neighbor %	CIELa*b*+ YCbCr + Moore Neighbor %	CIELa*b*+ YCbCr + Decorrelation Stretching+ Moore Neighbor (proposed method) %
c(Nga)	60.07	98.18	97.08	98.48	39.93	1.82	2.92	1.53
ည(Nya)	67.52	93.11	96.05	97.05	32.48	6.89	3.96	2.95
ဓ(Ma)	64.35	85.79	86.26	92.96	35.65	14.21	13.74	7.04
လ(La)	72.96	95.28	93.88	97.71	27.04	4.72	6.12	2.29
ဝ(Wa)	65.01	92.11	96.68	96.87	34.99	7.89	3.32	3.14
သ(Tha)	54.68	97.55	98.27	98.68	45.32	2.45	1.73	1.32
ဟ(Ha)	71.83	96.22	96.65	97.96	28.17	3.78	3.36	2.04
အ(Ah)	69.37	97.45	94.44	98.13	30.63	2.55	5.57	1.88

Total accuracy rate/ error rate	65.72	91.11	94.91	97.23	34.28	8.89	5.09	2.77

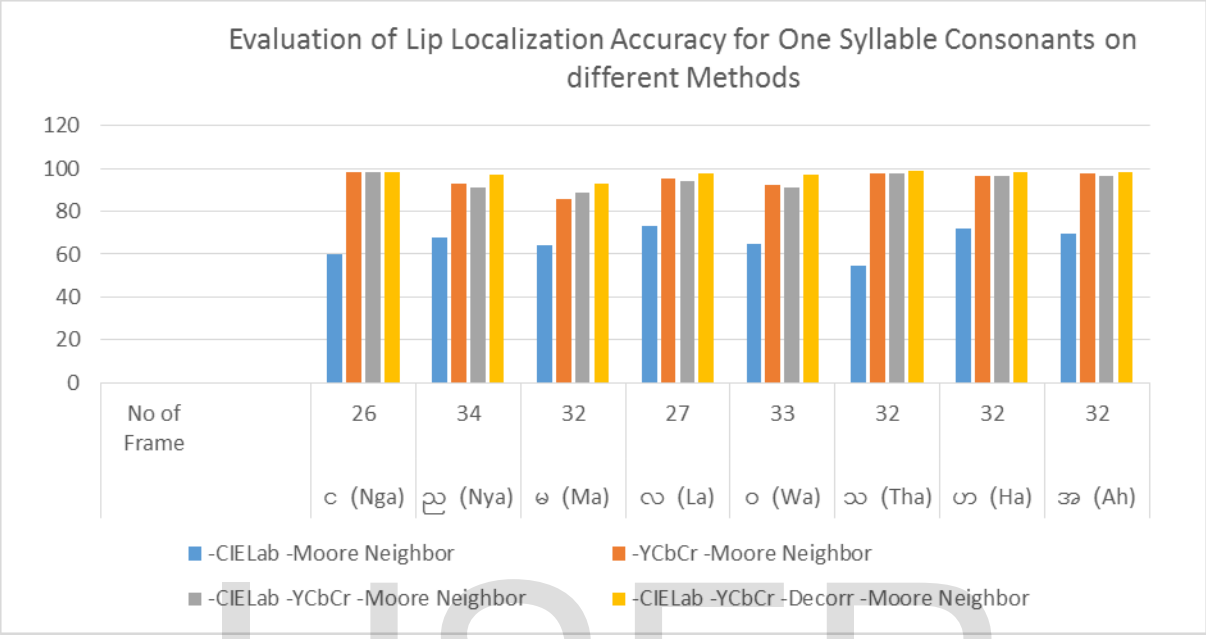

Fig. 12. Evaluation of Lip Localization Accuracy for One Syllable Consonants on different Methods.

TABLE 3
IMAGES FOR TWO SYLLABLE CONSONANTS

Two Syllable Consonants	Images
ကြ (Ka Gyi)	
ခြ (Kha Gway)	
ဂ (Ga Nge)	
ဃြ (Ga Gyi)	
စြ (Sa Lone)	
ဆ (Sa Lain)	
ဇ (Za Gwe)	
ဏြ (Na Gyi)	
ဒ (Da Dway)	
နှ (Na Nge)	
ပ (Pa Saug)	
ဘ (Ba Gone)	
ရ (Ya Gaug)	
ဠ (La Gyi)	

TABLE 4
IMAGES FOR ONE SYLLABLE CONSONANTS

strate that this approach performs accurate and significant localization for lip motion sequences in video. These results were perceived to be acceptable for lip movement recognition system.

5 CONCLUSIONS AND DISCUSSION

The propose system aims not only to recognize lip movements during the utterance of Myanmar consonants by the speaker but also to investigate dynamic motion of mouth opening and closing. This paper proposed efficient lip movement recognition approach towards an automatic lip reading system. The experiments is done on different methods for lip localization. This system proposed a solution for automatic lip localization based on lip movements. The experimental results demon-

REFERENCES

[1] Matthews, Iain, Timothy F. Cootes, J. Andrew Bangham, Stephen Cox, and Richard Harvey. "Extraction of visual features for lipreading." Pattern Analysis and Machine Intelligence, *IEEE Transactions* on 24, no. 2 (2002): 198-213.

[2] Zhi, Qi, A. D. Cheok, K. Sengupta, Zhang Jian, and KO Chi Chung. "Analysis of lip geometric features for audio-visual speech recognition." Systems, Man and Cybernetics, Part A: Systems and Humans, *IEEE Transactions* on 34, no. 4

[3] Kji lwano, Tomoaki Yosh inaga, Santoshi Tamura and Sadaoki Furui, "Audio-Visual speech Recognition Using Lip information Extracted from side-face Images", *EURASIP Journal and Audio, Speech, and Music Processing*, January 2007.

[4] Vicente P. Minotto, Carlos Bo. Lopes,Jacob Scharcanski, Claudi R. Jung and Bowon Lee, "Audiovisual Voice activity Detection Based on Microphone Arrays and Color Information", *IEEE Journal of selected topics in Signal Processing*, February 2013.

[5] Luca Lombardi, Waqqas ur Rehman Butt, Marco Grecuccio, "Lip Tracking Towards an Automatic Lip Reading Approach", *ResearchGate*, March 2014.

[6] P.Sujatha et al., "Novel Pixel-based Approach for Mouth Localization, *International Journal of Computer Applications*" (0975 – 8887), International Conference on Computing and information Technology, IC2IT, 2013.

[7] T. Coianiz, L. Torresani and B. Caprile, "2D Deformable Models for Visual Speech Analysis", Proceedings of Sprintger, Speech reading by Humans and Machines, D.G. stork & M. E. Hennecke Eds., NY, 1996.

One Syllable Consonants	Images
င(Nga)	
ည(Nya)	
မ(Ma)	
လ(La)	
ဝ(Wa)	
သ(Tha)	
ဟ(Ha)	
အ(Ah)	

[8] Meng Li and Yiu-ming Cheung, "Automatic Segmentation of Color Lip Images Based on Morphological Filter", ICANN, 2010.

[9] Alan Wee-Chung Liewa, Shu Hung Leungb, Wing Hong Laua, "Lip contour extraction from color images using a deformable model", The Journal of Pattern Recognition Society, Nov 2002.

[10] Stefan Badura and Michal Mokrys, "Feature extraction for automatic lips reading system for isolated vowels", ICTIC, March 23. - 27. 2015.

[11] N. Eveno, A. Caplier, P.Y. Coulon. "A new color transformation for lips segmentation", Proceedings of IEEE Fourth Workshop on Multimedia Signal, pp. 3-8, Cannes, France, 2001.

[12] Namrata Dave, "A Lip Localization Based Visual Feature Extraction Meth-

od", Electrical & Computer Engineering: An International Journal (ECIJ), Volume 4, Number 4, 2015.

[13] N. Otsu, "A Threshold Selection Method from Gray-Level Histogram", IEEE Transaction on Systems, Man, and Cybernetics. Vol. SMC-9, Pontificia Universidance Catolica Do Rio de Janeiro, 1979.

[14] Faridah, Balza Achmad, Binar Listyana S, "Lip Image Feature Extraction Utilizing Snake's Control Points for Lip Reading Applications", International Journal of Electrical and Computer Engineering (IJECE), Vol. 5, No. 4, pp. 720~728, August 2015.

[15] Y. Tian, T. Kanade, J. Cohn, Robust lip tracking by combining shape, color and motion, in: *Proceedings of the Asian Conference on Computer Vision*, pp. 1040–1045, 2000.

[16] N. Eveno, A. Caplier, P.Y. Coulon, Accurate and quasi-automatic lip tracking, IEEE Transactions on Circuits and Systems for Video Technology 14 (5) (2004) 706–715.