Movements Recognition Towards An Automatic Lip Reading System for One Syllable Myanmar Consonants

Thein Thein, Kalyar Myo San
University of Computer Studies,Mandalay
theinthein.cmw@gmail.com, kalyar.myosan@gmail.com

Abstract

*Lip reading is a process of extracting visual information, observing the movement of the lips of the speaker with or without sound. To extract visual information, reliable movements of the lips are necessary. The major challenge is to recognize lip movements because of many possible lip motions and lip shapes. The accuracy and robustness of a speech recognition system can be improved by using visual information from lip movements and the need for lip reading system is ever increasing for every language. Therefore, this paper presents Myanmar consonant recognition based on lip movements toward lip reading by using CIELa*b* color transformation, Moore Neighborhood Tracing algorithm and Otsu global thresholding technique. This research aims to develop a visual teaching method system for the hearing impaired persons precisely identifying the features of lip movement.*

1. Introduction

Lip reading was used for a variety of purposes, for learning voice hearing and improving speech recognition. In the lip reading system, the lips are read to extract visual information from the video input. However, lip movement recognition is active undergoing research topics with lot of improvements that recovers various difficulties faced in the research. Many researchers have proposed various methods for the accurate identification of the lip region and proposed various functions for obtaining significant recognition results. The purpose of this study is to propose a visual teaching method for Myanmar consonants recognition for the hearing impaired persons by precisely localizing the lip movements and activities when they produce the consonants.

In this paper, we propose approaches to recogize accurate lip region 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)) of Myanmar language. This paper aims to extract the accurate upper and lower lip boundary based on lip movement using CIELa*b* color space model, Otsu global thresholding method and Moore Neighborhood Tracing Algorithm.

This paper is organized as follows: the section 1 finishes a brief introduction given above, related works will present in section 2, section 3 presents framework of the proposed lip reading system, research method will describe in Section 4, section 5 gives lip feature extraction, section 6 will show the classification of lip movements, experimental results will be mentioned in section 7, finally conclusion and future work will be done.

2. Related works

Many researchers proposed various techniques for lip localization. Some of the techniques are applied on side view of the face to localize the lips [3]. Some of the techniques are also applied on front view of the face from which lip is localized [12]. W. Salah et al. [13], L. Luca et al. [4] are applied on front view of the face from which lip is localized. D. Namrata [2] implemented a novel color based approach for lip localization based visual feature extraction method which gave a good accuracy for their database.

D. Namrata implemented a novel color based approach for lip localization based visual feature extraction method which gave a good accuracy for their database [2]. X. Liu and Y.M. Cheung [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. R.E. Hursig et al. [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. The proposed algorithm is shown to effectively differentiate facial features, including lips, from their backgrounds and to bind the full extent of the lips within a face classified region of interest and the proposed algorithm making it more versatile within the unconstrained environment. S. Badura and M. Mokrys [9] proposed designs for automatic lip reading for the Slovak language. They work a deep analysis of feature extraction process based on recursive median filter.

S. Pathan and A. Ghotkar [10] proposed a method to recognize the different English phrases using geometrical features of lip shape. They presented a solution for automatic lip tracking and recognition of phrase. S. S. Morade and B.S. Patnaik [11] presented a novel active contour guided geometrical feature extraction approach. They performed to determine important statistical parameters such as area, height and lip's angle of geometrical method. A. Brahme and U. Bhadade [1] presented a methodology for detecting lip using Constrained Local Model (CLM) and extracted geometric features of lip shapes. M. Li and Y.M. Cheung [5] proposed automatic lip segmentation approach from a probability model in color space and morphological filter to estimate the model parameter; they used hue and saturation value of each pixel within the lip segment. P.Sujatha et al. [7] presented a new method for automatic lip detection using geometric projection method and adaptive thresholding.

3. Framework of Lip Reading System

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.

Figure 1 shows the processing stages of the proposed lip reading system and.

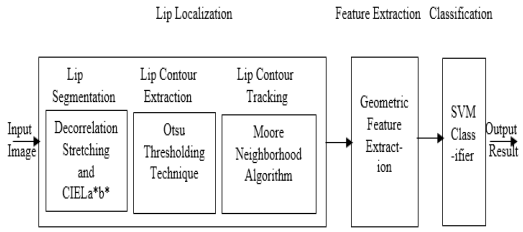

Figure 1. Framework of the proposed lip reading system

4. Research Method

4.1. Lip Localization

Lip localization is the important step in lip reading system to detect accurate lip boundary and to extract lip features. There are three kinds of models for localizing the lips. The one is low level or image based, 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 other one is high level or model based, uses integrity constraints and pixel information to segment the lip. It is finding the corner of the lips to detect the accurate lip. The final one is the hybrid model which is using the parameters of both the models. Among these techniques, this paper used hybrid approach to localize lip region for lip reading system.

4.1.1. Lip Segmentation

Lip segmentation is designed to separate the lip from the background skin color. Segmentation methods aiming at segmenting the lip shape, boundary or mouth area from the images of the input video. In this paper, Lip region segmentation required several process of image processing. Normalized cross-correlation method, Decorrelation stretching color enhancing method and CIEL*a*b* color transformation method are used for lip segmentation. The proposed segmentation process starts frame normalization by breaking the video image. Then, extract lip region by using normalize cross-coorelation as a preprocessing stage. Figure 2 show the original image and the result image after extracting the lip region. Figure 3 shows the normalized image frames.

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.

Decorrelation Stretching color enhancing method and CIEL*a*b* color transformation method are based on differences in color composition between lip as the object and skin as the background. Skin colors are marked more on color composition compare to brightness, even on different people. Color re remarkably constant even when exposed by a lot of illumination. So, in our experiments, we used Decorrelation Stretching color enhancing method with Stretch limit. And then, we used Kalman image filtering method to eliminate a lot of noise and unwanted information in image around mouth region. Color enhanced images are shown in Figure 4.

(a) **(b)**

Figure 2. (a)Original image, (b) Segmented lip region

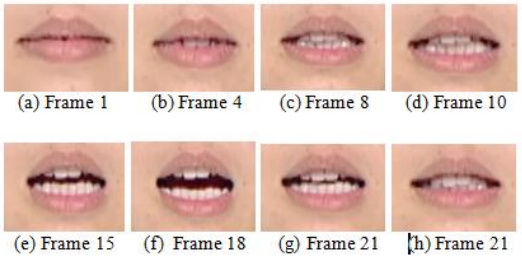

Figure 3. (a) to (h) Number of selected frames for utterance of Ma (one syllable consonant)

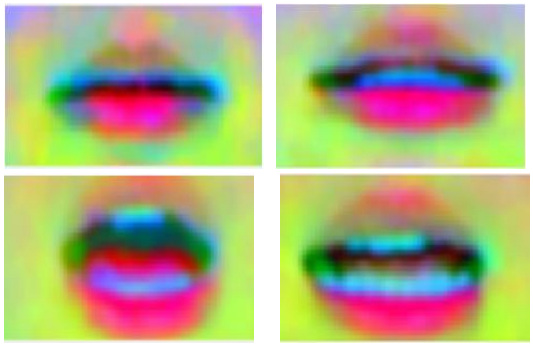

Figure 4. Color enhanced images for different frames

After enhancing image, RGB color image is transform into CIEL*a*b* color space based on first layer L channel, result is shown in Figure 5. For better lip localization, histogram equalization is used to color contrast. Therefore, the lip area appears much darker than the skin. The result images are shown in Figure 6.

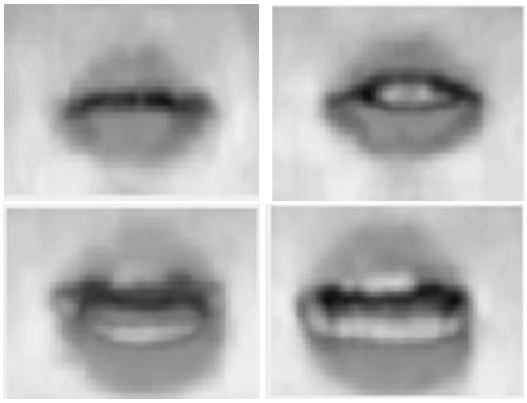

Figure 5. Color transformed images for different frames

Figure 6. Equalized Images for different frames

4.1.2. Lip Contour Extraction

Otsu thresholding is an automatic thresholding technique that commonly referred to as adaptive threshold [6]. Otsu thresholding is needed to perform image binarization to the image resulted from color transformation. This Otsu thresholding method calculates the value of threshold (T) for segmentation based on the input image. Otsu thresholding technique seeks the optimal threshold value to separate object from background by maximizing the variance between classes while minimizing the variance within classes. The maximum value of the variance between classes is defined in the following equations:

$$\sigma_w^2(T) = max_t \ \sigma_w^2(t) \quad (1)$$

$$\sigma_w^2(t) = w_1(t)\, w_2(t)\, (\mu_2(t) - \mu_1(t))$$

Where w_1 and w_2 are the probability of pixels in each class, while μ_1 and μ_2 are the mean gray scale of each class. The probabilities and means for each class are updated iteratively. In this paper, the upper and lower lip contour is extracted by using Otsu global thresholding technique; results are shown in Figure 7.

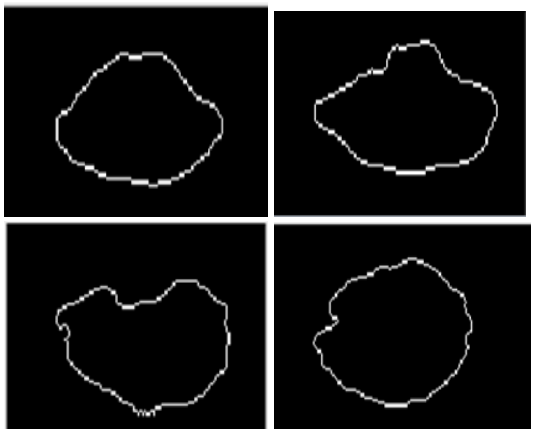

Figure 7. Extracted lip boundary contour images for different frames

4.1.3. Lip Contour Tracking

Information about lips position in each image (video frame) can enhance the effectiveness of lips reading [10]. Generally two basic approaches can be used for lip contour tracking:

• Tracing and localizing lips boundary in each frame.

• Tracking Upper and Lower lips regions over all frames.

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, after extracting the lip contour of the previous lip frame, we let it be the initial evolving curve embedded in Moore Neighborhood tracing algorithm to localize lip boundary and to lip contour tracking of the current frame. With this approach, we are able to estimate lip boundary in each frame of video sequence. An example of lip tracking results are shown in Figure 8.

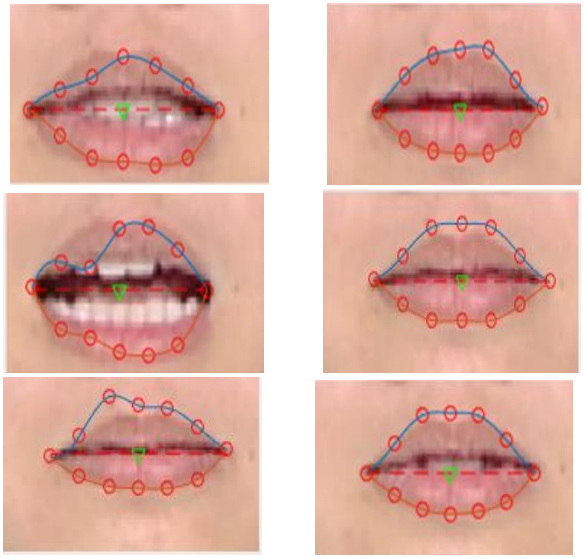

Figure 8. Lip tracking results for utterance of Ma (one syllable consonant) on only selected frame. The first column shows negative tracking results and the second column shows positive tracking results

5. Feature Extraction

In this paper, we take twelve coordinate points from lip border to extract six lip features. These coordinate points are basically one leftmost point, one rightmost point, five uppermost points and five bottommost points of the lip boundary, as illustrated in Figure 9.

In feature extraction step, appropriately normalize and rotate the outer lip contours using geometric feature extraction approach. The extracted lip features are lip height (H), width or distance of imaginary line(D), distance between imaginary line and bottom points of the lip boundary (F1, F3, F4), distance between imaginary line and uppermost points of the lip boundary(F2). The extracted features are the most informative for automatic lip reading. Figure 9 shows the extracted six lip features. These six features are calculated as the following equations based on lip model. Lip model is shown in Figure 10.

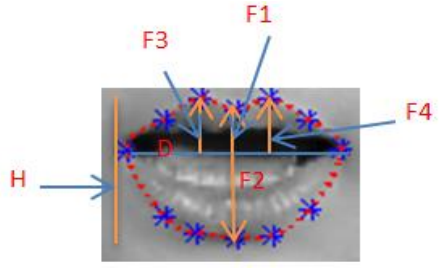

Figure 9. Six features of lip movement

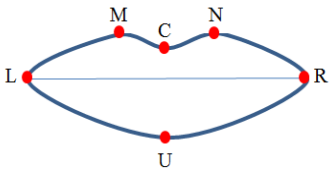

Figure 10. Lip model

$$H = \max(M_y - N_y) - U_y \quad (2)$$

$$D = R_x - L_x \quad (3)$$

$$F_1 = C_y - \left(\frac{L_y + R_y}{2} \right) \quad (4)$$

$$F_3 = M_y - \left(\frac{L_y + R_y}{2} \right) \quad (5)$$

$$F_4 = N_y - \left(\frac{L_y + R_y}{2} \right) \quad (6)$$

$$F_2 = \left(\frac{L_y + R_y}{2} \right) - U_y \quad (7)$$

The next step is to extract more valuable three features based on extracted the above six features of lip movements. Three features are aspect ratio, upper and lower lip ratio and lip's roundness. The mathematical equations are formulated as follows:

$$AR = \frac{D}{H} \quad (8)$$

$$ULR = \frac{F_1}{F_2} \quad (9)$$

$$RDN = \frac{4\pi A}{PM^2} \quad (10)$$

where, A is the area of lip and PM is the parameter of the lip.

6. Classification of Lip Movements

Final task of the proposed system is to classify the features of each video frame. The classification process can be trained and executed for different classes on own audiovisual database (one syllable consonants).

In this paper, the classification for Lip movement recognition is done with linear SVM classification learner. Support Vector Machine (SVM) classifier for linear and separable case is capable to find the optimal separating hyper plane between classes in sparse high dimensional spaces with relatively few training data. The SVMs were trained with 80 training samples and were tested using the 80 remaining samples of each video frame for all 12 speakers.

7. Experimental Results

7.1. Database for Lip Reading System

In this paper, experiments were tested on own audiovisual database of Myanmar consonants. Database consists of twelve speakers, four persons of male speakers and eight persons of female speakers. Both are white and black skin. Sony DVCan-DSR 300A professional video camera is used. Videos are recorded in mp4 format with 23frames/second. Recording distance is constant. Each image frame has resolution of 720×480.

7.2. Results of Lip Localization

Table 1. Localization accuracy rate for one syllable consonants

One Syllable Consonants	Accuracy Rate	Error Rate
ငNga)	90.9%	9.10%
ညNya)	88.09%	11.91%
မMa)	97.2%	2.80%
လLa)	89.19%	10.81%
ဝWa)	89.09%	10.91%
သTha)	93.18%	6.82%
ဟHa)	89.47%	10.53%
အAh)	92.45%	7.55%
Total accuracy rate/ error rate	91.20%	8.80%

Significant localization accuracy is needed to get accurate and precise lip movement recognition results for lip reading system. So, we experimented the segmentation process with CIELa*b* color space model. Tables 1 shows the localization accuracy rate for one syllable consonants. The lip localization accuracy rate have achieved a more satisfactory result to lip movement recognition.

7.3. Results of Classification

In the proposed system, classified the lip movement during the utterance of the consonants by using the linear SVM. As the experimental results, the classification accuracy rate of linear SVM is more

accurate than other classification methods for the proposed system. However, the one syllable consonants, ငNga), ဟ(Ha) and အ(Ah) may not be clearly classified. Table 2 shows the comparison of classification accuracy rates on different classification learners.

Table 1. Classification accuracy rate for different classifier

Classifier	Dataset	Classification Accuracy Rate
Quadratic SVM	Own Dataset of Myanmar Consonants	90.57%
Linear SVM	Own Dataset of Myanmar Consonants	91.89%
Fine KNN	Own Dataset of Myanmar Consonants	86.92%
Weight KNN	Own Dataset of Myanmar Consonants	88.88%
CNN	Own Dataset of Myanmar Consonants	86.79%

8. Conclusion and Future Work

This paper proposed efficient lip movement recognition approach towards an automatic lip reading system. The proposed 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 system uses geometric features and Moore Neighborhood tracing algorithm to recognize lip movements and SVM classification method is used to classify. The experimental result demonstrates that this approach performs accurate and significant recognition for lip motion sequences in video. In our experiment, we can recognize all of the test lip movement significantly and the results were perceived to be acceptable for lip reading. For future work, we will intend to explore more observable features and recognition phase for remaining, two syllable, three syllable and four syllable Myanmar consonants. We hope that this study will help to a new teaching and learning method for Myanmar language education.

References

[1] Brahme and U. Bhadade, "Lip Detection and Lip Geometric Feature Extraction using Constrained Local Model for Spoken Language Identification using Visual Speech Recognition," Indian Journal Science and Technology, Vol 9(32), August 2016.

[2] D. Namrata, "A Lip Localization Based Visual Feature Extraction Method," Electrical & Computer Engineering: An International Journal, ECIJ, Volume 4, Number 4, 2015.

[3] K. lwano, T.Y. Inaga, S. Tamura and S. Furui, "Audio-Visual speech Recognition Using Lip information Extracted from side-face Images," EURASIP Journal and Audio, Speech, and Music Processing, January 2007.

[4] L. Luca, R.B. Waqqas ur and G. Marco, "Lip Tracking Towards an Automatic Lip Reading Approach," ResearchGate, March 2014.

[5] M. Li and Y.M. Cheung, "Automatic Segmentation of Color Lip Images Based on Morphological Filter," ICANN, 2010.

[6] 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.

[7] 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.

[8] R.E. Hursig, J.X. Zhang, and C. Kam, "Lip Localization Algorithm Using Gabor Filters," International Conference on Image Processing, Computer Vision and Pattern Recognition, ICPR, 2012.

[9] S. Badura and M. Mokrys, "Feature extraction for automatic lips reading system for isolated vowels", ICTIC, March 23. - 27. 2015.

[10] S. Pathan and A. Ghotkar, "Recognition of spoken English Phrases Using Visual Features Extraction and Classification," International Journal of Computer Science and Information Technologies, IJCSIT, Vol.6 (4), 3716-3719, 2015.

[11] S.S. Morade and B.S. Patnaik, "Automatic Lip Tracking and Extraction of Lip Geometric Features for Lip Reading," International Journal of Machine Learning and Computing, Vol 3, No.2, April 2013.

[12] V. P. Minotto, Carlos B.o. 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.

[13] W. Salah, W. Mahdi and A.B. Hamadou, April 12-14, "Automatic Hybrid Approach for Lip POI localization: Application for Lip-reading System," ICTA'07, Hammamet, Tunisia.

[14] X. Liu and Y.M. Cheung, "A Robust lip tracking Algorithm Using Localized Color Active Contour and Deformable Models," ICASSP, 2011.