Viseme Recognition using lip curvature and Neural Networks to detect Bangla Vowels

Nahid Akhter, Amitabha Chakrabarty
Department of Computer Science and Engineering
BRAC University
Dhaka-1212, Bangladesh
lubnaonline81@yahoo.com

***Abstract*— Automatic Speech Recognition plays an important role in human-computer interaction, which can be applied in various vital applications like crime-fighting and helping the hearing-impaired. This paper provides a new method for recognition of Bengali visemes based on a combination of image-based lip segmentation techniques, use of curvature of the both inner and outer lips as well as neural networks. The method is divided into three steps. First step is a lip segmentation step that uses a combination of red exclusion method, HSV space and CIE spaces to produce illumination invariant images. Next, inner and outer lips are extracted separately using a new technique for curve-fitting. Second step is the feature extraction step, which makes use of quadratic curve-coefficients of the inner and outer lip contours. Finally, viseme recognition is done using a Neural Network. A dataset was created with 171 lip images of Bangla Visemes being spoken by different speakers and under different lighting conditions. The proposed method gave a viseme recognition result of 87.3%. Due to the use of non-iterative method as opposed to conventional methods, the algorithm was found to be faster in detecting lip contours.**

***Index Terms*—Active Appearance Model; Artificial Neural Networks; Bangla Viseme Recognition; Lip Reading; Speech Recognition.**

I. INTRODUCTION

Human-computer interaction is a research area that has fascinated scientists and engineers for a very long time. Within this arena, automatic speech recognition is of special interest as it forms the basis for important human applications, like teaching people with hearing or speech impairment to speak and communicate effectively. Moreover, a visual speech recognition system can help intelligence agencies track a remote conversation by using a camera, where auditory input or support is not available. Visemes are used by the hearing- impaired to view sound visually, thus effectively lip reading [1]. So far, a lot of research has been done on lip-reading in English and substantial work on French [2] and Chinese [3], as well as few other languages [4] [5], but not much research has been done on lip-reading in Bengali.

This paper provides a new approach to lip reading Bengali vowels using a combination of the curvature of the inner and outer lips and Neural Networks.

The first part of this paper gives an overview about conventional algorithms and technology normally used in lip-reading and contour-finding. The second part gives a description of a new proposed contour-finding algorithm. The third part draws a comparison between traditional methods of contour finding and the proposed contour-finding algorithm. The fourth part explains the method used by the proposed system for viseme recognition. The fifth part gives observations and results of experiments done using the system and the last part consists of conclusion, challenges and future endeavours.

The experiments were conducted in three phases. First phase selected the best lip localization method, second phase selected the best lip contour finding method and the third phase selected the best parameters to use for machine learning of the viseme identities.

In all, it was found that a combination of colour based methods, followed by a curve fitting method, and finally the use of neural networks were the optimum set of steps to detect a viseme spoken.

II. LITERATURE REVIEW

Lip segmentation or contour finding techniques may be image-based, colour-based, or model based [6]. Some image-based methods include DCT [5] [17] and PCA methods [13]. In addition, conversions to variouscolor spaces such as RGB (Red-Green-Blue), HSI (Hue-Saturation-Intensity) [7], and YCbCr (Luminance-Component blue- Component red) [18] or L*a*b space are used. Pixels of lip area have stronger red component and weaker blue component than other facial regions. Therefore, the chrominance component Cr has greater value than the Cb in the lip region. From all the available colour spaces, It has been found in [18] that the Saturation component of the HSI colour space, in combination with the Cr and Cb component of the YCbCr colour space provide a good base .In some cases, the image is converted to a contrast-enhanced black and white image. For better lip recognition, the histogram equalization algorithm is used to enhance contrast so that the lip area appears much darker than the skin. However, the disadvantage of Image-based systems is that they are restricted to changes in illumination, mouth rotation and dimensionality [1]. Image based techniques use the pixel information directly, the advantage is that they are computationally less expensive but are adversely affected by variations such as illumination.

In 1988, Kass et al [8] proposed the concept of a snake; a model-based technique. Which used an energy-minimizing spline guided by external constraint forces and influenced by image forces that pull it toward features such as lines and edges. These snakes lock onto nearby edges, localizing them accurately.

However, to solve energy minimizing crisis, snakes

require long computational time and large amount of calculations that make it unfeasible in stand-alone system to extract area function. Moreover, when the snake is initialized properly, it enters into the object region and then the repulsion force won't work. In addition, the need for the contour points to be initialized by the user first makes it not so feasible.

It has been found that model based techniques are more widely used in most contour-finding applications currently, in particular, the Active Contour Models (ACM), Active Shape Models (ASM) and Active Appearance Models (ASM) [19] [20].

Model based techniques are based on prior knowledge of the lip shape and can be quite robust. They learn the shape and appearance of lips from training data that has been manually annotated. From [19] it was concluded that the AAM approach produced the most reliable results in terms of lip localization with an error rate of just 0.3%. However, the prior process of manually fixing landmarks for the process makes using AAM and ASM a tedious and time-consuming task. Moreover, the contour-finding process itself takes place in iterations until a match is found, which again adds to the processing time, especially when large videos need to be processed. AAM also has the additional disadvantage of being too memory-intensive to run on an average smartphone or computer.

In addition, there are some hybrid techniques [9]. These methods combine both image based and model based techniques. Majority of the hybrid techniques proposed in the literature [9] use color based techniques for a quick and rough estimation of the candidate lip regions and then apply a model-based approach to extract accurate lip contours.

Feature extraction is another important factor in lip reading. A lot of lip reading algorithms use Principal Components as features for viseme recognition [14] [16], while many others use geometrical features such as height and width of mouth, area, perimeter, etc. The French ALiFe system in [2] used a feature called DA (Dark Area) which was area of the dark region inside the mouth.[21] had used a feature set consisting of a combination of a variety of descriptive features like Height and Width, Image Quality value, presence of tongue and Number of teeth pixels.

In 2006, Chen proposed in his paper [3] a feature extraction method for lips which used a variation of the red-exclusion method for lip detection, followed by a curve-fitting to get the contour of the inner lips only. This method was much faster than other contour-finding algorithms, however, the initial red exclusion method did not give accurate enough results.

III. Methodology

A. Dataset

Since this research primarily deals with viseme recognition of Bangla vowels, it was necessary to have a dataset consisting of images with Bangla vowels being spoken. Although there are a number of data sets for lip-reading available online, there were no existing dataset of Bengali visemes. So, it became necessary to create a Bangla viseme dataset from scratch.

The dataset used contained lip images of Bangla visemes being spoken. For simplicity, the dataset contained images of only three Bangla vowels, "B", "A" and "H" being spoken. For each vowel, a total of 57 images were collected. That means, in all, the data set contained 57 x 3 = 171 images. These images were of different speakers, and they were taken under varying lighting conditions, during different times of the day and in different locations.

The images were of varying sizes. This was done to show that the algorithm was independent of changes in image dimensions. Mouth images were in varying angles. This was to show that the algorithm was robust against changes in mouth orientation with the camera tilted to different angles, as is common in amateur photography by smartphones. Mouth images of different subjects were used to test the speaker-independence of the algorithm. Images were taken under varying lighting conditions, for example, some were taken in natural light, while others were taken indoors, to test the robustness of the system to variations in illumination.

Of the 171 images, 120 of them were used for training, while the rest 51 were used as the test set..

B. Proposed Method

The proposed method makes use of image-based techniques to roughly identify the lip area and then uses Chen's curve-fitting [3] to extract lip feature to be used for describing the lip shape, which will aid in lip recognition. For better accuracy, both inner and outer lip curves were used.

The algorithm is divided into three parts. The first part deals with extraction of the outer lip contour. Second part deals with extraction of the inner lip contour and third part deals with recognition of the uttered viseme.

1. The original image is first broken down into its RGB components. And the red component is thresholded using a threshold of 68. This masks out the inside of the lip. However, if teeth are present, they need to be masked out too. So, as described in [21], the teeth are also masked out by converting the RGB image to CIELUV and CIELAB colour spaces and thresholding the U and A components respectively using a combination of their means and Standard Deviations.

2. All three of these masks are combined together and morphological operators are used to clean up the final mask [22].

3. The image is cropped to the edges of this mask and the left most and right most points of the mask are found to correspond to the left and right corners of the lips. A total of 32 points are found along the edge of this inner lip mask, where 8 points each belong to the left side of the upper mouth, right side of the upper mouth, left side of the lower mouth and right side of the lower mouth.

4. These 32 points give a rough estimate of the inner lip shape. However, the lip edge in reality is a smooth curve. So, we interpolate these points using four quadratic curves.

The upper right lip was interpolated by a quadratic curve as:

$$a1x2+b1x+c1=0 \quad (1)$$

The upper left lip was interpolated as:

$$a2x2+b2x+c2=0 \quad (2)$$

and the lower lip was interpolated by:

$$a3x2+b3x+c3=0 \quad (3)$$

Finally we use the vector of coefficients (a1, b1,c1,a2,b2,c2,a3,b3.c3) to describe the inner-lip shape. The process of finding theses three inner lip contours is demonstrated in Figure 1.

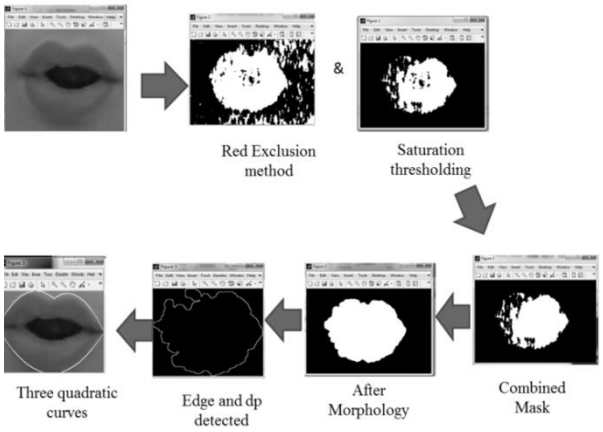

Figure 1: Extraction process of outer lip contours

5. To extract the outer lip shape, initially a different approach is used. For this, the red exclusion method of Chen [3] is used directly. A mask is thus obtained.

6. To further improve upon this mask, a second mask is found by converting the image to HSV space and masking out all the pixels that have saturation component less than the mean saturation.

7. Both masks obtained in steps 6 and 7 are combined by a logical AND operation and the final binary image obtained following morphological cleanup is used to get a preliminary outline of the outer lips.

8. The mask is then reduced to an edge using canny edge detection

9. The image is cropped to the edges and the left most and right most points of the mask are found to correspond to the left and right corners of the lips.

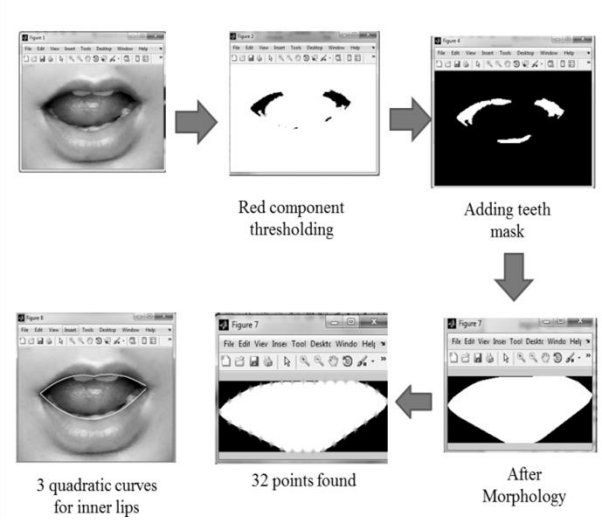

Figure 2: Extraction process of inner lip contours

10. Using a method described by [23] , the dip of the cupid's bow on the upper lip is localized to a point 'dp'. Let xt and xb be the two corners of the lips. Find the point xc that divides line xt–xb into equal segments. Select two boundary regions (depicted by red in Fig. 4A) that are within 20% of the distance of line xt–xb to the left and right of the point xc. Within this 20% region, find the lowest pixel. That will bet the dipping point dp of the upper lip cupid's bow, as shown in Figure 3.

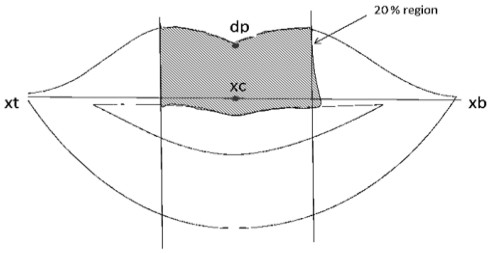

Figure 3: Selection of the dipping point of upper lip's cupid's bow

11. Steps 3 and 4 are repeated to fit curves of the outer lip and 9 more coefficients are found similarly. So to describe outer lip shape, now a vector (a4, b4,c4,a5,b5,c5,a6,b6.c6) is found.

This means each lip image can now be represented by a vector of 18 coefficients. Figure 4 shows a lip image with the 6 contours found.

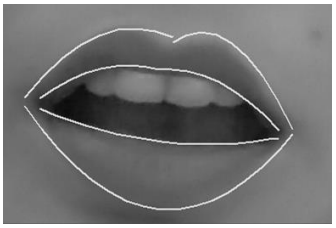

Figure 4: Screenshot of lip with contours showing

For pattern recognition, various machine learning tools are available, like Support Vector machines, KNN classifier and WEKA. For our experiments, we have used the Artificial Neural Networks tool available with the MATLAB package.

The extracted vectors of 18 coefficients (a1, b1, c1, a2, b2, c2, a3, b3, c3, a4, b4, c4, a5, b5, c5, a6, b6, c6) thus found in the previous section is finally normalized to values between 0 and 1 by the following formula:

$$n(i)=(c(i)-\min(i))/(\max(i)-\min(i)) \quad (4)$$

where the value of n(i) ranges from 0 to 1, and min(i) and max(i) denote the minimum and maximum value of the i-th vector, respectively.

The normalized vector for each lip image is then used as an input vector to the input layer of the single neural network system.

A Feedforward Neural Network was used with one hidden layer. The ANN had 18 input nodes, 15 hidden layer nodes and 3 output layer nodes (to identify 3 bangla vowels 'B'', 'A'' and 'H''). For both hidden and output layers, 'tansig' activation functions were used. The targets were set as follows:

'B': <1 0 0 >
'A' : <0 1 0>
'H' : <0 0 1>

The training parameters used were:
Epochs: 5000, goal: 10e-5

A diagram of the constructed network structure has been given in Figure 5.

Neural Network is trained and simulated using gradient descent method to minimize the error between the output values and the target values..

C. Analysis

The proposed method improves upon some of the drawbacks of the existing methods of contour extraction

It adds robustness and accuracy to image-based algorithms. Since it extracts the curvature of the lips, the results are independent of the size or quality of the picture, presence or absence of lipstick or mouth rotation. However, the images have to be front-facing.

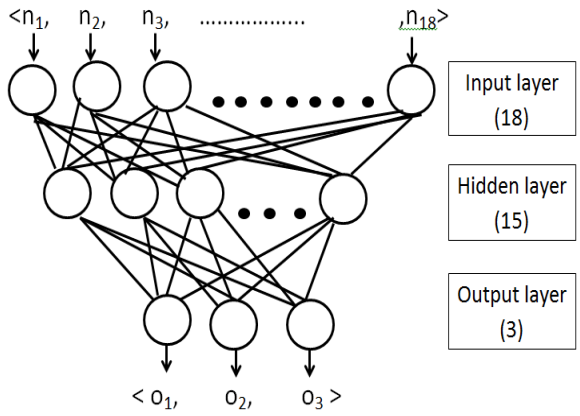

Figure 5: The three layer FeedForward Artificial Neural Network

The method does not require high quality or high resolution images. It does well with images taken on simple smartphone cameras or even images with small amount of noise.

In the same way, it does not need any iterations as in fuzzy clustering systems and saves time. The proposed method is based largely on Chen's method [3]. However, Chen's method extracts the features of the inner mouth only, whereas this method uses information of both inner and outer mouth to give a better representation of the mouth. Moreover, the preliminary steps of lip localization in this system used Chen's version of the red-exclusion method in combination with a conversion to HSV space to ensure a better extraction by fine-tuning the result obtained by the red-exclusion method alone.

Thus, the proposed system can be easily implemented in embedded systems such as Android or iOS to use with smartphones and tablets so that it can be used as and when needed.

IV. Results and Discussion

Our experiments were divided into three parts – A lip segmentation part, a contour extraction part and viseme recognition part. To demonstrate how conversions to different colour spaces affect the success of lip segmentation, an experiment was conducted by converting few images to different commonly used colour spaces for lip reading like RGB, HSV and YCbCr.

It was found that the Cr channel of YCbCr space and the S channel of the HSV space are most suitable for lip segmentation. However, on further experimentation, it was found that a combination of Saturation component and the Red Exclusion method gave the best results. This has been demonstrated in Figure 6.

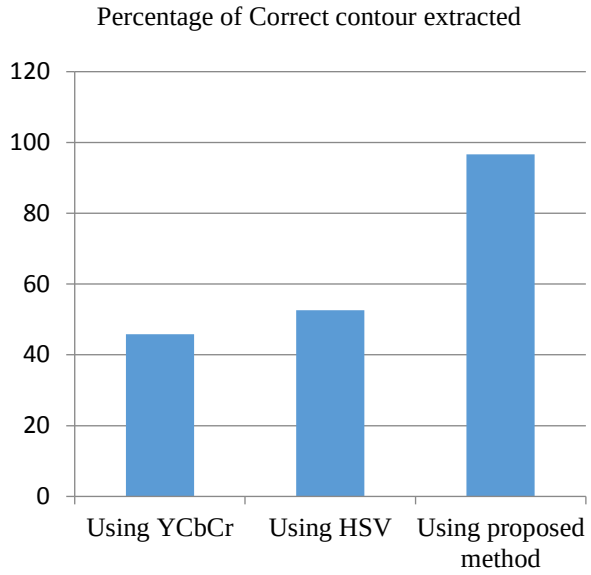

Figure 6: A comparison of the proposed method with image-based techniques like YCbCr and HSV methods

Experiments with different contour extraction algorithms were also performed, like ACM, ASM, AAM and curve fitting. In the end, curve fitting was found to be ideal, as it was simpler and computationally less expensive. Moreover, there was no need for prior initialization by the user with landmarks, as was necessary for the former 3 algorithms

Finally, for viseme recognition, Artificial Neural Networks were used. To evaluate our results, two ANNs were designed. One with 9 input nodes (representing inner lip features only), to test the accuracy using Chen's algorithm, and another with 18 inputs (representing both inner and outer lip features) to test the accuracy of the proposed method.

The initialized contour for the ACM method could not properly find the right side contour of the lip image for most of the images, as the snake failed to find the edge and entered into the object region. The Active Shape model was found to fail converging to the proper lip contour in most of the images. The AAM method was found to be too memory intensive and slow in processing.

However, the proposed algorithm was able to detect proper contours for almost 90% of the dataset images. Figure 7 shows a sample of some of the lip images with contours found.

For viseme recognition, two ANNs were constructed, one to see the result of using only inner mouth contours (a 9 feature vector) , and another to see the result of using both inner and outer lip contour (an 18 feature vector). ANN1 was used for the former and ANN2 was used for the latter.

There is no rule to determine the optimal number of neurons to be added in the hidden layer. However, through many experiments, it was found that the performance of the neural was optimal at 15 units. Adding any more neurons made no significant difference to the Mean Square Error.

Figure 7: A montage of some lip image contours found by the proposed method

To verify the validity of the proposed method, another single neural network was designed based on Chen's curve-fitting algorithm, for comparison. This network was fed only the inner mouth curves' coefficients as input. That means the input layer consisted of only 9 neurons. All other parameters were kept the same.

It was found that the error value during training for the former method was 0.112 at the end of 5000 iterations. Whereas for the proposed method, the error was only 0.072. This shows that using both inner and outer lip coefficients gives better results than only using inner lip coefficients.

A comparison was drawn between the accuracies of the two methods over the training and validation data sets. Results of this experiment are shown in Table 1.

Table 1
Results of Experiment

Viseme	Ratio of true positives using only inner lip contours (Feature Set 1)	Ratio of true positives using both inner and outer lip contours (Feature Set 2)
B	25 / 42	36/42
A	19/42	37/42
H	9/42	37/42
Accuracy	0.421	0.873

It was found that the result from former method (using feature Set 1) after Neural Network simulation was 42.1% while that achieved from the proposed method (with feature Set 2) was 87.3%.

V. CONCLUSION

This paper proposed a neural network as a multi-class pattern classifier to identify visemes of Bangla vowels being spoken. The success of a viseme classification system depends upon many factors, most importantly, the choice of lip-localization and contour-finding algorithm, the choice of features extracted and the pattern recognition system used. The paper used a combination of image-based techniques to localize the lip and quadratic coefficients of curvature of the lip contour as features for viseme recognition. Finally an Artificial Neural Network was trained to recognize the viseme spoken. The proposed system was found to have a good accuracy result of 87.3% for given dataset of Bangla visemes. Future works shall include identification of whole Bengali words or numbers spoken by tracking lip movements on video, so that it will be more useful in everyday applications.

REFERENCES

[1] L.P. Mei, "Interpretation Of Alphabets By Images Of Lips Movement For Native Language," Universiti of Teknologi, Malaysia, 2014.

[2] S. Werda, W. Mahdi and Hamadou, A., "Lip Localization and Viseme Classification for Visual Speech Recognition," International Journal of Computing and Information Sciences, Volume 5, No.1.,April 2007.

[3] Q. C. Chen, G. H. Deng , "An Inner Contour Based Lip Moving Feature Extraction Method for Chinese Speech," IEEE Xplore, March 2009.

[4] A. Sagheer, N. Tsuruta and R. Taniguchi, "Arabic Lip Reading System: A combination of Hypercolumn Neural Network Model with Hidden Markov Model", Proceedings of International Conference on Artificial Intelligence and Soft Computing, 2004, pp.311-316.

[5] A. N. Mishra, M. Chandra, "Hindi Phoneme-Viseme Recognition from Continuous Speech," International Journal of Signal and Imaging Systems Engineering, Volume 6 , No. 3., 2013.

[6] B. Naz and S. Rahim, "Audio-Visual Speech Recognition Development Era; From Snakes to Neural Network: A Survey Based Study," Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition Vol. 2, No. 1, 2011.

[7] S. Badura and M. Mokrys, "Lip detection using projection into subspace and template matching in HSV color space," in International Conference TIC, 2012.

[8] M. Kass, A. Witkin and D. Terzopoulos, "Snakes: Active contour models," International Journal of Computer Vision, 1987, pp. 321-331.

[9] U. Saeed and J. L. Dugelay, "Combining Edge Detection and Region Segmentation for Lip Contour Extraction," in AMDO'10 Proc. 6th International Conf. Articulated Motion and Deformable Objects, 2010, pp. 11-20.

[10] "Viola Jones Object Detection Framework", Wikipedia. [Online] From: https://en.wikipedia.org/wiki/Viola-Jones_object_detection_framework, 2015.

[11] J. M. Zurada, "Introduction to Artificial Neural Systems," 1992, pp. 1-21 .

[12] Md. Khalilur Rahman, "Neural Network using MATLAB (Powerpoint Presentation)," 2005.

[13] C. Bregler and K. Konig, "Eigenlips for robust speech recognition," Proc. ICASSP94, Adelaide, Australia, April 19-22, 1994, pp. 669–672.

[14] S. Gurbuz, K. Patterson, Z. Tufecki and J. N. Gowdy, "Lip-Reading from Parametric Lip Contours for Audio-Visual Speech Recognition," Eurospeech, 2001.

[15] R. Beale, and J. Finlay, "Neural Networks and Pattern Recognition in Human-Computer Interaction," Neural Networks and Pattern Recognition in Human-Computer Interaction, 1992, pp. 460.

[16] S. Lucey, S. Sridharan, V. Chandran, "Initialised Eigenlip Estimator for Fast Lip Tracking using Linear Regression," Proc. 15th International Conference on Pattern Recognition, vol.3, 2000, pp.178-18.

[17] Y. P. Guan, "Automatic Extraction of Lips based on Multi-scale Wavelet Edge Detection", in IET Computer Vision, vol.2, no.1, 2008, pp.23-33.

[18] N. Eveno, A. Caplier and P. Coulon,"Accurate and Quasi-automatic Lip Tracking," IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, 2004, pp. 706 – 715.

[19] H. Kalbkhani, and M. C. Amirani, "An Efficient Algorithm for Lip Segmentation in Color Face Images Based on Local Information," J. World. Elect. Eng. Tech 1(1), 2012, pp. 12-16.

[20] S.Badura and M. Mokrys, "Feature Extraction for Automatic Lip Reading System for Isolated Vowels," The 4th International Virtual Scientific Conf on Informatics and Management Sciences, March 23, 2015.

[21] I. Matthews, T. Cootes, "Extraction of Visual Features for Lip Reading", IEEE Transactions on Pattern Analysis and Machine Intelligence,Volume 24, February 2002.

[22] B. Hassanat, "Visual Speech Recognition,", Speech and Language Technologies,Volume 1, June 2011, pp.279-303.

[23] "Image Processing: Morphology-Based Segmentation using MATLAB with program code," [Online] From: www.code2learn.com/2011/06/morphology-based-segmentation.html

[24] S. H. Kang, S.H. Song, and S.H. Lee, "Identification of Butterfly Species with a Single Neural Network System," Journal of Asia-Pacific Entomology, 15(3), pp. 431-435.

[25] W. Rehman Butt, and L. Lombardi, "Comparisons of Visual Features Extraction Towards Automatic Lip Reading," University of Pavia, Italy. Researchgate, 2013