# A Survey-based Study on Lip Segmentation Techniques for Lip Reading Applications

Nahid Akhter, Amitabha Chakrabarty

Department of Computer Science and Engineering

BRAC University, 66 Mohakhali, Dhaka-1212, Bangladesh

***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 review of the different technologies used in lip reading and their evolution, from Active Contour Models to Artificial Neural Networks and Hidden Markov Model for temporal Viseme recognition. Special emphasis is placed on techniques in lip-contour detection and tracking, and a comparative study has been presented.

***Keywords*****—** *Speech Recognition; human-computer interaction; Viseme , Artificial Neural Networks.*

## 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 the entire human face [7] . Some applications of lip reading include crime fighting potential for computerized lip-reading, speech recognition systems in cars and lip reading systems in computer as an alternative to keyboard. This paper aims to give a run-down of some of the different lip reading techniques, technologies and algorithms that various computer scientists have incorporated thus far, make a comparison among them and present some suggestions.

This paper is organized as follows. First part furnishes the introduction given above. The second part deals with the various steps involved in lip reading. The third part enlists various techniques used in lip reading and their evolution. The fourth section gives an elaborate description of Artificial Neural Networks to be used for learned recognition of syllables. The fifth section provides a brief description of Hidden Markov Models for word recognition and the last section will provide a comparison between various lip contour detection methods.

## I. STEPS INVOLVED IN LIP READING

The entire process of lip reading can be broken down into a number of steps. These steps have been listed in chronological order in Fig 1. It starts with Image Acquisition by breaking the video into frames. This is followed by detection of the face, then the lip is detected from the face and the lip ROI is extracted by Lip segmentation. After that, techniques are used to extract important features of the lips and finally Viseme is recognized by a pattern recognition algorithm or technique and the word is recognize by some temporal pattern recognition technique. To detect a human face in a given frame, haarcascade classifiers in OpenCV or Viola Jones algorithm is in MATLAB is used [9].Another algorithm similar to Viola Jones algorithm is the KLT algorithm which can detect and even track tilted and rotated faces, which is not possible in case of Viola–Jones algorithm[10].

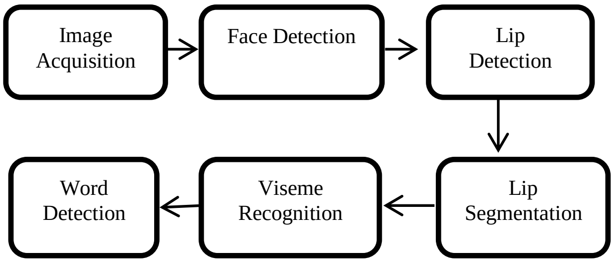

Fig 1. Steps in Lip Reading

For lip detection,Various color spaces such as RGB (Red-Green-Blue), HSI (Hue-Saturation-Intensity), and YCbCr (Luminance-Component blue- Component red) or L*a*b space are used. Pixels of lip area have stronger red component and weaker blue component than other facial regions as proposed in [43]. 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 [43] 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 histeq algorithm is used to enhance contrast so that the lip area appears much darker than the skin.

Lip segmentation techniques may be image-based, colour-based, or model based. The various techniques have been elaborated in Section 6 of this paper. The next step would be extracting the relevant information from the lip segment to be used in recognition of the viseme. Some research papers use pixels of the whole lip image or the just the inner pixels of the mouth as inputs. Other papers advocate the use of certain points on the lip such as the centre of the upper lip, distances between corners of the lip and a few pairs opposite points on the edges of the lip. Whichever technique is used, the end product should be a set of numbers that can be efficiently used as input for a Viseme recognizer.

Invariably and inarguably, the best method for recognizing lip visemes would be modern learning algorithms, called Artificial Neural Networks. These use a set of images to learn standard visemes corresponding to a particular lip image, so that it can recognize a viseme corresponding to a test image. Neural Network model like a Multilayer Perceptron is popularly used for pattern recognition as they can tolerate noise and, if trained properly, will respond correctly for unknown patterns. Sagheer et Al [42] proposed a Hypercolumn Neural Network model to recognize Arabic syllables. Finally, a series of visemes are strung together to recognize a meaningful word, this is done by using Hidden Markov Models.

## II. EVELUTION OF LIP READING TECHNIQUES

This section will review some of the popular lip reading techniques available in the literature. This will be just a brief of what each method is capable of doing using what kind of methods.

### *A. Snakes*

In 1988, Kass at el [5] 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, as illustrated in Fig2. These snakes lock onto nearby edges, localizing them accurately. Snakes consider different features for image energies like: color of pixels or sharpness of specified area, etc. and provide an account of visual problems like detection of edges, lines and subjective contours; motion tracking and stereo matching. They are guided by user-imposed constraint forces that navigate the snake near and around features of interest. This model is also called an ACM (Active Contour Model).

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. So, Hashimoto et al [13] introduced variation of the Active Contour Model known as Sampled-ACM. This model assumes area extraction problems as force balancing problems of sample points on the closed curves, which are controlled by three local forces: attraction Fa , Pressure Fp , and repulsion Fr. By calculating the sum of these three forces on each contour point, sampled-ACM can extract the area more rapidly.

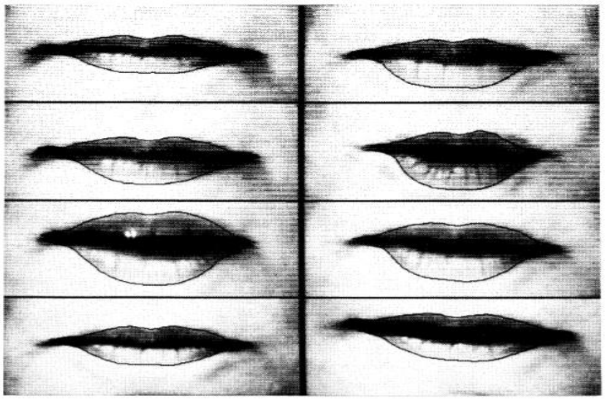

Fig.2.This diagram shows selected frames from a 2-second video sequence using snakes for motion tracking. After being initialized to the speaker's lips in the first frame, the snakes automatically track the lip movements with high accuracy.

Another problem was when the snake is not contacting the object boundary it will enter into the object region and then the repulsion force won't work. So, Sughara at. el. [14] introduced a new force called vibration factor, to improve accuracy against noises in image and combined hardware circuits in FPGA (Field Programmable Gate Array) [12] with the S-ACM vibration factor. This helped improve the area extraction function in standalone systems, but with only one given image.

In 2006, Toshio [15] proposed the Sampled-ACM with splitting characteristics. The proposed Sampled ACM reduced the number of memory accesses required and increased processing speed. Later, using the RASTA-PLP (Relative Spectral Transform - Perceptual Linear Prediction) method a system was built which was robust against any kind of distortions. The pattern recognizer is based on the first n principal components of a 24x16 gray-level matrix centered and scaled around the lips coded to get the outer boundary fairly with "Eigenlips" [16] in regard to the similar approach of Turk and Pentaland's "Eigenfaces" [17], and then it leads on to find the mutual information for audio-visual lip-reading using an MLP (Multi-Layer Perceptron) ANN.

Nowadays, Automatic speech recognition (ASR) is playing a vital role in the human computer interfaces (HCI). Speech recognizers of different kinds were developed in the last decade, like the Tangora system, which is now a STT (Speech to Text) engine [19] [20], created by IBM, which was a large-vocabulary natural-language isolated word recognition system. The first speech recognizer was developed in 1952, which could recognize a single spoken digit by eliminating signal distortion [18]. In 1988, Bahl, at the Watson Research Center described a new type method of word recognition from an acoustic representation of the word [21].

In 1984, Petajan [22] used simple image thresholding to extract binary mouth images, height, perimeter, area and width as visual features to produce their speech reading system. During the last decade, most systems have been based on AV-ASR. In 2000, IBM improved speech recognition by adding visual modality to the traditional audio for Large Vocabulary Continuous Speech Recognition.

### B. Neural Networks

Beala and Finaly [23] were the first to apply Neural Networks to the application of lip reading. Rumelhart et al [24], in 1986 used a back propagation system to train a neural network. This went on to be known as a BP network. A BP network can be used to learn and store a great deal of mapping relations of an input-output model. Its learning rule is to adopt the steepest descent method in which the back propagation is used to regulate the weight value and threshold value of the network to achieve the minimum error sum of square. This led to the evolution of systems like NETtalk (Sejnowski and Rosenberg, 1987) for pronunciation of English sentences.

Later, the TDNN (Time-delay Neural Network) network, designed by Stork [25] was introduced that deciphered the time delay and scalability of the system. This network was able to do automatic and acoustic lip reading efficiently in silent as well as acoustic noise environment. Then Yuhas et al. [26], in 1989 accomplished a vowel recognizer using the neural network for static images of mouth shape.

It is noticed that the ANN model developed is accurate enough to recognize the image even if the image is distorted or some portion of data is missing from the image. This model eliminates the long time-consuming process of image recognition.

### C. Hidden Markove Model

Later Hidden Markov Models (HMM) were merged with the NN. Further enhancement came about in the form of Modal Neural Network (MNN) [27] for the purpose of robust speech recognition, and Spiking neural network (SNN) designed to cope with neurons that fire multiple spikes in a multi-layer network.

In the classification of dynamic lip movements, geometrical information is extracted from video sequences. The performance of the geometrical-based method remains consistent and unaffected by rotational and brightness changes.

In [45], Yu, Jiang et al. proposed a sentence systematic approach to lip-reading whole sentences by using HMMs integrated with grammar. In this approach, a vocabulary of elementary words is considered. Based on the vocabulary, they define a grammar that generates a set of legal sentences. Each word of the basic vocabulary is modeled by an HMM and the individual HMMs are concatenated according to the rules of grammar. Each HMM corresponding to one of the basic words consists of six states as shown in Fig 3. The HMMs are trained using forward-backward algorithm based on Baum-Welch formula.

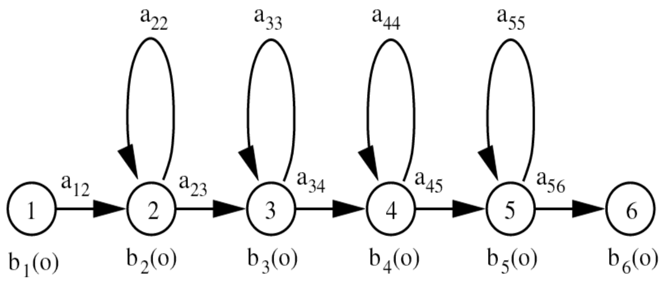

Fig 3. A six-state HMM. Here, O={$o_1$, $o_2$,…$o_6$} is a visual observation of a word, represented by a sequence of feature vectors; A={$a_{ij}$} is an NxN matrix of state transition probabilities from state i to state j; and B is a set of observation probabilities $b_j(o_t)$ for state j.

Simmons and Cox [46] developed an HMM based system that analyzed a small number of sentences to obtain several acoustic and visual training vectors. Then they created a fully connected 16 state discrete HMM, each state representing a particular vector quantized mouth shape, and producing 64 possible audio codewords. Subsequently, the trained HMM was employed in the Viterbi algorithm to generate the most likely visual state sequence, given the input audio observations.

## III. LIP SEGMENTATION TECHNIQUE

For lip reading system, feature extraction is considered as a crucial part. In general, there are two feature extraction methods: 1. “Pixel-based”, 2. “Lip Contour Based”. [2] Table 1 demonstrates the summary of feature extraction/ Lip segmentation methods.

| Model | Example | Distortion | Lip Extraction | Performance | Limitation |
|---|---|---|---|---|---|
| Image-Based | DCT, DWT, PCA | Real Environment | Visual only | High DCT is better than others | Restricted to illumination, mouth rotation, dimensionality |
| Model-based | ACM (snakes) , ASM, AAM Deformable templates | Noise and channel | Acoustic and visual | Low but robust, i.e. invariant to translation, rotation, scaling and illumination | Inner outer lip contours, colour of skin and lip matched, computationally expensive |

Image based techniques use the pixel information directly, the advantage is that they are computationally less expensive but are adversely affected by variation such as illumination. Under image-based techniques, there are colour-based and subspace-based techniques. It was found in [30] that the difference between red and green is greater for lips than skin and it was proposed to have a pseudo hue as a ratio of RGB values [31] have also proposed a RGB value ratio based on the observation that blue color plays a subordinate role so suppressing it improves segmentation. Color clustering has also been suggested by some, based on the assumption that there are only two classes i.e. skin and lips. However, if facial hair or teeth are visible, then this does not hold true [6].

In [34] a lip detector based on PCA (A Supspace-based technique) was proposed. Firstly outer lip contours are manually labelled on training data, PCA is then applied to extract the principal modes of contour shape variation, called eigencontour, finally linear regression is applied for detection.

LDA (Linear Discriminant Analysis) has been employed in [35] to separate lip and skin pixels, as shown in Fig 4. [36] have proposed a method in which a Discrete Hartley Transform (DHT) is first applied to enhance contrast between lip and skin, then a multi scale wavelet edge detection is applied on the C3 component of DHT.

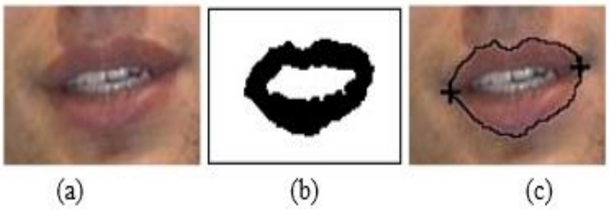

Fig4. (a) The original estimate of the mouth region. (b) Segmented lip region (black) using LDA. (c) The lip contour and the corners of the mouth

Model based techniques are based on prior knowledge of the lip shape and can be quite robust. Snakes have been a commonly used technique for lip segmentation, however, they need to be properly initialized. Moreover, they are unable to detect lip corners as they are located in low gradient regions. So, Eveno et al in [37] proposed a jumping snake which can be initialized far from the lip edge and the parameter adjustment is easy and intuitive[38] proposed a real time tracker that models the dynamic contours of lips using quadratic B-Splines learned from training data using maximum likelihood estimation algorithm. [39] have proposed Active Shape Models (ASM) and Active Appearance Models (AAM), which learn the shape and appearance of lips from training data that has been manually annotated. The introduction of deformable templates led on to proposal of a lip detection method based on Point Distribution Model (PDM) of the face. From [44] 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%. An example of an AAM fitting has been shown in Fig 5.

In addition, there are some hybrid techniques. These methods combine both image based and model based techniques. Majority of the hybrid techniques proposed in the literature 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.

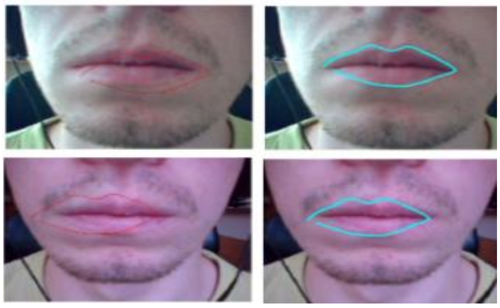

Fig 5. An example of AAM fitting. Left column proposes a negative case, right column proposes a positive case.

Usman Saeed and Jean-Luc Dugelay in [6] proposed a “fusion” of edge-based and region-based detection methods to carry out lip segmentation with comparatively better result than any of the two methods carried out individually. Here, given an image, it is assumed that a human face is present and already detected; the first step is to select the mouth Region of Interest (ROI) using the lower one third of the detected face. The next step involves the outer lip contour detection where the same mouth ROI is provided to the edge and region based methods. Finally the results from the two methods are fused to obtain the final outer lip contour. A flowchart of this system is given in Fig 6.

Another method used for lip detection is Fuzzy clustering. This was applied in [32] by combining color information and spatial distance between pixels in an elliptical shape function. [33] have used expectation maximization algorithm for unsupervised clustering of chromatic features for lip detection in normalized RGB color space. Nowadays, lip image segmentation is also done using Fuzzy Clustering as outlined by Leung et al [40]. In this method, multiple clusters are adopted to model the background region sufficiently and a spatial penalty term is introduced to effectively differentiate the non-lip pixels that have similar color features as the lip pixels but located in different regions. Experimental results demonstrate that the proposed algorithm has good segmentation results over other segmentation techniques.

## IV. CHALLENGES

Inspite of all these techniques, lip reading comes with its challenges [7]. Firstly, different sounds can be made with the lips in the same position. Secondly, moustaches and beards make lip reading more difficult or even impossible. Thirdly, we form many English sounds in the middle of our mouth, others come from the back of our mouth and even in our throat. These latter are impossible to speech read so far. Moreover, there are numerous homophones in English. Words as different as "queen" and "white" look the same on a person's lips. This accounts to very less accuracy (< 30%) in such a system. Other challenges include fast speech, poor pronunciation, bad lighting, faces turning away, hands over mouths, etc. Most importantly, In order for speech reading to be effective we have to know the subject being discussed.

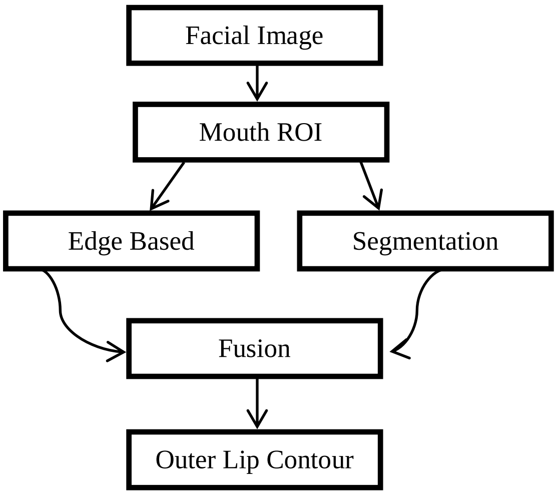

Fig 6. Overview of Saeed and Dugelay's proposed fusion method for lip segmentation

## V. CONCLUSION

This paper was created to outline the different research works that have been done in the field of lip reading down the years, while focusing on the various techniques used in lip segmentation, which is a very important step in the process of lip reading. For face and lip localization, Viola Jones algorithm has been unanimously proclaimed as the best way. However, for lip segmentation, there are various methods available, the most primitive of which have been snakes and Active Contour Models. Other techniques include image-based techniques that use differences in colour components of the lips and skin in various colour spaces as well as fuzzy clustering techniques. Techniques that combine both image and model-based methods for lip segmentation were also outlined. Of all the techniques, Artificial Neural Networks for viseme recognition and Hidden Markov Models for Speech Recognition have been found to be the best in modern technologies.

ANNs have been used to detect the viseme being spoken based on advanced learning of previous patterns. Nowadays different researchers are combining different probabilistic, statistic and ANN techniques to provide appreciable and accurate error free automatic Lip-reading systems. HMMs add to the ANNs functionality by providing solutions to the problem of both the acoustic and temporal modeling. Using HMMs, it is now possible to some extent, make predictions of words uttered by evaluating a series of visemes that are stochastically distributed.

Although ANNs try to ape the human central nervous system, and HMMs try to mathematically evaluate real-life phenomena, the performance of current speech recognition systems is far below that of humans. More research needs to be directed to this field as it will have massive impact in real-world applications like helping the deaf understand spoken words crime-fighting institutions pick up clues when auditory input is limited.

## REFERENCES

[1] L. R. .Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models", IEEE ASSP Magazine, 1986.

[2] Bushra Naz and Sabit Rahim, "B 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.

[3] Toshio Miyaki, Sughara et al, "Active Contour Model with Splitting Characteristics for Multiple Area Extractions and its Hardware Realization", SICE-ICASE International Joint Conference 2006.

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

[5] Michael Kass et al. "Snakes: Active contour models," International Journal of Computer Vision, pages 321-331, 1987.

[6] Usman Saeed and Jean-Luc Dugelay "Combining Edge Detection and Region Segmentation for Lip Contour Extraction", AMDO'10 Proceedings of the 6th international conference on Articulated motion and deformable objects, Pages 11-20.

[7] Lai Pei Mei. "Interpretation Of Alphabets By Images Of Lips Movement For Native Language", Universiti of Teknologi, Malaysia, 2014.

[8] Shi-Lin Wang et al, "Robust lip region segmentation for lip images with complex background", Science Direct, pages 3481 – 3491, 2007.

[9] "Viola Jones Object Detection Framework", Wikipedia, 2015.

[10] 'Neetu Saini and Hari Singh "Comparison of two different approaches for multiple face detection in color images', International Journal of Innovative Research in Eletrical, Eletronics, Instrumentation and Control Engineering, Vol 3, Issue 1, pages 2321-2004, 2015

[11] Jacek M. Zurada, 'Introduction to Artifiial Neural Systems", 1992 edition, pages 1-21

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

[13] M.Hashimoto, H.Kinoshita and Y.Sakai, "An Object Extraction Method Using Sampled Active Contour Model," IEICE Trans. D-II, Vol.J77-D-II, No.11, pp.2171-2178, 1994.v

[14] K.Sugahara, T.Shinchi and R.Konishi, “Active Contour Model with Vibration Factor,” IEICE Trans. DII, Vol.J80-D-II, No.12, pp.3232-3235, 1997.

[15] T.Miak,T.Kawamura, K.Sughara; “Active Contour Model with Splitting Characteristics for Multiple Area Extractions and its Hardware Realization,” SICE-ICASE International Joint Conference, 2006.

[16] C. Bregler and Y. Konig, “Eigenlips for robust speech recognition,” in Proc. of ICASSP94, Adelaide, Australia, April 19-22. 1994, pp. 669–672.

[17] M. Turk and A. Pentland, “Eigenfaces for Recognition,” Journal of Cognitive Neuroscience, Volume 3, Number 1, MIT 1991.

[18] S.Gurbuz, E.K.Patterson, Z.Tufekci, and J.N.Gowdy Lip-Reading from Parametric Lip Contours for Audio-Visual Speech Recognition;; Department of electrical and computer Engineering Cemson University;Clemson,SC 29634, USA.

[19] ‘Pioneering Speech Recognition’ from www-03.ibm.com

[20] ‘Speech Recognition”, Wikipedia, 2015

[21] ‘Automatic Speech and Speaker Reognition-Advanced Topics’ ,third Edition, Chin-Hui Lee, Frank K. Soong, Kuldip Paliwal, Spinger website, 1999

[22] E. D. Petajan, “Automatic lipreading to enhance speech recognition,” in Proc. Global Telecomm. Conf., Atlanta, GA, 1984, pp. 265–272.

[23] Russell Beale , Janet Finlay,“Neural networks and pattern recognition in human-computer interaction,” Neural networks and pattern recognition in human-computer interaction, Pages: 460, 1992.

[24] R. P. Lippmann, “Review of Neural Networks for Speech Recognition, Readings in Speech Recognition,” A. Waibel and Morgan Kaufmann Publishers, pp. 374-392, 1990.

[25] D.G. Stork, G.WolfP and E. Levinet, “Neural network lipreading system for improved speech recognition,” IJCNN, 1992.

[26] B.P.Yuhas, M. H.Goldstein, J.R. and T. J. Sejnowski, “Integration of acoustic and visual speech signals using neural networks,” IEEE Communications Magazine, 1989.

[27] H.Kabre, “Robustness of a chaotic modal neural recognition network applied to audio-visual speech,” Neural Networks for Signal Processing, page(s): 607 - 616, Sep 1997.

[28] “Inference in Hidden Markov Models”, Olivier Cappe, Eric Moulines, Tobias Ryden, Springer, Page 42, 2005

[29] Govind, ‘Introduction to Hidden Markov Models’ , Lecture 12, CEDAR, Buffalo (Powerpoint Presentation)

[30] Hulbert, A. Poggio, T.: Synthesizing a Color Algorithm from Examples. Science. vol. 239, pp. 482-485 (1998)

[31] Canzlerm, U., Dziurzyk, T.: Extraction of Non Manual Features for Video based Sign Language Recognition. In: Proceedings of IAPR Workshop, pp. 318-321 (2002)

[32] Leung, S.-H., Wang, S-L., Lau, W.-H.: Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. In: IEEE Transactions on Image Processing, vol.13, no.1, pp.51-62 (2004)

[33] Lucey, S., Sridharan, S., Chandran, V.: Adaptive mouth segmentation using chromatic features. In: Pattern Recogn. Lett, vol. 23, pp. 1293-1302 (2002)

[34] Lucey, S., Sridharan, S., Chandran, V.: Initialised eigenlip estimator for fast lip tracking using linear regression. In: Proceedings. 15th International Conference on Pattern Recognition, vol.3, pp.178-181 (2000)

[35] Nefian, A., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., Murphy, K.: A coupled HMM for audio-visual speech recognition. In: Proc. ICASSP, pp. 2013–2016 (2002)

[36] Guan, Y.-P.: Automatic extraction of lips based on multi-scale wavelet edge detection. In: IET Computer Vision, vol.2, no.1, pp.23-33 (2008)

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

[38] Kaucic, R., Dalton, B., Blake, A.: Real-Time Lip Tracking for Audio-Visual Speech Recognition Applications. In: Proceedings of the 4th European Conference on Computer Vision, vol. II (1996).

[39] Cootes, T. F.: Statistical Models of Appearance for Computer Vision. Technical report, University of Manchester (2004).

[40] Shu-Hung Leung, Shi-Lin Wang, and Wing-Hong Lau, “Lip Image Segmentation Using Fuzzy Clustering Incorporating an Elliptic Shape Function”, IEEE Transactions on Image Processing, Vol. 13, No. 1, January 2004.

[41] R. Padilla, C. F. F. Costa Filho and M. G. F. Costa, “Evaluation of Haar Cascade Classifiers Designed for Face Detection”, World Academy of Science, Engineering & Technology; 2012, Issue 64, p362.