Estimation of the Area of Mouth Opening during Speech Production

Sudipan Saha*
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai-400076
Maharashtra, India
sudipan@ee.iitb.ac.in

Prem C. Pandey
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai-400076
Maharashtra, India
pcpandey@ee.iitb.ac.in

ABSTRACT

Estimation of the area of mouth opening from a video of the speaker's face and processing of the simultaneously recorded speech signal can be used for estimation of vocal tract shape during speech production. A technique is presented to estimate the area of mouth opening, using a processing based on contrast enhancement with multi-threshold binarization and connected component detection for reducing the errors in estimating the vertical opening due to the presence of teeth and tongue in mouth opening. The method is tested on vowel utterances from 12 speakers.

Keywords

Lip contour, mouth opening, speech training aids, vocal tract shape

1. INTRODUCTION

Children with hearing impairment face difficulty in acquiring speech production due to lack of auditory feedback. A visual feedback of articulatory efforts is considered to be useful for speech training [18, 5, 4, 2]. A display of time-varying vocal tract configuration can be generated from an estimation of the vocal tract area function, i.e. area of the opening in the oral cavity along its length from the lips to the glottis. Although this function can be estimated using one of the direct imaging methods, e.g. X-ray [15] , MRI [21] etc. or the techniques based on acoustic measurements [20], these methods are not suitable for use in speech training aids as they interfere with speech production. Most of the aids use Wakita's LPC based inverse filtering of the speech signal for estimation of the vocal tract area function [27]. In this method, the oral cavity is modeled as an acoustic tube with a series of equal length sections with variable cross-sectional area, and analysis of windowed speech signal is used to estimate ratios of the areas at the section interfaces. Vocal tract area function is obtained from these ratios by assuming a constant area at the glottis end. This assumption is one of the main limitations of the technique [28] and introduces errors because of variation in the area at the glottis end during speech production. It has been shown that estimation of the vocal tract area function can be improved by using the area of mouth opening as the reference area of the section at the lip end [17]. Do *et al.* [6] proposed a method to estimate mouth opening from the speech signal, but performance of the method was found to be dependent on speech utterances and hence it is not suitable for speech training. The area of mouth opening can be estimated by processing of the frames of the video signal acquired simultaneously along with the speech signal. In addition to its application in speech training, estimation of the area of mouth opening has also been reported to be useful for lipreading [12], automatic speech recognition [11, 13], and facial expression recognition [19, 14].

Estimation of the area of mouth opening involves tracking the inner contour of the lips. Several methods for tracking the outer contours of the lips, involving separation of the lips from the rest of the face, have been reported [16, 3, 10, 1, 7, 24]. However, tracking the inner contour of the lips poses a different set of challenges as it involves distinguishing the lips from the region inside the mouth opening, which may have a high shade variability and may have the presence of the teeth and the tongue. Nayak *et al.* [17] proposed a method based on template matching, assuming the mouth opening to correspond to an asymmetrical ellipse. The method was tested for different utterances from a number of speakers. It was found that the presence of teeth and the tongue inside the mouth opening occasionally introduced errors in the estimated areas. Hence, a method is needed which can work in the presence of teeth and tongue between the lips. Speech training typically involves acquiring short utterances of speech followed by visual feedback of articulatory efforts in slow motion. Although this application does not require real-time processing, it is desirable that the delay in the feedback is not significantly larger than the utterance duration.

A technique is presented to estimate the area of mouth opening by estimating its horizontal and vertical opening, using a method which is not very sensitive to variation in the shade of the region inside the mouth opening and the presence of the teeth and the tongue. The input frame is processed to locate the mouth sub-image, which is further processed to obtain a line segment corresponding to the widest

*Corresponding author

horizontal opening. It is used to divide the sub-image into upper and lower parts and the vertical opening at the center of the line segment is located in each part. These values are used to estimate the area of mouth opening. The method is tested on vowel utterances from 12 speakers.

2. PROCESSING FOR LOCATING THE MOUTH SUB-IMAGE

From each frame in the video signal, face sub-image is obtained by Viola-Jones' method [26]. This method works on the assumption that the eyes are darker than the cheeks and bridge of the nose. While using this method, a few smaller regions also were found to get incorrectly identified as face sub-image in some of the frames. This problem was solved by accepting the region with the largest area from among all such regions as the face sub-image. An example of obtaining the face sub-image from the input frame is shown in Fig. 1(b).

Face sub-image is used as input to obtain mouth sub-image by using the method by Hsu *et al.* [8]. The color space $\mathrm{YC_bC_r}$ is used in this method, with the assumption that C_r is greater than C_b in the mouth region. The equations for conversion from RGB to $\mathrm{YC_bC_r}$ are as follows:

$$Y = 0.299R + 0.587G + 0.114B \tag{1}$$

$$C_b = -0.169R - 0.331G + 0.500B + 128 \tag{2}$$

$$C_r = 0.500R - 0.418G - 0.081B + 128 \tag{3}$$

From these values, C_r^2 and C_r/C_b are calculated and normalized to the range [0,255], the same as that of R, G, B. These normalized values of pixels in the face sub-image $\mathcal{F}$ are transformed according to the equation as given in [8] to calculate "mouthmap" as

$$\text{mouthmap} = C_r^2(C_r^2 - \eta\frac{C_r}{C_b})^2 \tag{4}$$

where η is given by

$$\eta = 0.95(\frac{\sum_{(x,y)\in\mathcal{F}} C_r(x,y)^2}{\sum_{(x,y)\in\mathcal{F}} \frac{C_r(x,y)}{C_b(x,y)}}) \tag{5}$$

It is observed that the mouthmap attains high value on the location of the lower lip as shown in Fig. 1(c). A high intensity region is defined by taking pixels falling within the highest 1 percentile of the lower segment of the face sub-image. The leftmost, rightmost, topmost, and bottommost of these pixels are used to define a rectangular region approximately corresponding to the lower lip. This rectangular region is extended 1.25 times its length on both left and right sides, 1.25 times its height on the bottom side, and 2.5 times its height on the upper side, as described in [17], to obtain the mouth sub-image. Fig. 1(d) shows the output of this step. Due to extension on the upper side, the detected mouth sub-image occasionally included the nostril region, as shown in Fig. 2(a). To solve this problem, an average of the heights of the detected mouth sub-images across the frames is obtained. In frames with detected sub-image height being greater than 1.5 times the average height, extension factor for the upper side is set as 2 in place of 2.5. The mouth sub-image after this corrective step is shown in Fig. 2(b).

3. ESTIMATING THE AREA OF MOUTH OPENING

The mouth sub-image is processed for obtaining a line segment corresponding to the maximal horizontal opening, assuming that it is aligned with the horizontal orientation of the image. At the center of this line segment, the vertical opening is estimated. These values are used to estimate the area of mouth opening. The methods used are described in the following subsections.

3.1 Estimation of the horizontal opening

To track the horizontal opening, image is first converted into gray-scale by taking the mean of the R, G, B values as

$$P = (R + G + B)/3 \tag{6}$$

Sum of the pixel values in each row is calculated as

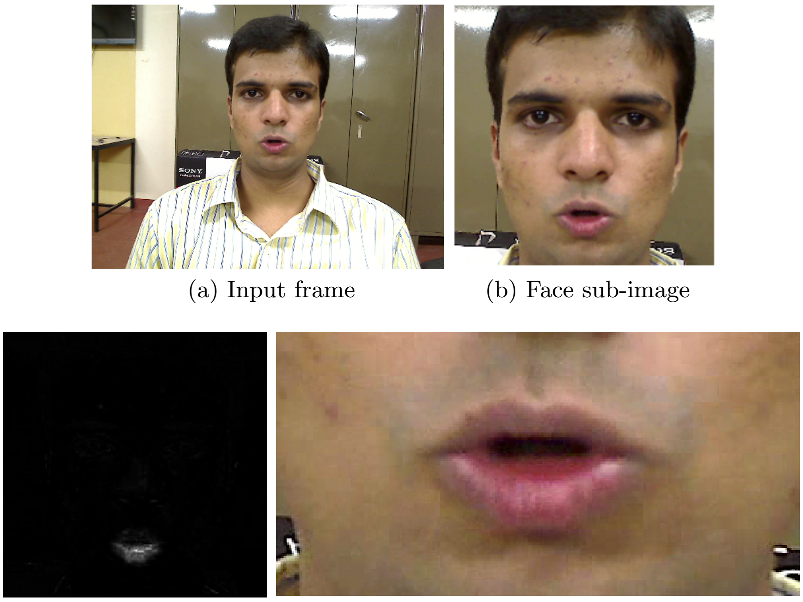
$$S(y) = \sum_{x=1}^{n_x} P(x,y) \tag{7}$$

where n_x is numbers of pixels in the row. Most of the pixels in the region inside the mouth opening have very low values. The row with the lowest sum, y_{hor}, is taken as the row corresponding to the maximal opening of the mouth.

Next we locate the left and right end points of the horizontal opening. The image $P(x,y)$ is thresholded to obtain a binary image, with a threshold set at 20% of the 99 percentile of the pixel values. The pixels with values greater than the threshold are set as black and others are set as white. This operation converts most of the pixels inside the mouth opening as white and almost all outside as black. Small unconnected white patches are removed by morphological operation of opening [9]. Let, the resulting binarized image be denoted as $T(x,y)$. In this binary image, the y_{hor}th row is scanned, and the leftmost white pixel position x_{left} is taken as the leftmost point of the horizontal opening. Similarly, the rightmost white pixel position is taken as x_{right}. These two positions are used to find the length of the maximum horizontal opening and its center. An example of the results from the processing steps for vowel utterances /i/ and /u/ are shown in Fig. 3 and Fig. 4, respectively.

3.2 Estimation of the vertical opening

To estimate the vertical opening at the center of the maximum horizontal opening, as estimated in the previous step, a method for distinguishing the inner lip contour from the teeth and the tongue is needed. The color conversion schemes for lip tracking [29, 30, 11, 17, 25] have been generally devised for distinguishing the lips from the rest of the face but they do not help in distinguishing the lips from the teeth and the tongue inside the mouth opening. Zhang *et al.* [30] used G and B for lip tracking. Wark *et al.* [29] used R/G, while Lewis and Powers [11] used $\log(G/B)$. Vezhnevets [25] proposed the use of CIE-LUV for lip pixel detection, and Nayak *et al.* [17] proposed to use the $R - 0.5G - 0.25B$ values for distinguishing the inner lip contour. Experimenting with different conversions showed that a color conversion itself can not be effective in distinguishing lips from teeth and tongue, and we need to use post-conversion processing. We propose a processing based on contrast enhancement using multi-threshold binarization and connected component



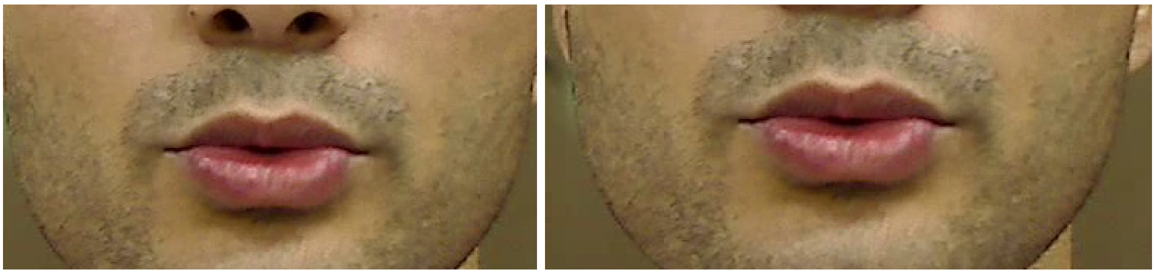
(a) Input frame

(b) Face sub-image

(c) Face sub-image plotted using “mouthmap”value

(d) Mouth sub-image

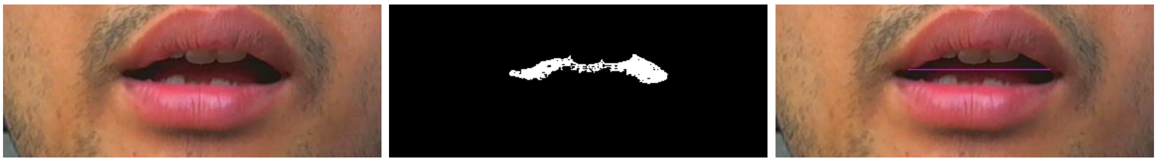
Figure 1: Obtaining face sub-image and mouth sub-image from input frame.



(a) Extending top by 2.5 times height

(b) Extending top by 2 times height

Figure 2: Mouth sub-image obtained with extending the top side by factor 2.5 and factor 2.

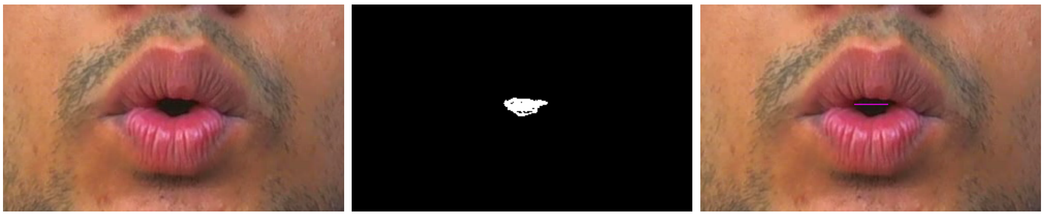


(a) Mouth sub-image

(b) Dark region inside mouth

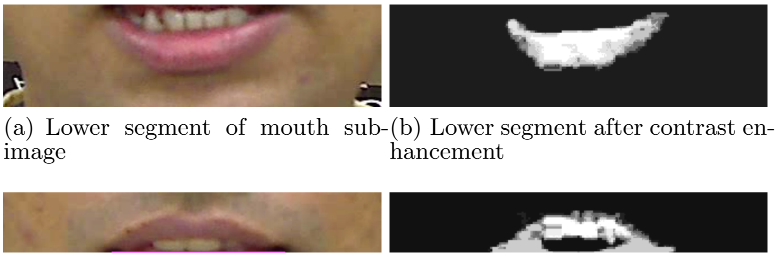
(c) Horizontal opening

Figure 3: Horizontal opening tracking in utterance of vowel /i/.



(a) Mouth sub-image (b) Dark region inside mouth (c) Horizontal opening

Figure 4: Horizontal opening tracking in utterance of vowel /u/.



(a) Lower segment of mouth sub-image (b) Lower segment after contrast enhancement

(c) Upper segment of mouth sub-image (d) Upper segment after contrast enhancement

Figure 5: Contrast enhancement process.

detection. The mouth sub-image is converted into Q channel of YIQ space, as used in [23] for lip tracking, with the help of the conversion equation

$$Q = 0.211R - 0.522G + 0.311B \tag{8}$$

The converted sub-image is segmented by the line segment corresponding to maximum horizontal opening. As the range of the values of the pixels corresponding to the lips in the lower and upper segments of the sub-image generally differ, these are processed separately.

Let us first consider the lower segment of the sub-image. The pixel Q values are normalized to the range [0,1], with 1 corresponding to 99 percentile of the pixel values and saturation for higher values. The pixels in the lip region were found to have values very close to 1. Further, these pixels have a gradual variation in the values. Hence such group of pixels can be classified together by binarizing the image and then detecting the connected components in it. A connected component is defined as a group of adjacent pixels which have same intensity value [22]. The pixels corresponding to the teeth and the tongue, which lie in close proximity to the lips, may also posses these two characteristics, but not simultaneously. The following iterative scheme is used to increase the contrast between the pixels corresponding to the lip and the rest.

1. Set an initial threshold value (say 0.99).
2. Using the set threshold, binarize the image.
3. Detect the connected components in the binarized image. Retain the connected component with the maximum number of pixels. Set rest of the pixel values as zero.
4. Reduce the threshold by a constant step size (say 0.01) and go to step 2 till we have reached a final threshold (say 0.90).
5. Sum the binarized values of the corresponding pixel positions from all the iterations.

After this processing, the value of a pixel is equal to the number of iterations in which it gets selected. Therefore, the pixels corresponding to the lip region have much higher values than the others. We use a similar set of operations on the upper segment of the sub-image. An example of the the contrast enhancement on both segments is shown in Fig. 5.

From the center of the horizontal opening in the lower segment, we calculate the first difference along the vertical and the position of the maximum first difference is taken as the position of the inner boundary of the lower lip and it is marked as y_{low}. Same process is followed on the upper segment to mark the inner boundary of the upper lip as y_{high}. The detected vertical opening for vowel utterance /a/ is shown in Fig. 6.

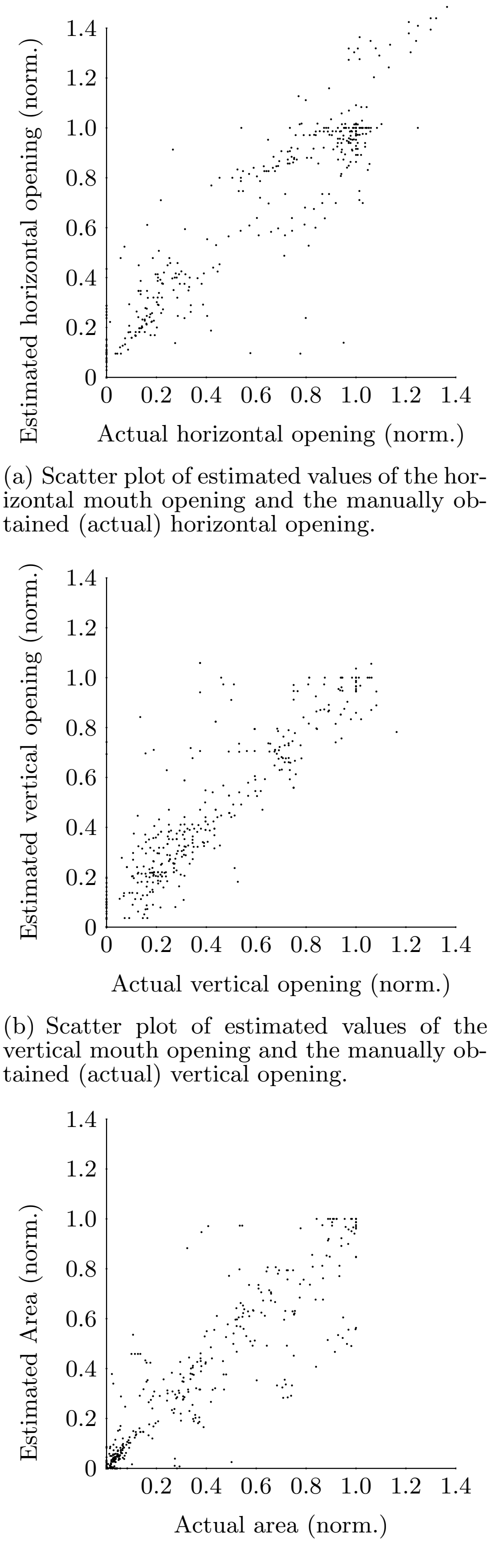
3.3 Area estimation

The vertical position of the horizontal opening of the mouth is given as y_{hor} and its left and right ends are given as x_{left} and x_{right}. The uppermost and the lowermost points of the mouth opening are given by y_{high} and y_{low} respectively. These positions are used for calculating the area of mouth opening by modeling it as two semi-ellipses with a common major axis.

4. RESULTS

The technique was tested on video recordings of speech production of vowel utterance /aiu/ by 12 male speakers[1],

[1] S. Saha. Video recordings of vowel utterances. SPI Lab, EE Dept., IIT Bombay, 2012. URL-`http://www.ee.iitb.ac.in/~spilab/sudipan/database_video`



(a) Scatter plot of estimated values of the horizontal mouth opening and the manually obtained (actual) horizontal opening.

(b) Scatter plot of estimated values of the vertical mouth opening and the manually obtained (actual) vertical opening.

(c) Scatter plot of estimated values of the area of mouth opening and the manually obtained (actual) values.

Figure 7: Scatter plot of the estimated values and the manually obtained (actual) values of the parameters of mouth opening: (a)horizontal opening, (b)vertical opening, (c)area of mouth opening

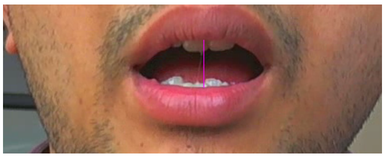


Figure 6: Estimated vertical opening

recorded using a 2 megapixels webcam at 15 frames per second. Presence of teeth and tongue in the region inside the mouth opening did not affect the area estimation. The horizontal opening values, the vertical opening values, and the area values estimated by processing of the video frames were compared with the corresponding manually estimated values. The scatter plot between these two sets of values is shown in Fig. 7, with the manually estimated values labeled as "actual" values. For each speaker, the area values were normalized with respect to the maximum area during vowel /a/. The horizontal opening values and the vertical opening values were normalized with respect to the horizontal opening value and the vertical opening value corresponding to the frame with maximum area, respectively. As the frame with maximum area need not have the maximum horizontal opening or maximum vertical opening, the normalized horizontal opening values and the normalized vertical opening values are greater than one in some cases. We see three distinct clusters, with normalized values of mouth opening areas of 1 for /a/, $0.5 - 0.8$ for /i/, and near 0.1 for /u/ (as reported earlier [17]). Other values correspond to the transitions between the vowels. Results are satisfactory for all the three clusters, except that estimated area values were lower than the actual values in some of the frames corresponding to /a/ and /i/. Estimated vertical opening values were higher than the actual values in some frames with small vertical opening. The correlation coefficients between the manually estimated values and those estimated using the proposed method are 0.93, 0.91, and 0.91 respectively, for horizontal opening length, vertical opening length, and the area of mouth opening.

5. CONCLUSION

A technique for estimation of the area of mouth opening during speech production has been presented. It has been devised to reduce the errors due to the presence of the teeth and the tongue within the mouth opening. The technique has been tested on utterances from 12 speakers and the estimated values are found to be in close agreement with the manually estimated values. It needs to be more extensively tested for a larger number of speakers with different conditions of illumination. The processing of the video frames needs to be integrated with the processing of the simultaneously recorded speech signal for estimation of the vocal tract area function for its use in speech training aids.

6. ACKNOWLEDGEMENT

This research is supported by the project "National Program on Perception Engineering," sponsored by the Department of Information Technology, MCIT, Government of India.

7. REFERENCES

[1] T. Akashi, Y. Wakasa, K. Tanaka, S. Karungaru, and M. Fukumi. Evolutionary video processing for lip tracking. *IC-MED*, 2(2):111–125, 2008.

[2] N. D. Black. Application of vocal tract shapes to vowel production. In *Proc. 10th Int. Conf. IEEE Engg. Med. Biol. Soc., New Orleans*, pages 1535–1536, 1988.

[3] S. W. Chin, K. P. Seng, L. Ang, and K. H. Lim. New lips detection and tracking system. In *Proc. Int. MultiConf. Engineers and Computer Scientists, Hong Kong*, pages 874–879, March 2009.

[4] R. G. Crichton and F. Fallside. Linear prediction model of speech production with applications to deaf speech training. *IEE Control and Science*, 121:865–873, 1974.

[5] P. M. T. de Oliveira and M. N. Souza. Speech aid for the deaf based on a representation of the vocal tract: the vowel module. In *Proc. 19th Int. Conf. IEEE Engg. Med. Biol. Soc., Chicago*, pages 1757–1759, 1997.

[6] C. Do, A. Aissa-El-Bey, D. Pastor, and A. Goalic. Area of mouth opening estimation from speech acoustics using blind deconvolution technique. In *Proc. Int. Conf. Audio-Visual Speech Process., Norwich*, pages 80–85, 2009.

[7] N. Eveno, A. Caplier, and P. Y. Coulon. A new color transformation for lips segmentation. In *4th Workshop Multimedia Signal Process., Cannes*, pages 3–8, 2001.

[8] R. Hsu, M. Abdel-Mottaleb, and A. K. Jain. Face detection in color images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):696–706, May 2002.

[9] A. K. Jain. *Fundamentals of Digital Image Processing.* Prentice Hall, Englewood Cliffs, NJ, 1989.

[10] J. Kadlec. Lip detection in low resolution images. In *Proc. 10th Conf. Competition Student EEICT, Brno*, pages 303–306, 2004.

[11] T. W. Lewis and D. M. W. Powers. Lip feature extraction using red exclusion. In *Selected Papers Pan-Sydney Workshop Visualisation*, pages 61–67. Australian Computer Society, 2001.

[12] M. Li and Y. Cheung. A novel motion based lip feature extraction for lip-reading. In *Proc. Int. Conf. Computational Intelligence and Security, Suzhou*, pages 361–365. IEEE Computer Society, December 2008.

[13] C. B. O. Lopes, A. L. Goncalves, J. Scharcanski, and C. R. Jung. Color-based lips extraction applied to voice activity detection. In *Proc. 18th IEEE Int. Conf. Image Process., Brussels*, pages 1057–1060, 2011.

[14] Y. Moses, D. Reynard, and A. Blake. Determining facial expressions in real time. In *Proc. 5th Int. Conf. Computer Vision, Cambridge*, pages 296–301, 1995.

[15] K. G. Munhall, E. Vatikiotis-Bateson, and Y. Tohkura. X-ray film database for speech research. *J. Acoust. Soc. Am.*, 98(2):1222–1224, 1995.

[16] J. C. Nascimento and J. S. Silva. Lip contour tracking using multiple dynamic models on a manifold. In *Proc. 18th IEEE Int. Conf. Image Process., Brussels*, pages 2369–2372, September 2011.

[17] N. S. Nayak, R. Velmurugan, P. C. Pandey, and S. Saha. Estimation of lip opening for scaling of vocal tract area function for speech training aids. In *Proc. 18th National Conf. Comm., Kharagpur*, pages P23.1–5, February 2012.

[18] S. H. Park, D. J. Kim, J. H. Lee, and T. S. Yoon. Integrated speech training system for hearing impaired. *IEEE Trans. Rehabil. Eng.*, 2(4):189–196, 1994.

[19] C. Saha and K. Ghosh. Estimation of facial expression intensity from a sequence of binary face images. In *Proc. Int. Conf. Image Informa. Process., Shimla, Himachal Pradesh*, pages 1–6, November 2011.

[20] M. R. Schroeder. Determination of the geometry of the human vocal tract by acoustic measurements. *J. Acoust. Soc. Am.*, 41(4):1002–1010, 1967.

[21] B. H. Story, I. R. Titze, and E. A. Hoffman. Vocal tract area functions from magnetic resonance imaging. *J. Acoust. Soc. Am.*, 100(1):537–554, 1996.

[22] R. Szeliski. *Computer Vision: Algorithms and Applications.* Springer, London, 2011.

[23] N. S. Thejaswi and S. Sengupta. Lip localization and viseme recognition from video sequences. In *Proc. 14th National Conf. Comm., Mumbai*, pages 456–460, 2008.

[24] Y. Tian, T. Kanade, and J. F. Cohn. Robust lip tracking by combining shape, color and motion. In *Proc. 4th Asian Conf. Computer Vision, Taipei, Taiwan*, pages 1040–1045, January 2000.

[25] V. Vezhnevets. Face and facial feature tracking for natural human-computer interface. In *Proc. Int. Conf. Graphicon, Nizhny Novgorod, Russia*, pages 86–90, 2002.

[26] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vis.*, 57(2):137–154, 2004.

[27] H. Wakita. Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. *IEEE Trans. Audio Electroacoust.*, 21(5):417–427, 1973.

[28] H. Wakita. Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art. *IEEE Trans. Acoust., Speech, Signal Process.*, 27(3):281–285, 1979.

[29] T. Wark, S. Sridharan, and V. Chandran. An approach to statistical lip modelling for speaker identification via chromatic feature extraction. In *Proc. 14th Int. Conf. Pattern Recog., Brisbane*, pages 123–125, 1998.

[30] J. Zhang, L. Wang, D. Niu, and Y. Zhan. Research and implementation of a real time approach to lip detection in video sequences. In *Proc. 2nd Int. Conf. Machine Learning and Cybernetics, Xian*, pages 2795–2799, 2003.