Lip Contour Detection for Estimation of Mouth Opening Area

Shreyansh Jain, Prem C. Pandey, and Rajbabu Velmurugan
Department of Electrical Engineering
Indian Institute of Technology Bombay, Mumbai, 400076, India
Email: shreyansh.jain@iitb.ac.in, {pcpandey, rajbabu}@ee.iitb.ac.in

Abstract—**In speech training aids for providing visual feedback of the articulatory efforts, time-varying vocal tract shape during speech production is generally obtained by linear prediction (LP) analysis of the speech signal and assuming a constant area at the glottis end as a reference. Its variation during speech production causes errors in the estimated vocal tract shape. The problem can be overcome by using area of the mouth opening as the reference. This area can be estimated by detecting the inner lip contour from the video recording of speaker's face during speech utterance. A technique for detection of inner lip contour, based on color transformation and template matching, is presented for reducing the errors caused by presence of teeth and tongue. Face detection by Viola-Jones algorithm, localization using a mouth detection technique, and outer lip contour detection are used to narrow down the search region for inner mouth opening. Presence of the teeth is masked by separate color transformations for upper and lower lip segments. For reducing the errors due to visibility of the tongue, which may not have any significant separation from the lips in the color space, a template matching technique is employed. It is used separately for the upper and lower lip segments to obtain the mouth opening area. The technique has been validated against graphically measured values of the mouth opening and found to be successful in estimating the mouth opening area, and it is not affected by skin hue and presence of teeth.**

Keywords*−*lip contour; mouth opening area; speech training aid; template matching

I. INTRODUCTION

For a person having normal hearing, the learning process for speech production involves receiving auditory feedback. Due to absence of such a mechanism in hearing impaired individuals, alternate means of learning have been developed. One such technique to provide feedback about the articulatory effort with the help of visual cues is to display the dynamically varying vocal tract shape [1], [2]. Most of the methods use inverse filtering of speech signal based on linear predictive coding proposed by Wakita [3]. In this method, the vocal tract is modeled as a lossless acoustic tube with equal-length segments of varying cross-sectional areas. The analysis gives reflection coefficients which are used to obtain the area ratios at the section interfaces. These are scaled by the area of a reference section to obtain the cross-sectional areas. Generally a constant area is assumed at the glottis end and is used as the reference area for scaling. However, the area at the glottis end varies during speech production, as also seen in MRI images [4] for different utterances. Hence the assumption of a constant reference area at the glottis end introduces gross errors during dynamic estimation of the vocal tract shape. One of the proposed alternatives is to use the area of mouth opening, estimated by detecting the inner lip contour from the video recording of speaker's face during speech utterance, as the reference area for scaling purposes. Nayak *et al.* [5] validated the use of mouth opening area as the reference area. The points corresponding to lip opening were manually marked and joined using straight line and the number of pixels within the polygon formed was used as its area. This was repeated for all frames of the video. The area values were normalized by the area obtained for the largest opening, which occurs during the utterance of vowel /a/. It was reported that the scaling of the vocal tract using the area of mouth opening resulted in better estimation of vocal tract shape area compared to the one obtained by using a constant reference area.

For use of the mouth opening area as a reference for estimating dynamically varying vocal tract, an image processing technique for consistent and accurate detection of the inner contour of the lip is required. The technique has to be robust against variations in illumination and the skin hues across speakers. Detection of the inner contour is also severely affected severely by presence of tongue and teeth.

A technique for detection of inner lip contour for estimation of mouth opening area and based on color transformation and template matching is presented. It is validated with reference to manually estimated values. The next section presents a review of lip detection methods. The proposed technique is described in the third section. Test results are presented in the fourth section, followed by conclusion in the last section.

II. LIP DETECTION METHODS

Several methods have been reported for lip detection and segmentation because of its numerous applications [5]-[9]. The main challenges are variation in skin hues, lighting conditions, face orientation, and visibility of the teeth and tongue inside the mouth opening. These methods may be grouped into those that require a lip model and those that do not [7].

The use of deformable template by Hennecke *et al.* [7] is a lip model based approach. A set of parameters is tuned in such a way that the template can be deformed to match the lip

The research is supported by the National Program on Perception Engineering Phase II, sponsored by the Department of Electronics & Information Technology, MCIT, Government of India.

978-1-4673-8564-0/15/$31.00 ©2015 IEEE
Proc. 5th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics 2015 (NCVPRIPG 2015), Patna, India, Dec. 16-19, 2015, paper ID 92.

contour in some optimal fashion, for minimizing a cost function. These deformable templates lack the flexibility to adapt to ‘non-standard’ mouth shapes. Though this approach is insensitive to lighting conditions, the presence of teeth or shadow below lower lip acts as a distractor for accurately determining the contour of the lips. The lack of flexibility of deformable templates can be improved using the active shape model proposed by Luettin *et al.* [8]. It uses *a priori* knowledge about typical deformations, from a training set. This method is constrained by the dataset used for modelling. For good results, a large dataset with samples representing conditions with different illumination, varying skin hues, presence of teeth and tongue etc. need to be labelled and used for training.

One of the approaches to permit higher flexibility in tracking lip contour is to use active contours. Unlike the previous two methods, it is not constrained by a dataset used for modelling. It requires the contour to be manually initialized in the first frame of the video. The contour locks down on the visual features of the lip. The motion of these features causes the resultant motion of the contour. The tracking is done irrespective of any deformation in the lips. This method is sensitive to lighting conditions, presence of teeth etc. which interfere with the tracking. Further, flexible contours can lead to generation of unrealistic shapes.

Nayak *et al.* [5] investigated a template matching based technique to estimate the area of mouth opening modelled as an ellipse. In this method, horizontal strip matching is done to determine the major axis, followed by strip length correction and asymmetric ellipse matching. Presence of teeth and tongue may result in lower estimate of the height of mouth opening, especially in cases where upper and lower teeth are in proximity of each other, like in /i/. Further, the technique is computationally expensive due to a large number of iterations involved in matching the template model in an unrestricted search region in mouth sub-image.

Saha and Pandey [9] used a technique based on multi-threshold binarization and connected component detection to reduce the error in mouth area estimation due to presence of teeth and tongue. It was found that the leftmost and rightmost point of the horizontal opening, estimated using this approach, were sensitive to the orientation of mouth as well as lighting condition. Also in some frames, corresponding to utterance of /a/ and /i/, actual area values were higher compared to the estimated ones. In some frames, especially the ones having small vertical opening, the height of opening estimated was higher than the actual value.

It can be seen that these existing techniques are sensitive to lighting condition changes and thereby do not provide a good estimate of the mouth opening area. They are also affected by the presence of teeth, tongue and shadow below lower lip. This motivates the proposed approach, discussed in the next section, which addresses these issues.

III. Proposed Technique for Mouth Area Estimation

In this approach, the elliptical template proposed by Nayak *et al.* [5] is used to estimate the area of mouth opening. The first step in the proposed technique is to determine the region of interest by localizing the search region. This is followed by outer lip contour detection to detect the horizontal opening corresponding to the major axis of the elliptical model of mouth opening. Detection of outer lip contour constrains the search region for the subsequent detection of inner lip vertical opening. The upper and lower lips are processed separately by applying a suitable color transformation for contrast enhancement and then employing template matching to model the mouth opening as an unsymmetrical ellipse.

A. Localization of the search region

The image in a given frame is processed to extract the region of interest, as shown in Fig. 1, on which detection techniques are applied. This is carried out in two steps: (i) Face detection and (ii) Detection of mouth sub-image.

The first step consists of Viola-Jones method, which is commonly used for robust face detection in real-time applications [10]. The input RGB image is converted into grayscale by averaging the three channels.

$$S_G = (R + G + B)/3 \qquad (1)$$

The grayscale image is transformed into an integral image by replacing each pixel by the sum of all the pixels to its above and left.

$$I_{SG} = \sum_{x' \leq x, y' \leq y} S_G(x', y') \qquad (2)$$

The integral image is evaluated in one pass over the grayscale image and it permits fast computation of features at different scales and locations to build a classifier. Viola-Jones method relies on the fact that the eyes are darker than the cheeks and the bridge region of the nose. In some images, the algorithm may identify small non-face parts of the image as face image as well. This problem is solved by using the largest area sub-image as the face image..

In the second step, the face image as extracted in the first step is processed to detect the mouth sub-image using the method by Hsu *et al.* [11]. For this purpose, the RGB image is converted into the color space YC_bC_r. The variation of chrominance components across the face is such that the eye region is characterized by high C_b and low C_r whereas the mouth region has high C_r. This variance is used to detect the mouth region by calculating the mouthmap M_m using the following transformation:

$$M_m = C_r(x,y)^2(C_r(x,y)^2 - 0.95(\frac{\sum_{(x,y)} C_r(x,y)^2}{\sum \frac{C_r(x,y)}{C_b(x,y)}})\frac{C_r(x,y)}{C_b(x,y)})^2 \qquad (3)$$

To extract the mouth region, mouthmap is further processed using various morphological operation of erosion followed by dilation. After the processing, it is observed that the region identified has high value in the lower lip. To extract the mouth sub-image from this location, the identified region is extended by 1.25 times its height on bottom and 1.25 times its width on both sides. The extension at the top side is 2.5 times the height so as to cover the upper lip in the mouth sub-image as well. In some cases, this extension on upper side may include the nostrils in the image. To avoid it, average of mouth-sub image

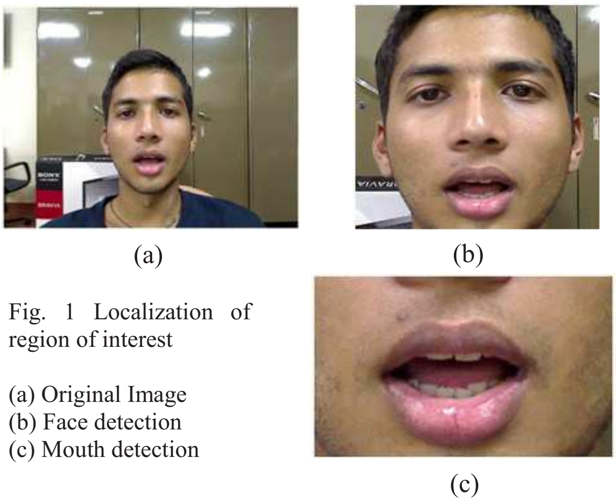

Fig. 1 Localization of region of interest

(a) Original Image
(b) Face detection
(c) Mouth detection

heights across the frames of video clipping is calculated. If the detected height in a frame is greater than 1.5 times the average height, the extension factor of 2 rather than 2.5 is used.

B. Horizontal and vertical opening estimation for outer lip

For estimating the horizontal opening, it is assumed that the horizontal orientation of the image is aligned with the lip opening. The opening estimated in this step is fixed as the horizontal opening for inner lip contour too. After estimating the horizontal opening, its center is determined to estimate a rough vertical opening for the outer lip contour. The vertical opening from the outer lip contour is used to place a bound on the inner lip contour.

A modified version of the technique proposed by Saha and Pandey [9] is used to estimate the horizontal opening for outer lip contour. The RGB image is converted into grayscale using (1). Sum of the pixel values are calculated across the rows. As most of the pixels inside mouth opening have very low intensities, the row with the lowest sum is taken as the one corresponding to the maximum horizontal opening of the mouth. It is marked as y_{center}. The left and right end points are determined by thresholding the image using a threshold set at 25% of the 99 percentile of the pixel value to obtain a binary image. The image is further processed by erosion and dilation morphological operations to remove any unconnected white patches which might have been left behind after thresholding. In the resulting binary image, the row corresponding to maximum horizontal opening along with its neighbouring rows are scanned and the leftmost and rightmost white pixel positions are taken as the markers of the lip opening. These points are used to find the center and width of the mouth opening. Vertical opening for the outer lip contour is estimated along the centre of horizontal opening. For this purpose, the image is converted from RGB to YIQ color space

$$Q = 0.211R - 0.522G + 0.311 \qquad (4)$$

Lip pixels have higher value in the Q channel whereas the face pixels have higher values in the I channel. Use of Q channel results in minimizing the errors which might be introduced in the estimation of the bottom end of vertical lip opening due to presence of shadow below the lower lip. A threshold of zero is used to binarize the Q channel (which may have positive and negative values). This is followed by performing morphological opening operation of erosion and dilation on the binary image using a diamond shaped structuring element. The binary image is then further transformed by removing small unconnected white patches. The top and bottom points of vertical opening are searched along the vertical line corresponding to centre of maximal horizontal opening. The topmost white pixel location is taken as the position of upper boundary of outer lip contour and is denoted as y_{top}. Similarly the bottommost white pixel corresponds to the lower boundary of the outer lip contour and is denoted as y_{bottom}. These locations are then used to constrain the search region corresponding to the estimation of inner lip contour as shown in Fig. 2(a).

C. Estimation of lip opening for inner lip contour

For estimating the vertical opening corresponding to the inner lip contour, the horizontal segment estimated earlier is used to segment the mouth region to process the segments separately. The lip opening is generally unsymmetrical about the horizontal line used to segment the mouth opening. It is modeled as a region bounded by two semi-ellipses sharing the horizontal opening as the major axis..

The search region for upper segment of the mouth opening is constrained between y_{top} and y_{center} and that for the lower segment between y_{center} and y_{bottom}. For determining the inner boundary, the image for the lower segment is converted into grayscale using the transformation proposed by Nayak *et al.* [5]:

$$Z_{lower} = R - G/2 - B/4 \qquad (5)$$

For the image of the upper segment, the following transformation was found to be more suitable:

$$Z_{upper}= R - (5/2)G + (5/2)B \qquad (6)$$

These transformations were found to increase the contrast between the lips and other part of the mouth across the images from several speakers and varying lighting conditions and were successful in darkening the area surrounding the lip, and masked the teeth in most of the cases.

As shown in Fig. 2, the segments are modeled as white semi-ellipses and template matching is employed to search between the outer lip contour boundary and the central horizontal opening to determine the corresponding minor axis length at which the minima of correlation is achieved. At the point of correlation minima, inner mouth extremities are localized as localized as $y_{inner\text{-}top}$ and $y_{inner\text{-}bottom}$. These points along with the left and right mouth ends of the horizontal segment are used to estimate the area of mouth opening as the sum of the area of the two semi ellipses.

IV. RESULTS

The technique presented in the previous section was used to estimate the mouth opening area from recordings from three speakers, for the vowels /a/, /i/, and /u/. For reference, the mouth opening areas in these frames were graphically calculated [5]. It was observed that the horizontal bar length

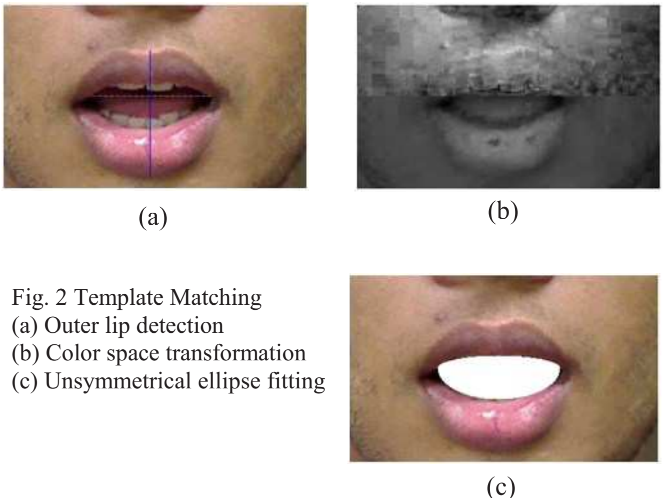

Fig. 2 Template Matching
(a) Outer lip detection
(b) Color space transformation
(c) Unsymmetrical ellipse fitting

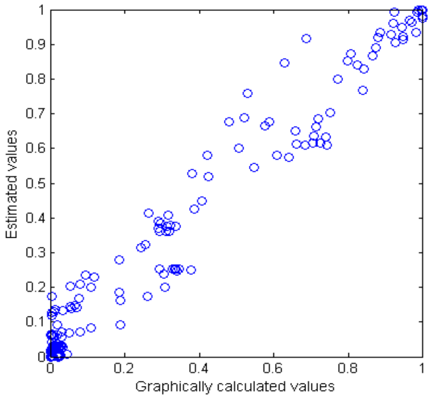

Fig. 3 Scatter plot of area of mouth opening estimated using template matching versus graphically calculated values

during utterance of /u/ was overestimated due to the relatively small mouth opening. Interference due to presence of tongue might introduce error in estimated vertical opening in some cases, especially in the scenario where there is low contrast between the tongue and the lower lip.

The estimated and graphically calculated areas were normalized to the range 0−1, using the following relation:

$$A(i) = \frac{A(i) - \min_i(A(i))}{\max_i(A(i)) - \min_i(A(i))} \qquad (7)$$

where $A(i)$ denotes the mouth opening area for the frame i. A scatter plot of the two set of the normalized values is shown in Fig. 3. A high density of data is observed to lie in very low area region due to intervals of closed lips across all the video sequences. A correlation coefficient of 0.97 is obtained between the estimated values and graphically estimated values.

V. Conclusion

A technique for estimation of mouth area has been presented and investigated. The estimated area values are to be used for scaling purposes in LPC analysis for modelling dynamic vocal tract shape. An extensive testing is needed to evaluate the technique for its suitability for video recordings under varying illumination and from persons with speech impairment.

References

[1] R. G. Crichton and F. Fallside, “Linear prediction model of speech production with applications to deaf speech training, ” Proc. IEE, vol. 121, no. 8, pp. 865-873, 1974.

[2] S. H. Park, D. J. Kim, J. H. Lee, and T. S. Yoon, “Integrated speech training system for hearing impaired”, IEEE Trans. Rehabilitation Engineering, vol. 2, no. 4, pp. 189-196, 1994.

[3] H. Wakita, “Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms,” IEEE Trans. Audio Electroacoustics, vol. AU-21, no. 5, pp. 417–427, 1973.

[4] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammart, J. Kim, S. Lee, K. S. Nayak, Y. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. K. Ghosh, A. Katsamanis, and M. Proctor, “Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research, ” J. Acoust. Soc. Am., vol. 136, pp. 1307-1311, 2014.

[5] N. S. Nayak, R. Velmurugan, P. C. Pandey, and S. Saha, “Estimation of lip opening forscaling of vocal tract area function for speech training aids, ” in Proc.18th National Conf. Commun., Kharagpur, 2012, pp. 521-525.

[6] P. Gacon, P. Y. Coulon, and G. Bailly, “Non-linear active model for mouth inner and outer contours detection,” in Proc. 13th European Signal Processing Conference, Antalya, Turkey, 2005, pp. 1-4.

[7] M. Hennecke, K. Prasad, and D. Stork, “Using deformable templates to infer visual speech dynamics,” in Proc. 28th Asilomar Conference on Signals, Systems and Computer, vol.1, Pacific Grove, California, 1994, pp.578-582.

[8] J. Luettin, N. A. Thacker, and, S. W. Beet, “Active shape models for visual speech feature extraction, ” in Speechreading by Humans and Machines, Berlin: Springer, 1996, pp. 383-390.

[9] S. Saha and P. C. Pandey, “Estimation of the area of mouth opening during speech production,” in Proc. 8th Indian Conference on Vision, Graphics and Image Processing, Mumbai, India, 2012, no. 315

[10] P. Viola and M. Jones, “Robust real-time face detection,” Int. J. Comput.Vis., vol. 57, no.2, pp. 137–154, 2004.

[11] R. Hsu, M. Abdel-Mottaleb, and A. K. Jain, “Face detection in color images,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 5, pp. 696–706, 2002.