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
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 using the method by Hsu et al. [8]. The color space YCbCr is used in this method, with the assumption that Cc is greater than Cb in the mouth region. The equations for conversion from RGB to YCbCr are as follows:

\[ Y = 0.299R + 0.587G + 0.114B \]
\[ C_b = -0.169R - 0.331G + 0.500B + 128 \]
\[ C_c = 0.500R - 0.418G - 0.081B + 128 \]

From these values, \( C_c^2 \) and \( C_c/C_b \) are normalized to the range \([0, 255]\), the same as that of \( R, G, B \). These normalized values of pixels in the face sub-image \( F \) are transformed according to the equation as given in [8] to calculate “mouthmap” as

\[ \text{mouthmap} = C_c^2 (C_c^2 - \frac{C_c}{C_b})^2 \]

where \( \eta \) is given by

\[ \eta = 0.95 \frac{\sum_{(x,y) \in F} C_c (x,y)^2 \frac{C_c}{C_b(x,y)}}{\sum_{(x,y) \in F} C_c(x,y)} \]  

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. The resulting binary image, the location of the lower lip is taken as the leftmost point of the horizontal opening.

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 binary image is denoted as \( T(x, y) \). In this binary image, the \( y_{bottom} \) row is scanned, and the next row is taken as the row corresponding to the maximal opening of the mouth. 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. 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

\[ P = (R + G + B)/3 \]  

Sum of the pixel values in each row is calculated as

\[ S(y) = \sum_{x=1}^{nx} P(x, y) \]

where \( nx \) 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, \( y_{bottom} \), is taken as the row corresponding to the maximal opening of the mouth.

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 \( \log(G/B) \). Vezhnets [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
Figure 1: Obtaining face sub-image and mouth sub-image from input frame.

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

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

Figure 6: Estimated vertical opening

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.

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


