Estimation of Lip Opening for Scaling of Vocal Tract Area Function for Speech Training Aids

Nagesh S. Nayak, Rajbabu Velmurugan, Prem C. Pandey, Sudipan Saha
Department of Electrical Engineering
Indian Institute of Technology Bombay, Mumbai, 400076, India
Email: nageshsnayak@sensibol.com, {rajbabu, pcpandey, sudipan}@ee.iitb.ac.in

***Abstract*—For visual feedback of articulatory efforts in speech training aids, the vocal tract shape can be estimated by LPC analysis of the speech signal. The vocal tract is modelled as a concatenation of equal length sections and the ratios of the areas at section interfaces are calculated and these are scaled using the area of a reference section. The lip opening area as estimated from a video recording of the speaker's face can be used as a reference area for obtaining the vocal tract shape during speech utterances with transitional tract configuration. A technique for estimating the area of the lip opening based on template matching is investigated. It satisfactorily tracked the horizontal and vertical opening of the lips in the video images of speakers with different skin hues, recorded under good lighting conditions.**

***Keywords*—vocal tract shape; speech training aid; lip opening; template matching.**

I. Introduction

Several speech training aids have been developed for providing a visual feedback of articulatory efforts to the hearing impaired children [1]–[3]. Most of these aids use Wakita's inverse filtering method [4] for estimation of the vocal tract shape directly from the speech signal. The vocal tract is modelled as a lossless acoustic tube with a series of discrete and equal length sections with variable cross-section area. The LPC analysis is used for calculating the reflection coefficients at the section interfaces from the autocorrelation coefficients of the windowed speech signal and the area of a reference section can be used for scaling the ratios to get the vocal tract area function. One of the main limitations of the technique, as outlined by Wakita [5], is the lack of a method for estimating the area of a reference section from the speech signal for converting the area ratios into area values during varying configuration of the vocal tract. Generally, the scaling is carried out by assuming a normalized area of unity at the glottis end, but this method introduces errors due to variation in the area at the glottis end. Estimation of the vocal tract shape for a dynamic display can be improved by using the area of opening at the lips, estimated from the video images of the speaker's face, as the reference for scaling. With a simultaneous acquisition of the audio and video signals, the lip area values estimated from the video frames can be used for interpolating the values corresponding to the LPC analysis frames.

The main problems in lip tracking arise due to different skin hues and the variations in lighting conditions [6]. Other difficulties are caused by different physical shapes and orientations of the speaker's face, position of the face from the camera, and visibility of the teeth and tongue. Yuille *et al.* [7] used deformable templates to find the best fit for the lips, but it requires *a priori* knowledge of the shape. Wang *et al.* [6] used active shape models, for lip tracking, which are trained using a large database. Kass *et al.* [8] and Chen *et al.* [9] used active contours/snakes for tracking the outer lips, but initializing the active contours to track the inner lips is difficult.

The objective of our investigation is to validate the use of the area of the opening at the lips for scaling the vocal tract area function and to devise a technique based on template matching for tracking the lip opening.

II. Scaling of the Vocal Tract Area Function

The investigation was carried out to compare the vocal tract shapes, for steady state vowel utterances, obtained by direct imaging with those obtained by estimation of the vocal tract area function from the speech signal and scaled by (i) unity area at the glottis and (ii) lip opening area. Wakita's method [4] was used for estimating the area ratios, with the use of minimum windowed energy index for improving the consistency [10]. Audio-visual signals were recorded from six male speakers under different lighting conditions, using a 2 megapixels webcam at 15 frames/s. The lip opening areas were calculated manually. In each frame, the points corresponding to the lip opening were marked using a cursor on the screen, and the number of pixels within the polygon formed by joining the marked points was saved as the lip opening area for the frame. The area values were normalized with respect to the area during the vowel /a/ (generally the largest opening). A cubic-spline interpolation was applied on these values to get the values at the frame positions used in the LPC analysis of the speech signal [11]. The area values of the vocal tract were displayed as an areagram, a 2-D plot of the square-root of the area of discrete sections of the vocal tract as a function of time along the x-axis and distance from the glottis on the y-axis (0: glottis, 1: lips). The areas are indicated as grayscale values, with white indicating the maximum opening and black indicating a closure.

An example of the results is shown in Fig. 1 for the vowel sequence /-a-i-u-/. It shows the speech waveform, wideband spectrogram, plot of the manually estimated lip opening, glottis-scaled areagram, and lip-scaled areagram. The

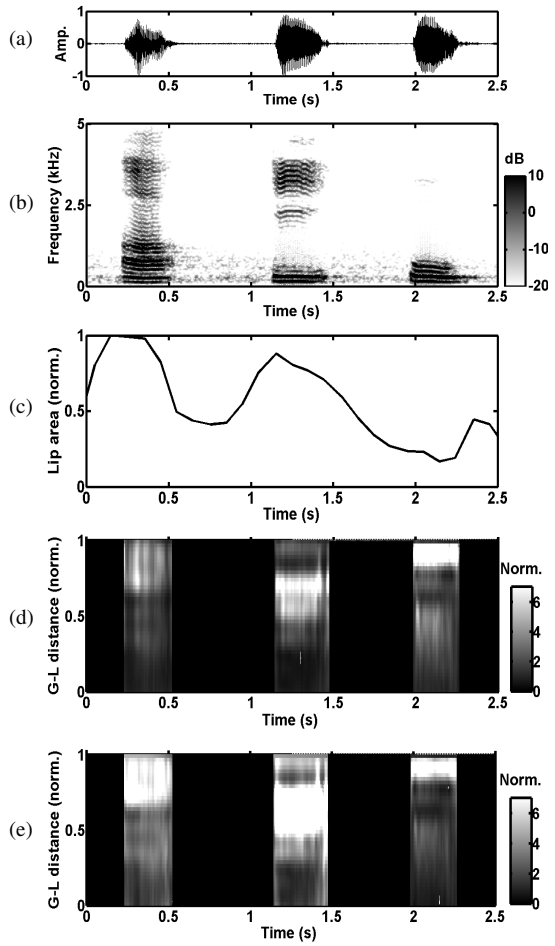


Figure 1. Areagrams for vowel sequence /-a-i-u-/ with glottis-scaling and with lip-scaling for speaker A: (a) speech signal, (b) wideband spectrogram, (c) lip opening area (norm.), (d) glottis-scaled areagram, and (e) lip-scaled areagram.

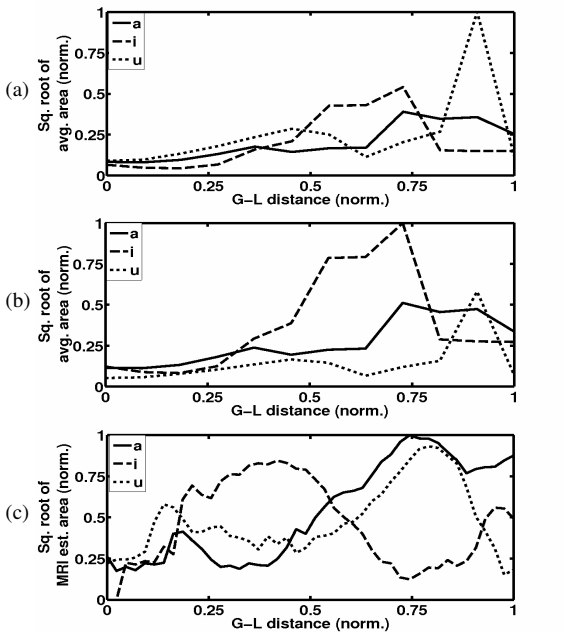


Figure 2. Vocal tract for the vowels /a/, /i/, and /u/ (a) estimated from the speech signal of speaker A after glottis-scaling, (b) estimated from the speech signal of speaker A after lip-scaling, and (c) estimated from MRI based on data reported in [12].

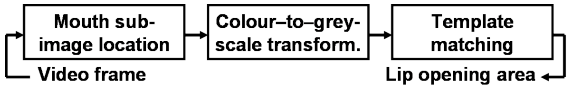


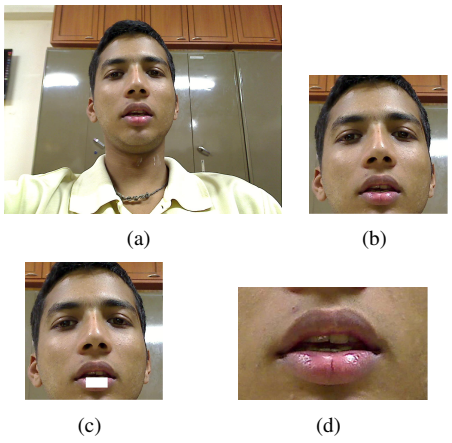
Figure 3. Block diagram of template matching.

glottis-scaled areagram shows the three different vocal tract configurations for the three vowels, but the difference are more distinctly visible in the lip-scaled areagram. The vowel /a/ shows a wide opening at the lip end of the vocal tract and a significant improvement in the area values. For /i/, the area values show a consistent vocal tract shape as well as place of constriction. For the vowel /u/, the lip rounding as well as the large area behind the lips are visible. Similar patterns were observed for the other five speakers.

To validate whether the area at the glottis remains the same during different steady state segments, square-root of the area values corresponding to the centre region of each vowel in the /-a-i-u-/ utterance were averaged and these values were compared with the values obtained from one of the direct imaging methods. The plots of the values after glottis-scaling and lip-scaling for speaker A are shown in Fig. 2(a) and 2(b), respectively. The square root of the area values obtained using MRI by Story *et al.* [12] for the vowels /a/, /i/, and /u/ are shown in Fig. 2(c). For the vowel /a/, a large area is observed at the front of the vocal tract. The area values for vowel /i/ show a constriction at the front with a large area at the back of the vocal tract. For the vowel /u/, the area at the lip is small followed by a large area immediately behind the lip. The lip area values are the largest for the vowel /a/, followed by the vowel /i/, and the vowel /u/ which has the lowest area due to lip rounding. It was observed that the area values at the lips obtained for all the six speakers after scaling using the lip area are consistent with those obtained using MRI. These resulted in different area values at the glottis end, across the speakers and utterances, indicating the invalidity of assuming a fixed area at the glottis. Since scaling of vocal tract using the area of opening at the lips results in a better vocal tract shape when compared with direct imaging methods, the next step is to automate the lip opening area estimation.

III. Tracking of the Lip Opening by Template Matching

A template matching technique is investigated for tracking the lip opening in the video frames. The block diagram of the method is shown in Fig. 3. A sub-image containing the mouth is located to restrict the search area for template matching. The colour mouth sub-image is transformed to a grayscale image to increase the contrast between the lip opening and the surroundings, and template matching is performed on the transformed image for tracking the lip opening.



(a) (b) (c) (d)

Figure 4. Results from various stages in a video frame during the vowel /i/: (a) Original frame, (b) Face detection using Viola-Jones method, (c) Mouth location by Hsu *et al.*, with the located region marked in white, and (d) Detected mouth sub-image.

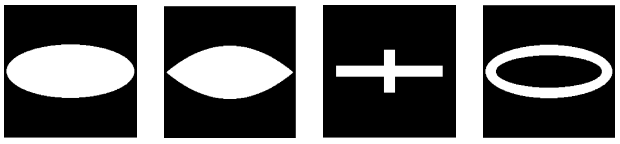
(a) (b)

Figure 5. Image transformation in a video frame during vowel /i/: (a) mouth sub-image and (b) transformed mouth sub-image.

A. Mouth sub-image location and image transformation

Tracking of the lip opening by template matching on the facial image often resulted in errors. Therefore, the search area is restricted to a sub-image around the mouth, using Viola–Jones' face detection method [13] and the method by Hsu *et al.* to locate the mouth region [14]. For most speakers, the mouth region located was on the lower lip, as shown marked in white in Fig. 4(c). To choose the mouth sub-image, the identified region is extended 1.25 times its length on both sides, and 1.25 times its height on the bottom. Since the initially identified region did not include the upper lip, the extension on the top is 2.5 times the height. This extended region forms the mouth sub-image, as shown in Fig. 4(d).

The mouth sub-image contains the lips, skin, part of the upper and lower teeth, and tongue depending on the type of utterance. Across the video images of several speakers and lighting conditions, use of $X = R - 0.5G - 0.25B$ as the colour-to-grayscale transformation was found to increase the contrast between the lips and other regions in the mouth sub-image, and hence this function is used as the colour-to-grayscale transformation. The transformed image for a mouth sub-image in Fig. 5(a) is shown in Fig. 5(b). In almost all the cases, the transformation resulted in the opening between the lips getting darker than the surrounding area and the teeth getting almost masked.



(a) (b) (c) (d)

Figure 6. Different templates : (a) white ellipse, (b) white area bounded by two parabolas, (c) white cross, and (d) white elliptical ring.

B. Template matching

The lip opening can be modelled as a dark ellipse with its major and minor axes corresponding to the width and height of the lip opening, respectively. The types of templates investigated for tracking the lip opening included a white ellipse, white area bounded by two parabolas, white cross with horizontal and vertical strips, and white elliptical ring as shown in Fig. 6. Considering the computational complexity involved in using these templates and the results obtained, the white ellipse was found to be the most appropriate template. Correlation of the white elliptical template with the grayscale image is calculated as a function of the position of the center of the ellipse and lengths of the major and minor axes. The lip opening is identified by finding the values of these variables corresponding to the minimum value of the correlation. Restrictions are imposed on the range of the variables for reducing the computation and decreasing the possibility of errors. The maximum length of the minor axis equals the vertical opening of the lips during vowel /a/. The maximum length of the major axis equals the horizontal opening of the lips during the vowel /i/.

Sudden changes in the ellipse minor axis between successive frames are occasionally observed due to incorrect positioning of the major axis. To avoid these errors, a horizontal white strip is used as a template to locate the region between the two horizontal lip ends. A white rectangle having variable length and width equal to 10% of the maximum value of major axis is used as the template. The left and right edges of the lip-skin boundary are located by finding the position and length of the rectangle corresponding to the minimum of the cross-correlation over the grayscale mouth sub-image. For closed lips, the rectangle overlaps with the dark region where the upper and lower lips meet. An example of fitting of the rectangular template is shown in Fig. 7(b). Its length is used as the major axis of the white ellipse and its minor axis is varied for template matching.

To reduce the number of iterations, the intial estimate of the minor axis is found by using another white rectangle, perpendicular to and centred on the horizontal rectangle as shown in Fig. 7(c). It may be noted that the presence of teeth may cause error in estimating the height of the opening, particularly when the upper and lower teeth are close to each other, e.g. in /i/. An example of fitting of the white ellipse template is shown in Fig. 7(d). The pre-localization of the major axis significantly reduces the computation and prevents the template from being matched in the skin region, which

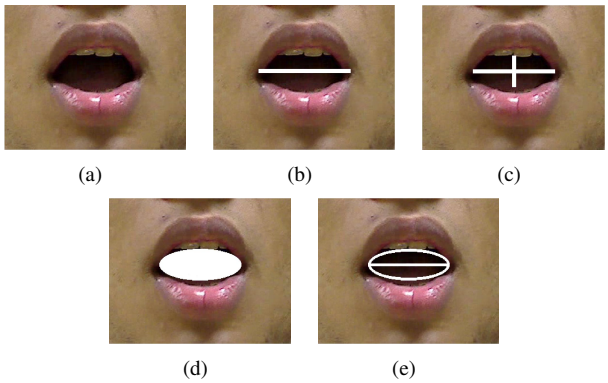


Figure 7. Template matching steps for a video frame during vowel /a/ : (a) mouth image, (b) horizontal strip fitting, (c) vertical strip fitting, (d) white ellipse template fitting using horizontal strip as major axis, and (e) white ellipse template fitting shown as an elliptical ring.

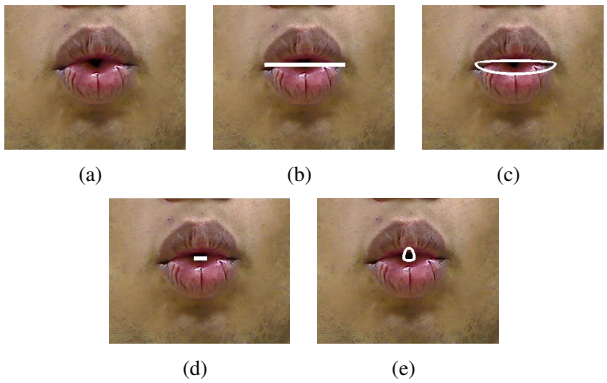


Figure 9. Lip opening estimation for /u/ : (a) original image, (b) horizontal strip template fitting, (c) filled ellipse fitting (incorrect), (d) horizontal strip template fitting (corrected), and (e) filled ellipse template fitting using corrected horizontal strip as major axis.

Figure 8. Comparison of the (a) symmetric and (b) asymmetric ellipse template matching for a video frame during vowel /a/.

happens otherwise due to variations in lighting and skin colour. Fig. 7(e) shows the filled ellipse as an elliptical ring with its major axis position along the horizontal strip.

The lip opening is generally not symmetric on the two sides of the horizontal strip leading to small errors in the ellipse minor axis estimation. This error is reduced by using two semi ellipses with the same major axis. An example of the lip opening estimation using the symmetric and asymmetric elliptical templates is shown in Fig. 8.

The mouth opening is sometimes much smaller than the length as estimated by the horizontal strip matching. An example for the vowel /u/ is shown in Fig. 9. This results in a large error in the ellipse fitting as it uses the major axis as estimated by the horizontal strip matching, as seen in Fig. 9(c). As a solution to this problem, the position of the strip centre is kept fixed as determined by the first localization step while its length is varied and the value resulting in the minimum correlation with the transformed image is taken as the corrected major axis of the ellipse. The result for the ellipse fitting using corrected major axis is shown in Fig. 9(e). Use of the grayscale image for obtaining the position of the horizontal strip avoids the possibility of the horizontal strip tracking the dark region lying in the shadow below the lower lip or the nostrils.

On the basis of these investigations, the steps for template matching are (i) horizontal strip matching, (ii) strip length correction, and (iii) asymmetric ellipse matching.

IV. Results and Discussion

The estimation of the lip opening using the combination of the horizontal strip and filled ellipse template was carried out for the video recordings from six speakers. A scatter plot of the estimated lip area values using the template matching method versus the graphically calculated values for the vowels /a/, /i/, and /u/ for the recordings from six speakers used in our investigation is shown in Fig. 10. For each speaker, the values are normalized with respect to the area for /a/. The normalized value of the graphically calculated areas for /a/ for all the speakers is 1. For /i/, the normalized area values ranged from 0.5 to 0.8. For /u/, the values were approximately 0.1 due to a low area value caused by lip rounding. The other area values corresponded to the transition regions. The correlation coefficient between the two variables was 0.77.

From the scatter plot, it is observed that the maximum error occurred for the cluster corresponding to the vowel /u/. For most cases of the vowel /i/ for all the speakers, the estimated area values were larger than the reference values, but the error was less as compared to the vowel /u/. For the vowel /a/, the estimated area was either on the higher or the lower side of the actual area. The larger estimate may be due to the elliptical template which may overcompensate for the lip opening. The lower estimates resulted due to the presence of teeth or tongue in the mouth sub-image because the colour-to-grayscale transformation with fixed weights was not able to increase the lip contrast equally well for speakers with different skin tones.

V. Conclusion

Use of lip opening area for scaling the area ratios, obtained from the LPC-based inverse filtering of the speech signal, to get the vocal tract area function has been validated. The scaling improved the estimation of the area values at the places of maximum opening without significantly affecting the places of constriction. A technique for tracking the opening of the lips by using a white asymmetric ellipse as template has been investigated. Use of a mouth sub-image within the face

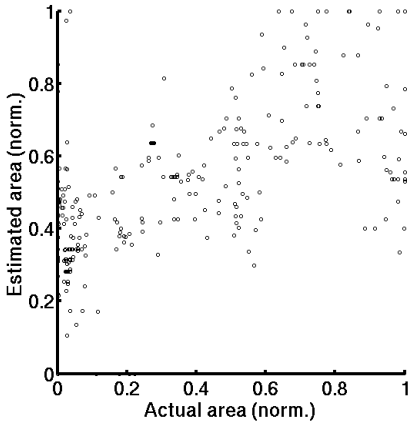


Figure 10. Scatter plot of the lip area values estimated using the template matching method with the actual area values measured graphically. Correlation coefficient = 0.77.

detected in each frame of the video helped in reducing the errors. The technique satisfactorily tracked the lip opening in the video images, recorded under good lighting conditions.

A quantification of the tracking ability of the technique for a larger number of speakers with utterances having different lip shapes has to be carried out. Further refinements in the technique are needed to reduce the errors in estimating the lip opening for some of the frames. Most of these errors may be reduced by adapting the colour-to-grayscale transformation to the speaker's skin tone. Use of active contours to track the outer lip and the template matching inside the area determined by the active contour may also be explored. Further investigations are needed for decreasing the computational complexity. The estimated lip area values have to be used for scaling the area ratios estimated by LPC analysis for dynamic estimation of the vocal tract area functions during speech utterances with transitional vocal tract configurations.

ACKNOWLEDGMENT

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

REFERENCES

[1] H. Levitt, J. M. Pickett, and R. A. Houde, (Eds.) "Speech training aids," part VII in *Sensory Aids for the Hearing Impaired*, New York: IEEE Press, 1980, pp. 349–419.

[2] A. Watanabe, S. Tomishige, and M. Nakatake, "Speech visualization by integrating features for the hearing impaired," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 454–466, 2000.

[3] S. H. Park, D. J. Kim, J. H. Lee, and T. S. Yoon, "Integrated speech training system for hearing impaired," *IEEE Trans. Rehabil. Eng.*, vol. 2, no. 4, pp. 189–196, 1994.

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

[5] 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.*, vol. ASSP-27, no. 3, pp. 281–285, 1979.

[6] S. L. Wang, W. H. Lau, and S.H. Leung, "Automatic lip contour extraction from color images," *Pattern Recogn.*, vol. 37, no. 12, pp. 2375–2387, 2004.

[7] A. L. Yuille, P. W. Hallinan, and D. S. Cohen, "Feature extraction from faces using deformable templates," *Int. J. Comput. Vis.*, vol. 8, no. 2, pp. 99–111, 1992.

[8] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, 1988.

[9] J. Chen, B. Tiddeman, and G. Zhao, "Real-time lip contour extraction and tracking using an improved active contour model," in *Proc. 4th Int. Symp. Adv. Vis. Comput. (ISVC '08)*, 2008, part II, pp. 236–245.

[10] K. S. Nataraj, Jagbandhu, P. C. Pandey, and M. S. Shah, "Improving the consistency of vocal tract shape estimation," in *Proc. National Conf. Commun. (NCC)*, 2011, Bangalore, India, paper SpPrII.4.

[11] N. S. Nayak, "Estimation and display of vocal tract shape for speech training aids," M.Tech. thesis, Electrical Engineering, IIT Bombay, India, 2011.

[12] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *J. Acoust. Soc. Am.*, vol. 100, no. 1, pp. 537–554, 1996.

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

[14] 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.