Place of Articulation from Direct Imaging for Validation of Its Estimation from Speech Analysis for Use in Speech Training

K. S. Nataraj and Prem C. Pandey
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai 400076, India
Email: {natarajks, pcpandey} @ ee.iitb.ac.in

***Abstract*—Place of articulation obtained by analysis of the speech signal is useful for visual feedback of articulatory efforts for speech training of hearing impaired children and for improving pronunciation by learners of second languages. Its estimation by direct imaging of the oral cavity is needed for validating the estimation from the speech signal. For such applications, an automated technique is presented for estimating the place of articulation by graphical processing of the upper and lower contours of the oral cavity image. It iteratively estimates the axial curve as an axis of symmetry of the oral cavity, such that the curve approximately bisects the normals to it. Distance between the contours along the normal to the axial curve gives the oral cavity opening and position of the smallest opening provides the place of articulation. The values estimated using the automated technique closely matched those obtained by manual marking of the visually estimated place of maximum constriction for the oral cavity images of vowels, stops, and fricatives, from the XRMB and MRI databases.**

Keywords—axial curve; oral cavity opening; place of articulation; speech training

I. INTRODUCTION

Hearing-impaired children have difficulty in acquiring ability to control the articulators involved in speech production, due to lack of auditory feedback. Speech training aids can provide non-auditory feedback to help the process of speech acquisition. A visual feedback of articulatory efforts has been found to be useful in improving vowel articulation by the hearing-impaired children [1], [2]. It can also help a second-language learner in improving pronunciation [3]. Place of articulation, i.e., the place of maximum constriction in the oral cavity is the most significant information for speech training. It can be estimated using imaging techniques, acoustic measurements, and analysis of the speech signal. However, only the estimation obtained by analysis of the speech signal can be used for speech training. In the commonly used methods, analysis of the speech signal is based on linear predictive coding with the oral cavity modeled as a lossless acoustic tube with plane wave propagation [2], [4], [5]. Although the lower and upper contours of the oral cavity have varying curvatures, the cavity is modeled as a tube with equal-length segments of varying cross-section areas. Place of articulation estimated by speech analysis needs to be validated with reference to the value obtained from direct imaging of the oral cavity during speech production.

Direct imaging provides upper and lower contours of the oral cavity. Irregular shapes of the contours cause difficulties in locating the maximum constriction and finding its distance from the lips in a consistent manner. As a solution to this problem, an automated technique involving graphical processing of the contours is presented. The acoustic wave propagation is assumed to be along an axial curve between the two contours with the normal to the curve representing the wave front. The axial curve is iteratively estimated as an axis of symmetry, approximately bisecting the normals to it. Segment of the normal to the axial curve between the two contours provides an estimate of the oral cavity opening. Values of oral cavity opening as a function of distance from the lips are used for estimating the place of articulation.

The second section provides a review of the direct imaging techniques and some of the earlier methods for estimation of place of articulation. The proposed automated technique is presented in the third section. The test results are presented in the fourth section, followed by conclusion in the last section.

II. PLACE OF ARTICULATION BY DIRECT IMAGING

Commonly used direct imaging techniques to capture the movement of the articulators in the mid-sagittal plane (side view) during speech production are ultrasound imaging, electropalatography (EPG), X-ray microbeam (XRMB), electromagnetic articulography (EMA), and magnetic resonance imaging (MRI) [6]–[11]. Speech production databases have been developed using the last three techniques.

The XRMB technique [6] uses narrow X-ray beams for recording articulatory movements in the mid-sagittal plane by tracking the gold pellets glued to the articulators. It permits simultaneous recording of speech signal. The database [7], developed at the University of Wisconsin, provides articulatory plots consisting of four pellet points (T1-T4) on the tongue and one each on the upper lip (UL), lower lip (LL) and incisor (MNi), at 160 frames/s and audio recordings for vowels, vowel-consonant-vowel syllables, sentences, and paragraphs, from 25 male and 22 female speakers.

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 88.

The EMA technique [8] tracks the movement of tongue, jaw, lips, and teeth by measuring the alternating voltages in the receiver coils placed on them due to the magnetic fields generated by transmitter coils on a plastic helmet. The Multi-channel Articulatory (MOCHA) database [12], developed at the Edinburgh Speech Production Facility, provides articulatory data at 500 frames/s and audio recordings for 460 sentences, from one male and one female speaker.

The MRI technique can be used to capture images of the oral cavity [10] without radiation hazard. It permits recording of the speech signal, but the signal gets corrupted by a high intensity gradient noise. Bresch et. al. [13] used a setup with FPGA-based hardware to synchronize the audio and image acquisition and two optical microphones to capture the speech signal and the ambient noise. The noise in the speech signal was suppressed by offline processing using the ambient noise as a reference. The database [14], developed at the University of Southern California, has MRI data and noise-suppressed audio recordings for 460 sentences (same as in the MOCHA database) from five male and five female speakers, and EMA data from two of the male and two of the female speakers.

The XRMB and EMA techniques provide high frame rate and noise-free audio recordings. However, the instrumentation provides the oral cavity contour information only at discrete points and accurate place of the tongue tip is unknown as the tongue tip pellet or receiver coil is placed about one cm away from the actual tip. The XRMB technique poses a radiation risk. The MRI technique provides oral cavity contour information along the entire length of the oral cavity, but the frame rate is low and the audio recordings are noise corrupted.

Several methods have been reported to estimate the oral cavity shape and place of articulation from the images in the direct imaging databases [15]–[17]. Story [15] used an iterative bisection method to obtain the axial curve of the oral cavity contours from the XRMB data. As shown in Fig. 1, the iteration is initiated using points A and B, the midpoints between the contours at two ends of the oral cavity. The oral cavity is segmented by perpendicular bisector Lp of line segment AB, cutting the upper and lower contours at D and E, respectively. The segment DE is taken as the mouth opening at the interface of the two segments. The midpoint of DE is used to get the line segments AC and CB. Repetition of the segmentation process results in the four segments AG, GC, CI, and IB. The process is iterated for obtaining the desired numbers of segments. Using 5 iterations providing 33-point axial curve, oral cavity shapes were obtained for 11 vowels and vowel-vowel sequences from four speakers. Bresch *et. al.*[16] used the iterative bisection method using 3 iterations to obtain 9-point axial curve and corresponding oral cavity shapes from the MRI images and fitted a cubic smoothing spline through these points. Oral cavity configuration for the vowel /a/ was obtained and the corresponding resonant frequencies, estimated using VTAR software [18], matched well with those obtained from analysis of the speech signal.

Jagbandhu [17] used a segmentation method for estimating the place of articulation from the XY articulatory plots of the XRMB database. In this method, the upper and lower contours are divided into 50 equal segments. The midpoints of line

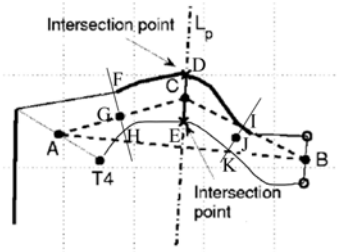

Fig. 1. Iterative bisection method used by Story [15]

segments joining the corresponding points on the two contours are obtained as the points forming the axial curve. Normal to the axial curve at a given point is drawn by taking the mean of the slopes of the line segments on the left and right sides of the point as the slope of the curve. The length of the normal between the two contours is taken as the oral cavity opening at the point. The method was applied for finding the place of articulation for /ΛCa/ utterances with stop consonants /b/, /d/ and /g/ from 35 speakers. The results were validated by comparing them with those obtained by manual measurement along the length of the oral cavity, using tongue profile method [5], by adding the straight line distances along the curve joining the lower lip marker and the pellet markers.

To evaluate the methods reported in [15]–[17], they were used for estimating the oral cavity shapes from MRI images. Results for /a/, as used in the investigation by Bresch *et. al.* [16], are shown in Fig. 2 as an example. In Fig. 2(a) for the iterative bisection method, the normals drawn to the axial curve (marked as dotted lines) deviate from the bisector lines (marked as dark lines). As this method starts with a global slope and captures the local axial slope in subsequent iterations, it may not be able to capture the local axial slope at bisectors in the initial iterations. Further, the bisector lines after a few iterations start crisscrossing due to the sharp bends in the axial curve. Sharp bends can be avoided by limiting the number of iterations and applying smoothing spline [16] as shown in Fig. 2(b). A significant deviation between the smoothened axial curve (marked as dotted) and the curve joining the midpoints of the normals (marked as dark) indicates that the axial curve may not divide the oral cavity symmetrically. Fig. 2(c) shows the result for the segmentation method [17]. The estimated axial curve (marked as dotted) deviates from the curve joining the midpoints of the normals (marked as dark) in the region where the two contours are not symmetric. Similar problems were observed in application of these methods on many other images. Hence, there is a need for an automated method for estimating the axial curve which approximately divides the normals into two equal segments, such that the deviation between the axial estimate and curve joining the midpoints of normals is minimal.

III. Proposed Technique

An automated technique is developed for estimating the place of articulation by graphical processing of the upper and lower contours of the oral cavity. It assumes the acoustic wave propagation to be along an axial curve between the two contours with the normal to the curve representing the wave front. The axial curve is iteratively estimated as an axis of

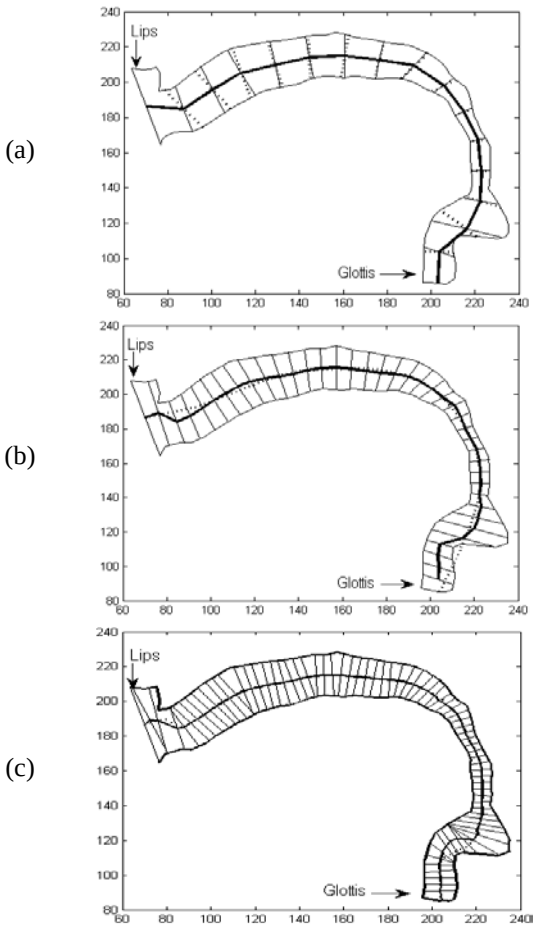

Fig. 2. Estimation of oral cavity opening from the lower and upper contours in an MRI image of /a/ using (a) iterative bisection method [15], (b) iterative bisection method and smoothening spline [16] with smoothening factor of 0.99, and (c) segmentation method [17]. The x and y distances are in number of pixels.

symmetry, approximately bisecting the normals to it. Length of the normal between the two contours provides an estimate of the oral cavity opening.

A set of equidistant points are marked on the lower and upper contours. Midpoints of the lines joining the corresponding points on the two contours form the initial estimate of the axial curve. A least-squares approximation based B-spline is fitted on these points for smoothening the curve by reducing sharp bends. For this purpose, Matlab function "spap2", with internal generation of the knot vector for specified number of knots is used. The smoothening controls the curve locally where it does not bisect the normals without significantly affecting it at other places. Normals are drawn at a set of equidistant points on the smoothened curve. The points of intersection of normals with upper and lower contours are obtained and their midpoints form the revised estimate of the axial curve. The process of smoothening and revising the axial curve is repeated to improve the approximation, until the RMS difference between the successive estimates is less than a specified fraction of the mean oral cavity opening

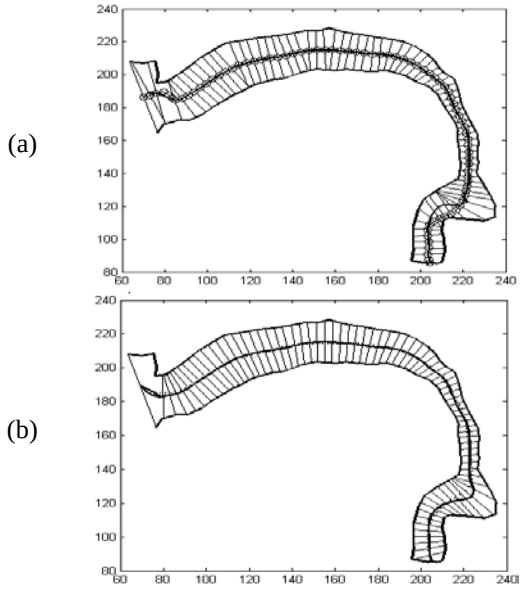

Fig. 3. Application of the proposed Iterative Axial Curve method for estimation of the oral cavity opening from the lower and upper contours in an MRI image of /a/: (a) first iteration, (b) final iteration. The x and y distances are in number of pixels.

An increase in the number of knots in the B-spline decreases the approximation error but it also reduces the smoothening of the axial curve, leading to crisscrossing of normals at sharp bends. Iterations are started with the knot vector length of 12 and stopped if the RMS difference is less than 5% and number of crisscrossing is 5 or less. If this criterion is not satisfied after 30 iterations, the knot vector length is decremented by one and the iterative process is repeated. It was found that the condition for stopping the iterations was satisfied with knot vector length of 5 or higher.

Fig. 3 shows an example of the processing for /a/. Fig. 3(a) shows the initial axial curve as a continuous curve and the revised estimate as a curve with circle markers. Result from the final iteration is shown in Fig. 3(b). The dotted curve is the estimated axial curve and the dark curve is formed by joining the midpoints of the normals. The two curves are almost superimposed. Thus the deviation between the two curves is much smaller than the corresponding deviations observed in the results shown in Fig. 2 as obtained by the earlier methods.

IV. Test Results

The technique presented in the previous section was evaluated using oral cavity images for 8 phonemes (3 vowels, 3 stops, and 2 fricatives) by two speakers (one male, one female) from the XRMB database [7] and two speakers (one male, one female) from the MRI database [14] as the test material. Accuracy of the estimated values was obtained with reference to the manually marked and measured values of the visually estimated place of maximum constriction in prints of the oral cavity images. An axial curve was marked for symmetrically dividing the cavity in upper and lower halves and the place of maximum constriction was marked on this curve. The distance of the maximum constriction from the lips along the axial curve was measured to get the place of articulation. The distances measured in the print were scaled

back to mm in case of XRMB articulatory plots and to number of pixels in case of MRI images. Means of the values as measured by 5 observers (fellow researchers in the lab) were used as the reference values.

A consistent manual marking of the place of articulation for the medial vowel /a/ was difficult due to a relatively large opening and a lack of well defined place of maximum constriction. For other seven phonemes, the places of articulation could be consistently marked. For these phonemes, the values obtained by manual measurement and those estimated by the automated technique are given in Table I. Very small standard deviations of the manually measured values indicate consistency of manual marking across observers. A relatively large standard deviation for /i/ from Speaker 2 was due to two nearly similar places of maximum constriction. For the bilabial stop /p/, the place of articulation is clearly visible in the images and hence the standard deviation is negligible.

The errors in the estimated values with reference to the manually measured values are of the same order as the standard deviation of the manually measured values. Hence the two set of values may be considered to have a close match for all phonemes from both the databases. The two set of values for oral cavity opening (not given in the table) at the place of articulation also showed a good match.

TABLE I. PLACE OF ARTICULATION: (A) MANUALLY MEASURED (MAN.) MEAN AND STANDARD DEVIATION (SD) OF THE VALUES OBTAINED BY 5 OBSERVERS AND (B) ESTIMATED (EST.) VALUES USING THE AUTOMATED ITERATIVE AXIAL CURVE TECHNIQUE.

Pho-ne-me	Place of articulation (distance of maximum constriction along the axial curve from the lips): mm for XRMB, pixels for MRI											
	Speaker 1 (F, XRMB)			Speaker 2 (M, XRMB)			Speaker 3 (F, MRI)			Speaker 4 (M, MRI)		
	Man.		Est.	Man.		Est.	Man.		Est.	Man.		Est.
	Mean	SD		Mean	SD		Mean	SD		Mean	SD	
/i/	53.9	0.8	54.2	35.8	10.2	53.7	47	4	47	108	2	105
/u/	60.2	0.7	60.5	69.4	0.6	70.2	124	1	126	108	2	106
/p/	0.0	0.0	0.0	0.0	0.0	0.0	8	1	10	6	1	5
/t/	30.7	0.4	30.3	23.3	0.5	22.9	36	2	36	49	5	49
/k/	63.4	0.4	63.2	61.9	0.4	62.1	135	7	133	119	4	119
/s/	27.2	1.9	24.8	23.3	0.5	22.6	36	2	35	47	1	46
/sh/	31.9	0.7	32.7	26.1	0.5	25.4	47	1	50	60	1	61

IV. CONCLUSION

A technique has been presented for automated estimation of the place of articulation from the upper and lower contours of the oral cavity image. The axial curve of the oral cavity is iteratively estimated as an axis of symmetry, approximately bisecting the normals to it. The estimation of the place of articulation has been validated for utterances with eight phonemes for four speakers. It needs to be further validated for utterances from a larger number of speakers. Its use for validation of the place of articulation estimated by analysis of speech signal will help in the development of speech training aids. It can also be useful as a tool in speech production related research.

REFERENCES

[1] R. G. Crichton and F. Fallside, “Linear prediction model of speech production with applications to deaf speech training,” Proc. IEE Control Sci., vol. 121, 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. Rehab. Eng., vol. 2, no. 4, pp. 189–196, Dec. 1994.

[3] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy–technology interface in Computer Assisted Pronunciation Training," Computer Assisted Language Learning, vol. 15, pp. 441 – 467, 2002b.

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

[5] P. C. Pandey and M. S. Shah, “Estimation of place of articulation during stop closures of vowel-consonant-vowel utterances,” IEEE Trans. Audio, Speech, Language Process., vol. 17, no. 2, pp. 277-286, 2009.

[6] S. Kiritani, K. Ito, and O. Fujimura, "Tongue-pellet tracking by a computer-controlled X-ray micro beam system," J. Acoust. Soc. Am., vol. 57, no. 6, pp.1516 -1520, 1975.

[7] J. R. Westbury, “X-ray microbeam speech production database user’s handbook (version 1.0),” 1994. [Online]. Available: www.haskins.yale.edu/staff/gafos_downloads/ubdbman/ (Last access: June 2015).

[8] J. Z. Summer, “Articulograph AG100 electromagnetic articulation analyzer,” 1997. [Online]. Available: http://www.linguistics.ucla.edu/faciliti/facilities/physiology/Emamual.html#1 (Last access: June 2015).

[9] A. Rouco and D. Recasens, “Reliability of electromagnetic midsagittal articulometry and electropalatography data acquired simultaneously,” J. Acoust. Soc. Am., vol. 100, no. 5, pp. 3384–3389, 1996.

[10] S. Narayanan, K. Nayak, A. Sethy, and D. Byrd, “An approach to realtime magnetic resonance imaging for speech production,” J. Acoust. Soc. Am., vol. 115, pp. 1771–1776, 2004.

[11] M. Stone, ‘‘A three-dimensional model of tongue movement based on ultrasound and X-ray microbeam data,’’ J. Acoust. Soc. Am., vol. 87, pp. 2207–2217, 1990.

[12] A. A. Wrench and H. J. William, “A multichannel articulatory database and its application for automatic speech recognition,” in Proc. 5th Seminar on Speech Production: Models and Data, 2000, Bavaria, pp. 305–308.

[13] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, ‘‘Synchronized and noise-robust audio recordings during real-time MRI scans,” J. Acoust. Soc. Am., vol. 120, pp. 1791–1794, 2006.

[14] 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. 136, pp. 1307–1311, 2014.

[15] B. H. Story, “Time dependence of vocal tract modes during production of vowels and vowel sequences, ” J. Acoust. Soc. Am., vol. 121, no. 6, pp. 3770-3789, 2007.

[16] E. Bresch, J. Adams, A. Pouzet, S. Lee, D. Byrd, and S. Narayanan, “Semi-automatic processing of real-time MR image sequences for speech production studies,” in Proc. Int. Seminar Speech Prod., 2006, Ubatuba, pp. 427-434

[17] Jagabandhu, “A visual feedback of vocal tract shape for speech training,” M.Tech. dissertation, EE Dept.., IIT Bombay, India, 2012.

[18] X. Zhou, Z. Zhang, and C.Y. Espy-Wilson, “VTAR: A Matlab-based computer program for vocal tract acoustic modeling,” J. Acoust. Soc. Am., vol. 115, no. 5, pp. 2543, 2004.