Rapid Semi-automatic Segmentation of Real-time Magnetic Resonance Images for Parametric Vocal Tract Analysis

Michael I. Proctor[1,2], Danny Bone[1], Nassos Katsamanis[1], Shrikanth Narayanan[1,2]

[1]Viterbi School of Engineering, University of Southern California, USA
[2]Department of Linguistics, University of Southern California, USA

mproctor@usc.edu

Abstract

A method of rapid semi-automatic segmentation of real-time magnetic resonance image data for parametric analysis of vocal tract shaping is described. Tissue boundaries are identified by seeking pixel intensity thresholds along tract-normal gridlines. Airway contours are constrained with respect to a tract centerline defined as an optimal path over the graph of all intensity minima between the glottis and lips. The method allows for superimposition of reference boundaries to guide automatic segmentation of anatomical features which are poorly imaged using magnetic resonance – dentition and the hard palate – resulting in more accurate sagittal sections than those produced by fully automatic segmentation. We demonstrate the utility of the technique in the dynamic analysis of tongue shaping in Tamil liquid consonants.

Index Terms: speech production, vocal tract segmentation, MRI, tongue shaping, articulatory analysis

1. Introduction

Real time magnetic resonance imaging (rtMRI) promises to be a viable tool for studying human speech production [1]. An important process in the analysis of MRI speech data is segmentation of the vocal tract. Tissue boundaries can be delineated manually with great accuracy (e.g. [2], [3], [4]), but such approaches are labor intensive, subject to inter-frame inconsistencies, and unsuitable for real-time video sequences, which can consist of hundreds of image frames. A robust technique has been developed to automatically segment the upper airway into anatomical regions of linguistic interest [5]; however, this method is computationally intensive and requires considerable supervision to initialize.

Two major problems common to all of these approaches are (i) locating the dentition, which does not image in the MR sequences typically used for upper airway analysis, and (ii) correcting tissue segmentation compromised by motion blur, low signal-to-noise ratios, or inherent scarcity of soft tissue. Additionally, further methods of data reduction or transformation are required when tract boundaries identified by an unconstrained segmentation procedure are to be quantified and compared.

1.1. Rationale

The goal of the current study is to develop a method of rapidly segmenting the airway in midsagittal MR images, with a minimum of supervision, in a manner which would be directly applicable to the parametric analysis of vocal tract shaping during the production of intervocalic consonants. Because the preferred method of analysis requires the superimposition of an analysis grid on the vocal tract [6], we make use of this grid from the outset to constrain the tissue identification algorithm. Additionally, we wish to provide a method by which the speech researcher may guide the algorithm with datasets which are problematic for fully automatic approaches, by selectively specifying reference boundaries to improve airway segmentation.

2. Method

2.1. Image Acquisition

All data analyzed in this study were acquired using an rMRI protocol developed specifically for the dynamic study of speech production [1]. Subjects' upper airways were imaged in the midsagittal plane with video reconstruction rates varying between 22.4 and 32 f.p.s., then reintegrated with audio simultaneously recorded at 20 kHz [7], to allow for dynamic audio-visualization of subjects' speech production.

2.2. Analysis Grid Construction

Building on the method developed by Öhman [6] and Maeda [8], a composite analysis grid was superimposed on each image frame to be segmented. Anatomical landmarks were chosen at (i) the glottis, (ii) the highest point on the palate, (iii) the alveolar ridge, and (iv) the lips. A lingual origin was located equidistant from the palate and the rear wall of the pharynx, close to the centre of mass of the tongue in resting position. Horizontal pharyngeal gridlines were superimposed at regular intervals from the glottis to the level of the lingual origin.

A semi-polar grid was constructed over the mid-oral vocal tract, extending from the mid pharynx through to the alveolar ridge, by projecting equi-spaced radial gridlines from the lingual origin. A second origin was located above the incisors, from which radial gridlines were projected through the anterior oral and sublingual cavities. To complete the grid, vertical lines were superimposed over the region of the vocal tract anterior to the teeth, extending beyond the lips (Fig. 1).

2.3. Segmentation

For each frame of interest, tissue boundaries were located by traversing the superimposed analysis grid and characterizing the change in pixel intensity along the paths defined by each gridline. A typical intensity profile calculated along a tract-normal gridline will feature several local minima, one of which will be located at the bottom of an 'intensity well' corresponding to the tract airway (Fig. 2). In a midsagittal MR image acquired with adequate SNR, tissue boundaries will typically be located symmetrically around this centerpoint, in the vicinity of the interval of steepest change in pixel intensity.
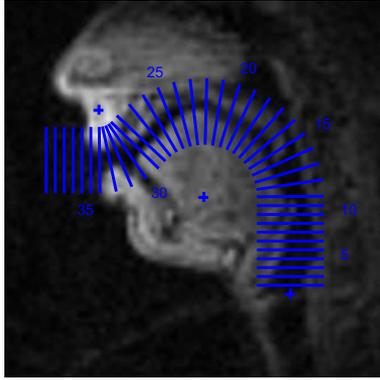
Figure 1: *Composite semi-polar analysis grid superimposed on a midsagittal MR image of a male vocal tract.*
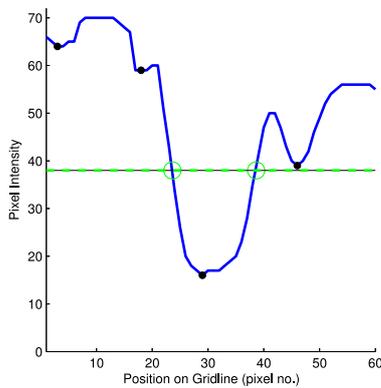

Figure 2: **Intensity profile of a tract-normal gridline**, *showing four local minima and a central 'well' corresponding to the darker pixels associated with the vocal tract airway. A horizontal line intersects the profile at the pixel intensity thresholds corresponding to tissue boundaries in this region of the tract.*

2.3.1. Graph Construction

The pixel coordinates and intensity values of all local minima identified on the intensity profile of each gridline were calculated for each frame. A graph was constructed in which each intensity minimum, the glottis, and the mid-labial point, was represented by a node. Each node on gridline g_i was connected to each node on the immediately adjacent gridlines g_{i-1} and g_{i+1}. Nodes on the first gridline were also connected to the glottal node; nodes on the final gridline to the mid-labial node.

For a typical male vocal tract, imaged with adequate SNR, the set of all pixel intensity minima in the vicinity of the tract airway can be represented as graph of approximately 150 nodes and 500 edges distributed over 30 equi-spaced gridlines (Fig. 3).

2.3.2. Centerline Determination

Having identified all intensity minima in the vicinity of the airway, the vocal tract centerline was estimated for each frame by finding an optimal path through the graph from the glottis to the lips. The weight w_{ij} of the edge connecting node n_i to n_j was calculated as the weighted sum of the destination node intensity I_j and the euclidean distance d_{ij} between the coordinates of the corresponding pixels (Eq. 1):
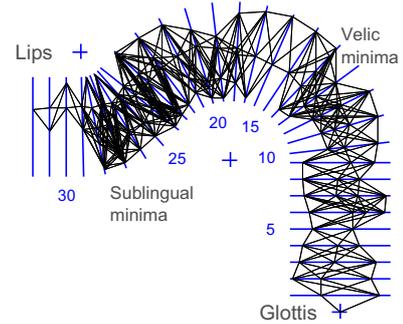

Figure 3: *Construction of a **graph connecting pixel intensity minima on adjacent gridlines**. Terminal nodes on the graph are defined at glottal and mid-labial points on each image frame.*

$$w_{ij} = \alpha.I_j + (1 - \alpha).d_{ij} \qquad (1)$$

where α is a weighting factor used to preferentially select shorter paths ($\alpha \to 0$) or paths defined over darker pixel nodes ($\alpha \to 1$). From the graph connecting all grid-constrained intensity minima, an optimal path corresponding to the tract centerline was calculated using Dijkstra's algorithm [9].

2.3.3. Tissue Boundary Estimation

Tissue boundaries were estimated by locating the steepest points on the intensity function within a specified threshold of the intensity range for a given gridline. For the intensity function $I(p)$ defined over the set of pixels $p \leftarrow [1..n]$ centered at pixel $p = c$ with intensity I_c, the initial estimate for tissue boundary intensity I_{th} was calculated using Eq. 2:

$$I_{th} = I_c + \beta.(\min(\max_{p \leftarrow [1..c]}(I(p)), \max_{p \leftarrow [c..n]}(I(p))) - I_c) \quad (2)$$

where β is a thresholding factor chosen according to the contrast and SNR of a given image sequence. The difference function $I'(p)$ was then used to locate the steepest region in each half of the intensity function within a specified range γ of I_{th}. For each gridline, with total intensity range $\Delta I(p)$, the inner and outer tissue boundary intensities I_{bi} and I_{bo} were located using using Eqs. 3 and 4.

$$I_{bi} = \max_{|I(p) - I_{th}| < \gamma.\Delta I(p)} I'(p), \ p \leftarrow [1..c] \qquad (3)$$

$$I_{bo} = \min_{|I(p) - I_{th}| < \gamma.\Delta I(p)} I'(p), \ p \leftarrow [c..n] \qquad (4)$$

2.3.4. Labial Segmentation

Labial tissue boundaries were segmented automatically by extending the analysis grid beyond the point of maximum lip protrusion and backtracking until the intensity of any local maxima exceeded a specified threshold, defined as a ratio of the difference between mean pixel intensities of gridlines superimposed entirely on regions of background noise and those straddling some amount of soft tissue. Upper and lower lip thresholds were calculated independently in order to capture any asymmetries in labial protrusion.

2.3.5. Palatal and Dental Correction

Because of the scarcity of soft tissue in the region of the hard palate, MR imaging often fails to resolve the upper mid-oral cavity boundary with sufficient contrast, if at all. As a result, the segmentation algorithm typically fails to locate the true palate when deployed with parameters which best identify tissue boundaries in other regions of the tract. Likewise, dentition cannot be imaged with the pulse sequences used in these studies, yet it is important to be able to approximate the location of subjects' teeth when segmenting the midsagittal airway.

For each image sequence, the palatal contour was automatically identified in the frame which best resolved the tissue in this region, and manually adjusted where necessary. Upper dental boundaries were estimated from frames in which the teeth were surrounded by soft tissue. Dental and palatal reference boundaries were then superimposed on each other image frame to guide the automatic segmentation algorithm.

A DFT-based algorithm [10] was used to determine whether the subject's head had moved from the postures assumed in the palatal or dental reference frames, allowing for correction during the registration process. Vocal tract contours segmented with and without palatal correction are illustrated in Fig. 4.
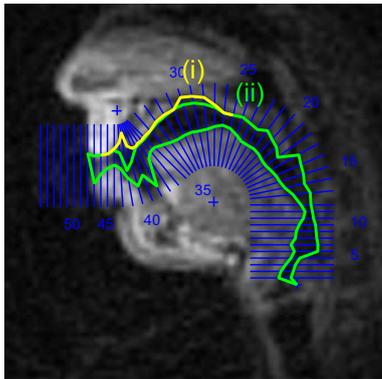

Figure 4: **Correction of palatal contours**. *Male Tamil speaker. (i) Tissue boundaries derived by unsupervised segmentation; (ii) modified boundaries incorporating reference palate.*

2.3.6. Tongue Smoothing

Because of the abrupt changes in curvature characteristic of midsagittal vocal tract contours – introduced by anatomical discontinuities around the epiglottis, velum and dentition – it difficult to deploy a global smoothing algorithm capable of reducing noise without also compromising the accuracy of the segmentation of these regions. Vocal tract contours corresponding to the tongue edge, on the other hand, are well suited to refinement through filtering and guided reconstruction because of their anatomical homogeneity.

We make use of the discrete cosine transformation (DCT) to reduce noise in the lingual tissue boundary segmented using Eq. 3. From the subset of gridlines covering the tongue, we extract a 1-dimensional lingual contour – expressed as set of radial distances from the innermost gridline endpoint – and low-pass filter the curve by eliminating all DCT components whose magnitude does not exceed a specified threshold, before reconstructing the tongue contour from the inverse DCT.

3. Results

3.1. Parameter Selection

For the MRI data analyzed so far, best segmentation results were achieved with gridline spacings in the range $5 < dGL < 8$ mm and radial gridline spacings in the range $4° < d\theta < 8°$. The centerline estimation algorithm (Eq. 1) performed best when a centerline weighting factor was specified in the range $0.1 < \alpha < 0.25$ (preferring shortest total path over relative pixel darkness of local minima). Tissue boundary detection (Eqs. 2–4) was most effective when using an intensity-thresholding factor in the range $0.45 < \beta < 0.7$, and an inflection-search constraint factor in the range $0 < \gamma < 0.1$.

For the majority of sequences, in which anterior pharyngeal displacement was negligible, more robust automatic segmentation of the rear pharyngeal wall was achieved by averaging the tissue boundaries located across all image frames in the sequence and using the mean pharyngeal contour as a reference boundary. Tongue smoothing was found to be most effective using DCT components responsible for 75% of the lingual curvature, or 85% of curvature when analyzing data which contains a significant number of retroflexed consonants.

3.2. Quantifying Segmentation Accuracy

To examine the accuracy of the automatic segmentation algorithm on a broad range of MR Image data, 50 image frames taken from video sequences of five different speakers were analyzed. 10 frames each from three male and two female speakers were randomly selected from a multilingual rtMRI speech database. The image set encompassed a variety of articulatory postures, including mid-vocalic, mid-consonantal, transition and rest frames.

The midsagittal airway in each test frame was first segmented manually by a phonetician experienced in analyzing MR image data; tissue boundaries were then identified automatically in the same image using an 8mm-spaced grid and segmentation parameters $\alpha = 0.15$, $\beta = 0.55$, $\gamma = 0.05$. The difference between the sagittal distance functions extracted from these two boundaries was calculated for each image, and averaged for each speaker (Fig. 5). Mean segmentation errors (root mean square differential displacement in mm per gridline) were calculated for all speakers (Table 1).
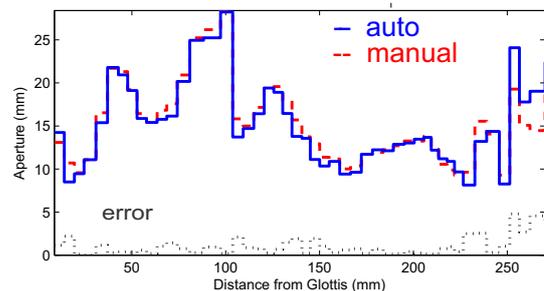

Figure 5: *Mean midsagittal distance functions calculated from (i) manual, and (ii) automatic segmentation of airways. (iii) Mean segmentation error (mm/gridline): 10 image frames randomly extracted from a 9 sec. utterance by Subject M1 (Tamil).*

The data in Table 1 show that most segmentation errors occur in the anterior region of the vocal tract, where rapid anatomical transitions occur around the lips and teeth. The success

of the algorithm in this region depends largely on the accuracy with which the reference dental boundaries can be registered with each frame and reconciled with the local oral intensity thresholds. Automatic segmentation is more consistently successful when identifying the more homogenous tissue boundaries in the mid-oral and pharyngeal regions. Automatic segmentation accuracy was poorest when tissue boundaries deviated from normal alignment with respect to the analysis gridlines, as was observed in the articulation of heavily retroflexed obstruents, and certain types of velic configurations.

Speaker	Pharynx	Mid-oral	Dental	Total
M1 (Tamil)	0.774	0.950	1.314	1.043
M2 (English)	0.754	0.329	1.770	0.947
M3 (German)	0.536	0.816	1.559	0.877
W1 (Serbian)	0.778	0.629	1.026	0.767
W2 (Mandarin)	1.284	0.792	2.398	1.310
Mean error	0.825	0.703	1.613	0.988

Table 1: *Mean segmentation errors (mm/gline) in 50 frames.* $(d_{GL} = 8mm; , \alpha = 0.15, \beta = 0.55, \gamma = 0.05, th_{DCT} = 0.85)$

4. Applications

Because it affords rapid, accurate, unsupervised identification of tissue boundaries across long sequences of video frames, the method described here is beginning to provide important insights into the dynamics of articulation. An ideal application for this technique is the analysis of tongue shaping in classes of consonants which are hypothesized to involve complex coordination of lingual gestures: liquids and fricatives. A segmented midsagittal MR image frame acquired from a male speaker in a study of Tamil liquid consonants is illustrated in Fig. 6.
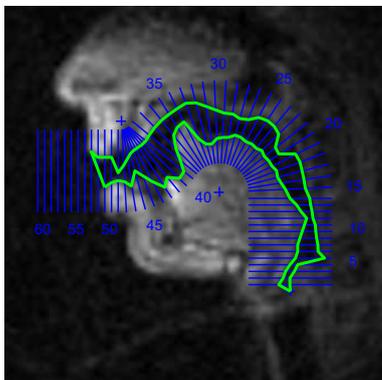

Figure 6: *Auto-segmented Tamil retroflex rhotic* [ɻ]. $(d_{GL} = 6mm, \alpha = 0.15, \beta = 0.55, \gamma = 0.05, th_{DCT} = 0.85.$

5. Discussion

When deployed with MR image data acquired with adequate signal-to-noise ratios, the current algorithm has proven to be an effective tool for the rapid automatic segmentation of a wide variety of vocal tracts with sufficient accuracy to allow for direct sagittal distance function analysis. There is a need, however, for refinement of the method to improve its ability to automatically segment vocal tracts in noisier data – a characteristic of MR images acquired or reconstructed at higher framerates.

Two ways in which the the algorithm might be rendered more robust are to incorporate information about the timecourse of articulatory displacement in the search for tissue thresholds, and to further constrain the analysis grid using some method of anatomically-informed principal components analysis – an approach which has been explored in previous work ([1], [3]).

5.1. Future Directions

While all of the examples shown in this paper involve segmentation of the midsagittal airway, the same approach to tissue boundary identification – using automatic thresholding of intensity profiles over an anatomically-guided graph – can be used to process MR data acquired from other imaging planes.

Automatic segmentation of sets of parasagittal slices is being used to construct three-dimensional models of the vocal tract, which are invaluable for studying tongue shaping in liquid consonants. A modified verion of the algorithm has been deployed to extract tissue boundaries from axial images of the pharynx: a technique which can be used to quantify differences in the articulation of voiced and voiceless consonant pairs.

6. Conclusion

The method described here addresses a need for linguists working with real-time MRI data, by providing a tool for rapidly and accurately segmenting vocal tract image frames, with a minimum of supervision, in a manner which is consistent with the requirements of parametric vocal tract analysis.

7. Acknowledgements

8. References

[1] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *JASA*, vol. 115, no. 4, pp. 1771–1776, 2004.

[2] V. Lecuit, "Sagittal cut to area function transformations: A comparative study," Mémoire, Université Libre de Bruxelles, 1992.

[3] D. Demolin, T. Metens, and A. Soquet, "Three-dimensional measurement of the vocal tract by mri," *Proc. 4th Intl. Conf. on Spoken Language Processing, Philadelphia, USA*, pp. 272–275, 1996.

[4] M. Stone, E. P. Davis, A. S. Douglas, M. N. Aiver, R. Gullapalli, W. S. Levine, and A. J. Lundberg, "Modeling tongue surface contours from cine-mri images," *Journal of Speech and Hearing Research*, vol. 44, no. 5, pp. 1026–1040, 2001.

[5] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 28, no. 3, pp. 323–338, Mar 2009.

[6] S. E. G. Öhman, "Numerical model of coarticulation," *J. Acoust. Soc. Am*, vol. 41, no. 2, pp. 310–320, 1967.

[7] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *J. Acoust. Soc. Am*, vol. 120, no. 4, pp. 1791–1794, 2006.

[8] S. Maeda, "Un modèle articulatoire de la langue avec des composantes lineaires," *10ème Journées d'Etude sur la Parole*, pp. 1–9, 1979.

[9] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, no. 1, pp. 269–271, 1959.

[10] M. Guizar-Sicairos, S. T. Thurman, and J. R. Fienup, "Efficient subpixel image registration algorithms," *Optics Letters*, vol. 33, no. 2, pp. 156–158, 2008.