INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Vocal Tract Airway Tissue Boundary Tracking for rtMRI using Shape and Appearance Priors

Sasan Asadiabadi and Engin Erzin

Multimedia, Vision and Graphics Laboratory,
College of Engineering, Koç University, Istanbul, Turkey
[sabadi15,eerzin]@ku.edu.tr

Abstract

Knowledge about the dynamic shape of the vocal tract is the basis of many speech production applications such as, articulatory analysis, modeling and synthesis. Vocal tract airway tissue boundary segmentation in the mid-sagittal plane is necessary as an initial step for extraction of the cross-sectional area function. This segmentation problem is however challenging due to poor resolution of real-time speech MRI, grainy noise and the rapidly varying vocal tract shape. We present a novel approach to vocal tract airway tissue boundary tracking by training a statistical shape and appearance model for human vocal tract. We manually segment a set of vocal tract profiles and utilize a statistical approach to train a shape and appearance model for the tract. An active contour approach is employed to segment the airway tissue boundaries of the vocal tract while restricting the curve movement to the trained shape and appearance model. Then the contours in subsequent frames are tracked using dense motion estimation methods. Experimental evaluations over the mean square error metric indicate significant improvements compared to the state-of-the-art.

Index Terms: speech production, vocal tract, contour tracking

1. Introduction

Knowledge on the time-varying dynamic changes of the vocal tract is the basis to understand the human speech production system. In the discrete concatenated tube models of the vocal tract, the resonance frequencies or, more generally, the transfer function of the vocal tract are estimated from the cross sectional area function of the tract. Knowing the time evolution of the vocal tract transfer function implies the speech being produced. Therefore the study of human vocal tract shape evolution could play a significant role in many speech processing applications.

Intensive research has been conducted on segmentation of the vocal tract (VT) airway tissue boundary. Among all of the medical imaging technologies developed, attention toward the MRI has been growing since MRI is safe for the subject and can provide full mid-sagittal view images of the vocal tract as well as a 3-D outlook.

Several parametric models have been constructed for the vocal tract using X-ray and MRI images of the tract ([1],[2],[3]). One of the first statistical models of the vocal tract was introduced by [1] using 1000 mid-sagittal X-ray images of the vocal tract. A semi-polar coordinate system for measuring the the lateral outlines of the vocal tract was proposed by Maeda. The hand-labeled contours extracted from the X-ray images were projected onto the semi-polar grid system and the variation of the contour points were analyzed using Principle Component Analysis.

Bresch and Narayanan [4] presented a novel approach to segmentation of upper airway real-time MRI in a frequency domain. Their algorithm uses an anatomically informed model for the vocal tract, fitting to a new image through a gradient descent optimization procedure. Despite the visually accurate results in their paper, the algorithm's computational complexity is so high that it is impractical for real-time research.

A semi-automatic rapid VT airway tissue boundary segmentation is proposed by Proctor [5]. In his work the image intensity profile along Maeda's gird line system is investigated. A graph is constructed by connecting all the intensity profile local minimas on all of the grid lines. The central airway path through the tract is then estimated by finding the optimal path to minimize a defined score through the constructed graph, using the Dijkstra algorithm. The airway tissue boundary are estimated by locating the first point on the grid lines on either side of the center line crossing a threshold.

A rapid VT airway tissue boundary segmentation was proposed by Kim et al. [6]. In this work, the image intensity profile along Maeda's gird line system [1] is investigated. A graph is constructed by connecting all of the pixels on each grid line. The central airway path through the tract is then estimated by finding the optimal path to minimize a defined score through the constructed graph using the Viterbi algorithm. The airway tissue boundary is estimated by locating the first point on the grid lines on either side of the center line crossing a threshold. This algorithm suffers from a quite high run-time and the accuracy of detecting the airway tissue boundaries, is directly dependent on the accuracy of center-line estimation.

In this paper, we present a robust vocal tract airway tissue boundary tracking for real-time MRI. For this purpose, the USC-TIMIT database [7], which comprises mid-sagittal MRI videos, is utilized. The traditional active contour models (ACM) [8] and active shape models (ASM) [9] are used to detect the airway tissue boundaries of the vocal tract. We define a new snake energy function to detect the airway tissue boundaries using the trained shape and appearance models as well as other energy terms for training imperfection compensation. An advantage of using both ASM and ACM over using ACM only, as presented in [10], is smaller tracking errors in the frames when articulator occlusion occurs.

2. Methods

In this section, we present details of the statistical shape and appearance models, as well as the the proposed airway tissue boundary tracking algorithm.

2.1. VT shape model

Vocal tract shape can be considered as a prior information for a generalized airway boundary tracking problem. Hence, we define a shape model for the VT to utilize it as a prior in the

http://dx.doi.org/10.21437/Interspeech.2017-1016

tracking system.

The landmark points of each training image is stored as a column vector:

$$X = [x_1, x_2, ..., x_n, y_1, y_2, ..., y_n]^T. \quad (1)$$

Assume that we have N such vectors X_j, which are generated for various possible shapes of the VT and stacked in a $2n \times N$ matrix D as

$$D = [X_1|X_2|...|X_N] \quad (2)$$

The vocal tract size and position in the recorded videos might vary from one subject to another, therefore alignment of the shapes into a common co-ordinate frame is necessary to eliminate the changes in the tract size and position. The popular Procrustes method [11] is used for the alignment purpose.

Principal Component Analysis (PCA) is used to model the distribution of the extracted aligned data by detecting the major directions of variation. The variation around the mean is described by the eigenvectors of the covariance of matrix D, which can be computed as

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \quad (3)$$

$$S = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})^T \quad (4)$$

Then the model can generate new shapes as

$$X = \bar{X} + Pb, \quad (5)$$

where $P = [P_1|P_2|...|P_t]$ contains t eigenvectors of the covariance matrix S, corresponding to the t largest eigenvalues λ_i and b is a t-dimentional model parameter vector, which is computed as $b = P^T(X - \bar{X})$.

2.2. VT appearance model

We utilize image structure as a second prior for the tracking problem. A statistical model is trained to model the image structure around the VT airway tissue boundaries. For the i-th training image and the j-th model point in the image, k pixels are sampled on each side of the point along the normal, and the normalized derivative image intensities at the sampled pixels are stored in a $2k + 1$ column vector g_{ij} as

$$g_{ij} = [g_{ij,1}, g_{ij,2}, ..., g_{ij,2k+1}]^T \quad (6)$$

For a model point v_j the normalized derivative gray-level matrix G_j is constructed from the g_{ij} vectors over all the training images as

$$G_j = [g_{1j}, g_{2j}, \ldots, g_{Nj}]. \quad (7)$$

As utilized for the shape model, a PCA analysis is carried out to obtain a statistical model for each point v_j by computing mean ($\bar{G}_j$) and covariance (S_{gj}) of the $(2k + 1)$-by-N matrix G_j.

The quality of fit of a new sample G_s to the model trained for point v_j is given by the Mahalanobis distance,

$$f(G_s) = (G_s - \bar{G}_j)^T Sg^{-1}(G_s - \bar{G}_j), \quad (8)$$

which defines the distance from the sample to the mean model. Note that the covariance matrix in the equation above might be singular, i.e. not invertible. Matrix S_g can be factored using eigenvalue decomposition to

$$S_g = Q\Lambda Q^T, \quad (9)$$

where Q and Λ being the eigenvectors and eigenvalues of S_g, respectively. Then equation (8) can be rewritten as

$$f(G_s) = (G_s - \bar{G}_j)^T Q\Lambda^{-1}Q^T(G_s - \bar{G}_j). \quad (10)$$

The projection of $(G_s - \bar{G}_j)$ onto all eigenvectors present in Q is computes as

$$b_g = Q^T(G_s - \bar{G}_j). \quad (11)$$

Note that b_g is a $2k+1$ column vector and Λ is a $(2k+1)\times(2k+1)$ diagonal matrix with eigenvalues of S_g on the diagonal. By inserting (11) into (10), the Mahalanobis distance can be written as

$$f(G_s) = b_g^T\Lambda^{-1}b_g = \sum_{i=1}^{2k+1}\frac{b_{g,i}^2}{\lambda_{g,i}}. \quad (12)$$

For a new frame, given that an initial shape model is known, n_s pixels are sampled each side of the j^{th} model point along the normal. The $2n_s + 1$ points (including the current model point) are the candidate locations where the model point can move to. For each candidate point, a $2k + 1$ pixel derivative gray-level profile centered to that candidate point is computed along the normal, and the quality of fit of each candidate location is computed using (12). The point with highest quality of fit (minimum Mahalanobis distance) will be the new position of the j^{th} model point.

2.3. VT contour detection and tracking

In the airway boundary tracking problem, an initial estimate of the vocal tract contour is essential to fit the trained model to a new image. For each MRI video to be segmented, we initialize the lower and upper contours at the first frame and at the upcoming frames an initial contour is predicted using the optical flow from the estimated contour at the previous frame, and segmentation is performed automatically.

2.3.1. The proposed tracking algorithm

To estimate the VT airway-tissue boundaries, the initial prediction of the contours are evolved iteratively by minimizing an energy function, as defined in (13), at each frame. At each iteration dynamic programming (DP) is used to find a set of vertices, which are globally minimizing the defined energy function over the contour. To have smoother curves, we interpolate the snake at each iteration with a cubic spline and sample equidistant points on the interpolated curve to represent the smoothed contour. The contour is then transfered to the trained shape model space to check the range of model parameters.

The proposed contour tracking algorithm is summarized in the following steps:

1. DP to minimize the energy function.
2. Interpolation and smoothing.
3. Aligning the contour to the mean shape model.
4. Finding shape model parameter (b) and removing parameters beyond the allowed range.
5. Un-aligning the contour to the frame scale.
6. If not converged go to step 1. If converged go to next step.
7. Predict next frame contour using optical flow and go to step 1.

2.3.2. The energy function

The curve evolution is executed by minimizing an energy defined as

$$F(V) = \sum_{j=1}^{n}(E_{mah}(v_j) + \alpha E_{con}(v_j) + \gamma E_{edge}(v_j)... \quad (13)$$
$$+\delta E_{mot}(v_j)),$$

where $V = (v_1, v_2, ..., v_n)$ is a set of points minimizing the function $F()$, and E_{con} and E_{edge} are respectively the traditional snake continuity and external energy defined as

$$E_{con}(v_j) = |d - ||v_{j+1} - v_j|||^2 \quad (14)$$
$$E_{edge}(v_j) = -|\nabla I(x, y)|^2. \quad (15)$$

Here $d = \frac{1}{n}\sum_{j=1}^{n} ||v_{j+1} - v_j||$ is the average distance between all snake points and $I(x, y)$ is the image intensity at pixel position (x, y). E_{mah} is the Mahalanobis distance energy defined in (8), which moves the curve to the maximum trained appearance model fit locations, and E_{mot} is the motion energy tending the points toward positions with a more reliable motion vector estimation. E_{mot} is defined as

$$E_{mot}(v_j) = ||v_j - \hat{v_j}||^2, \quad (16)$$

where $\hat{v_j} = (v_j + f_j) + b_j$ with f_j being the forward motion vector (from frame k to frame $k + 1$) at point v_j and b_j being the backward motion vector (from frame $k + 1$ to frame k) at point $v_j + f_j$, computed using the Lucas and Kanade method [12].

3. Experimental evaluations

In our experimental evaluations, the USC-TIMIT database [7] is used to train a shape and an appearance model for the VT lower and upper contours. The database comprises mid-sagittal MRI videos recorded at a frame rate of 23.13 frames/sec and a spatial resolution of 68×68 pixels over 20×20 cm (approximately 2.9 mm pixel width). To model the movements of the vocal tract, the tract airway-tissue boundaries are manually extracted for a male and a female speaker, from a set of training images. A set of landmark points are manually labeled on each lower and upper tract profile for N training frames as shown in Figure 1. The labeled data is used to train the shape and appearance models.

In the labeling process, a total of 1400 frames are manually segmented from four different speakers. Among them $N = 1000$ frames are used for the model training, and the remaining 400 frames are used to test performance of the proposed system.

The number of shape model parameters to explain different portions of the variance in the training set is given in Table 1. Figure 2 shows the deformation of the first four modes of variations from the mean shape for lower and upper VT contours by varying the model parameters between 3 and 2 of their standard deviation.

Table 1: *Number of shape model parameters required to cover different portions of the training set variance.*

	92%	**96%**	**98%**	**99%**
Upper-profile	6	11	16	23
Lower-profile	7	13	21	29

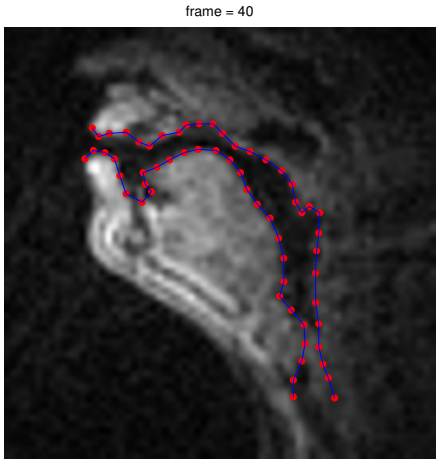

Figure 1: *Sample landmark points and the vocal tract contour from manual segmentation.*

Table 2: *Values for the parameters used in the experiments.*

Parameter	**Value**
k	8
n_s	5
α	0.6
γ	0.8
λ	0.2

3.1. Tracking results

To evaluate the performance of the proposed tracking system, a root mean squared error (RMSE) analysis is conducted for a set of Maeda's grid lines [1] over the sequence of test images. The distance from a manually detected contours to the automatically estimated ones is computed by finding the intersection points of the grid line with both contours at any grid line. The manual and automatic contours are interpolated with a cubic spline to find the intersection points with a higher accuracy.

The error analysis is performed on test image sequence of 400 frames. For this purpose, two videos containing 50 frames are selected for four speakers from the USC-TIMIT database. The selected videos contain sentences with high variations in the phones involving various constriction degree and locations of the articulators also including resting frames.

The vocal tract is divided into three sub-regions and the mean squared error is evaluated in each sub-region. Following [6], the sub-regions are defined in a similar way as (1) grid lines 1-19 for pharyngeal region, (2) grid lines 20-72 for velar and tongue region and (3) grid lines 73-92 for labial constriction region. The values of the parameters used in the experiments are listed in Table 2.

Figure 3 plots RMSE comparison of the proposed method with the baseline method in [6]. The RMSE results show that the proposed algorithm performs with higher accuracy, spe-

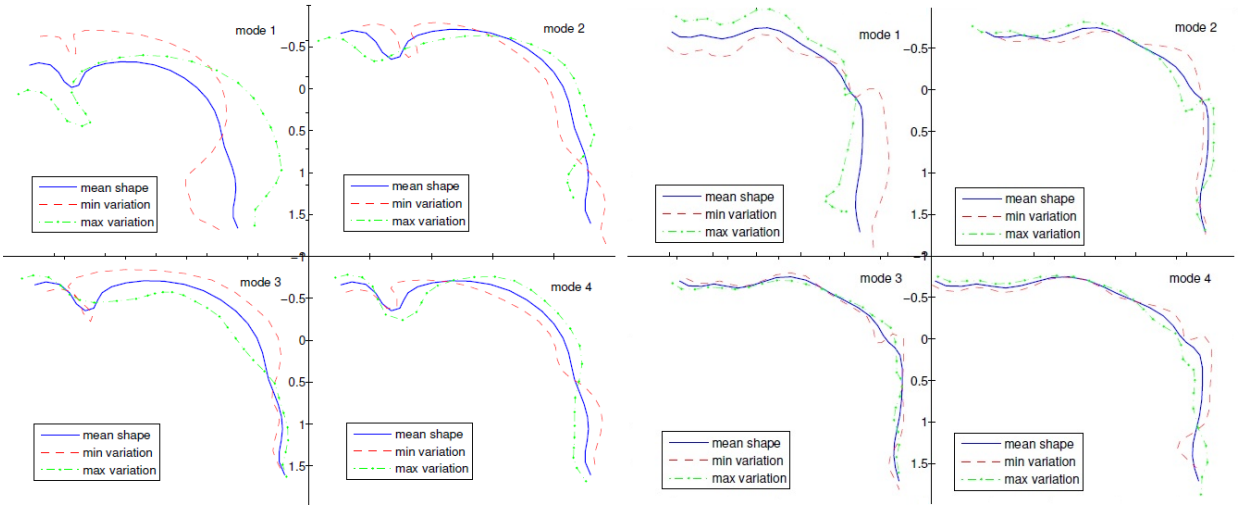

Figure 2: *Samples from the first four PCA modes of the shape model.*

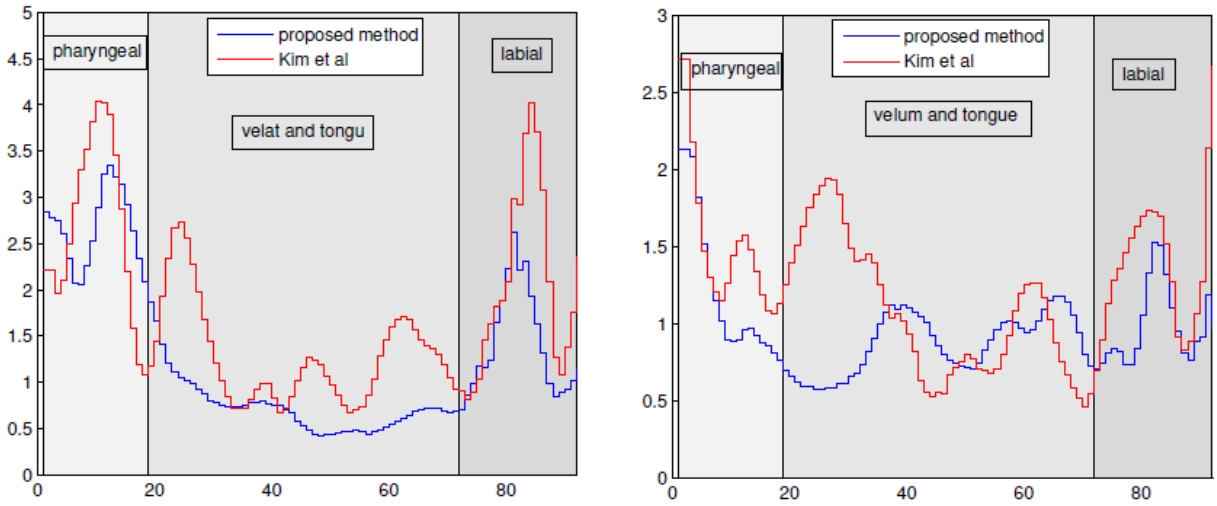

Figure 3: *Root mean squared error in the sub-region of the vocal tract.*

cially in the tongue region which is the most dynamic articulator in the vocal tract. The computed error in the pharyngeal region is almost equal for both algorithms. The complex anatomy of the epiglottis and the poor quality of MRI videos in the pharyngeal region, are believed to be the reason of poor accuracy in the epiglottis sub-region. Overall the proposed algorithm reduces the error in terms of RMSE as reported in Table 3.

4. Conclusions

In this paper we presented a robust contour tracking method applied to human vocal tract airway tissue boundary detection and tracking in a sequence of MRI images. A parametric shape model is built for the vocal tract using statistical modeling. The vocal tract is described with 21 or 16 control parameters in the shape model for lower and upper profiles respectively. The traditional active shape models are used to extract the image intensity variation around the vocal tract contours. The vocal tract contours are then detected and tracked in a sequence of images using the newly defined snake energy function. The experimental analysis shows the higher accuracy of the proposed algorithm over the baseline methods in the root mean square error metric.

Table 3: *RMSE for lower and upper boundary in different sub-regions [pixel units].*

Region	Proposed method	[6]
1-lower	2.26	2.29
1-upper	0.94	1.28
2-lower	0.61	1.57
2-upper	0.72	0.96
3-lower	1.11	1.73
3-upper	0.85	1.27
total-lower	1.06	1.58
total-upper	0.80	1.10

5. References

[1] S. Maeda, "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," *Speech Production and Modellling*, pp. 131–149, 1990.

[2] H. Yehia, K. Takeda, and F. Itakura, "An acoustically oriented vocal-tract mode," *IEICE TRANSACTIONS on Information and Systems*, vol. E79-D, no. 8, 1996.

[3] Y. Laprie and J. Busset., "Construction and evaluation of an articulatory model of the vocal tract," in *19th European Signal Processing Conference - EUSIPCO*, 2011.

[4] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance imaging," *IEEE Transaction on Medical Imaging*, vol. 28, no. 3, 2009.

[5] M. Proctor, D. Bone, N. Katsamanis, and S. Narayanan, "Rapid semi-automatic segmentation of real-time magnetic resonance image for parametric vocal tract analysis," in *Interspeech*, vol. 1, no. 4, 2010.

[6] J. Kim, N. Kumar, S. Lee, and S. Narayanan, "Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data," in *Tenth international seminar on Speech Production, ISSP10*, 2014.

[7] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)." *The Journal of the Acoustical Society of America*, vol. 136, no. 3, p. 1307, 2014.

[8] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *international journal of computer vision*, pp. 321–331, 1988.

[9] T. Cootes and C. Taylor, "Active shape model search using local grey-level methods: a quantitative approach," *Proc. British Machine Vision Conference*, pp. 639–648, 1993.

[10] 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 *international seminar on Speech Production, ISSP*, 2006.

[11] C. Goodall, "Procrustes methods in the statistical analysis of shape," *Journal of the Royal Statistical Society B*, vol. 53, no. 2, 1991.

[12] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *in Proc. Int. Joint Conf. Artificial Intelligence*, pp. 1674–679, 1981.