Interspeech 2018
2-6 September 2018, Hyderabad

Air-Tissue Boundary Segmentation in Real-Time Magnetic Resonance Imaging Video using Semantic Segmentation with Fully Convolutional Networks

Valliappan CA, Renuka Mannem, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science, Bangalore-560012, India.
valliappanc@iisc.ac.in, mannemrenuka@iisc.ac.in, prasantg@iisc.ac.in

Abstract

In this paper, we propose a new technique for the segmentation of the Air-Tissue Boundaries (ATBs) in the vocal tract from the real-time magnetic resonance imaging (rtMRI) videos of the upper airway in the midsagittal plane. The proposed technique uses the approach of semantic segmentation using the Deep learning architecture called Fully Convolutional Networks (FCN). The architecture takes an input image and produces images of the same size with air and tissue class labels at each pixel. These output images are post processed using morphological filling and image smoothing to predict realistic ATBs. The performance of the predicted contours is evaluated using Dynamic Time Warping (DTW) distance between the manually annotated ground truth contours and the predicted contours. Four fold experiments with four subjects from USC-TIMIT corpus (with $\sim$2900 training images in every fold) demonstrate that the proposed FCN based approach has 8.87% and 9.65% lesser average error than the baseline Maeda Grid based scheme, for the lower and upper ATBs respectively. In addition, the proposed FCN based rtMRI segmentation achieves an average pixel classification accuracy of 99.05% across all subjects.

Index Terms: real-time magnetic resonance imaging, air-tissue boundary segmentation, Fully Convolutional Networks.

1. Introduction

The real-time magnetic resonance imaging (rtMRI) video of the vocal tract in the midsaggital plane during speech is an important tool for speech production research. The rtMRI has an advantage of capturing the complete vocal tract in a non-invasive manner [1] which makes it more effective than the existing methods like Electromagnetic articulograph [2], Ultrasound [3], X-Ray [4]. The rtMRI video frames provide spatio-temporal details of the speech articulators that could help in modeling speech production [5] and several speech related applications [6], it is very important to have an accurate air-tissue boundary (ATB) segmentation. For example, to develop a text-to-speech system, Toutios [7] used the estimated ATB from the rtMRI videos. Patil et al. [8] used the rtMRI data for comparing the articulatory control of beatboxers to understand the usage of articulators in achieving acoustic goals. Studies involving morphological structures of vocal tracts [9] and analysis of vocal tract movement [10] using rtMRI videos have ATB segmentation as a pre-processing step. Thus, it is very important to estimate the ATBs in the rtMRI videos before they can be used in the study of the different articulators and dynamics of the vocal tract [11–14].

The problems of ATB segmentation of rtMRI images have been addressed by several works in the past using various approaches. Several robust ATB estimation techniques have been proposed using a composite analysis grid line superimposed on each rtMRI video frame [15–18]. Several other varied approaches were proposed. For example, Lammert et al. used a region of interest (ROI) based technique [19] and a data-driven approach using pixel intensity for the ATB segmentation problem [20]. A statistical method was presented by Asadiabadi et al. using the appearance and shape model for the vocal tract [21]. A semantic edge detection based algorithm for contour prediction was proposed by Somandepalli et al. [22]. Toutios et al. [23] and Sorensen et al. [24] used factor analysis technique to estimate the compact outline of the vocal tract. Zhang et al. [25] used multi-directional Sobel operators in order to construct boundary intensity map in the rtMRI video frames. Techniques such as [15], [18], [20], [21] are advantageous over the others because of their unsupervised and semi-automatic approach. However, a more precise and reliable ATBs can be obtained in a supervised learning approach where the model learns boundary shapes from the limited training images across different subjects rather than estimating in an unsupervised manner.

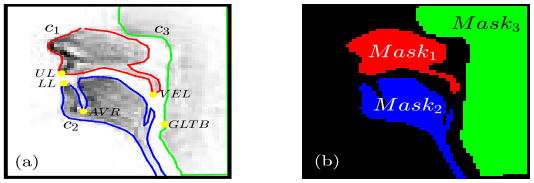

Figure 1: *(a) Illustration of the air tissue boundaries* (C_1, C_2, C_3) *(b) Closed contour polygon / Mask*

In this paper, we consider a supervised approach and propose a deep learning based semantic segmentation technique for automatic ATB segmentation both inside and outside the vocal tract from the rtMRI video. The proposed technique uses the state of the art semantic image segmentation architecture called Fully Convolutional Network (FCN) [26]. This method has several advantages over the existing techniques including its robustness to imaging artifacts and grainy noise, which could be challenging for naive low-level gradient based approaches. The areas in the upper airway in the rtMRI video frames have high pixel intensity regions (corresponding to tissue region) as well as low pixel intensity regions (corresponding to airway cavity region in the vocal tract). Hence estimating ATBs from rtMRI images can be visualized as a problem of finding the boundary that separates high pixel intensity region from low pixel intensity region. The FCN architecture is trained to learn such intensity difference in the rtMRI frames. We propose several post-processing steps to apply on the FCN output to predict smooth and realistic ATBs.

The performance of the proposed technique is evaluated using Dynamic Time Warping (DTW) distance between the ground truth contours (manually annotated) and predicted contours. In addition to the DTW distance, the pixel classification accuracy is also provided to supplement the semantic seg-

10.21437/Interspeech.2018-1939

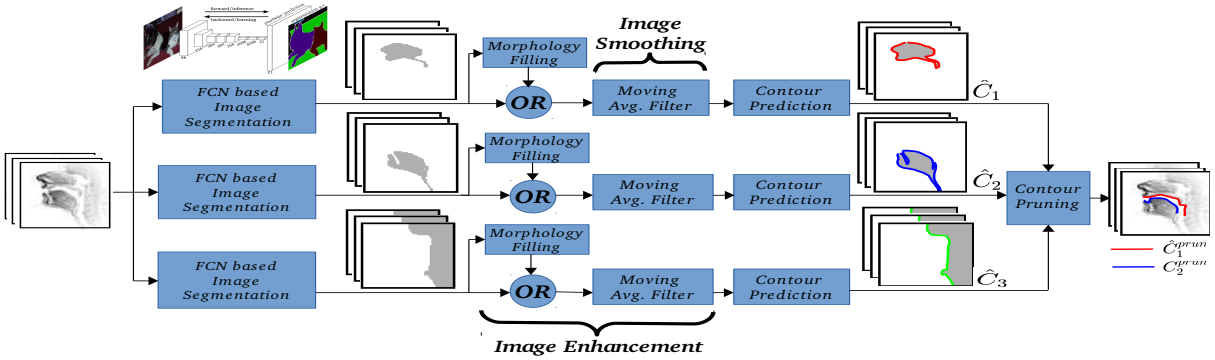

Figure 2: *Illustration of the steps in the proposed FCN based approach*

mentation based approach for ATBs segmentation. The DTW distance is compared with a Maeda grid (MG) based baseline scheme proposed by Kim et al. [15]. The MG scheme uses a grid based approach to estimate the ATBs in the rtMRI video frames. The DTW distance using the FCN based technique is found to be 8.87% and 9.65% lesser than that using the baseline MG scheme, for the lower and upper ATBs respectively. In addition, the FCN based ATB segmentation has an average pixel classification accuracy of 99.05% across all four subjects from USC-TIMIT corpus used in this work. This clearly implies that the proposed method performs better than the baseline MG scheme. The FCN based technique also estimates the ATBs outside the vocal tract unlike the earlier unsupervised approaches, thereby providing a more detailed and intricate depiction of the boundaries in the rtMRI frames.

2. Dataset

In this work, we use USC-TIMIT [27] corpus, a rich database of the rtMRI videos of upper airway in the midsagittal plane. The database consists of five female and five male subjects speaking 460 sentences from MOCHA-TIMIT [28] database. Each frame has a spatial resolution of 68×68 $(2.9mm \times 2.9mm)$ and the video was recorded at 23.18 frames/sec. For this work we chose to work on 16 rtMRI videos (one for each sentence) from each of two female (F1, F2) and two male (M1, M2) subjects. The selected 16 videos have $1462, 1270, 1642, 1399$ image frames from subjects F1, F2, M1, M2 respectively.

A MATLAB-GUI was used for manual annotation of three major contours representing the complete ATB in rtMRI images as shown in Figure 1 [29]. Along with the contours, upper lip (UL), lower lip (LL), tongue base (AVR), velum tip (VEL) and glottis begin (GLTB) were also marked for each rt-MRI frame using the GUI. As shown in Figure 1, the contour-1 $(C1)$ is a closed contour starting from upper lip (UL); it runs through the hard palate and joins the velum (VEL) and goes around the fixed nasal tract. Contour-2 $(C2)$ is a single closed contour which covers the jawline, lower lip (LL), tongue blade and extends below the epiglottis. Contour $(C3)$ marks the pharyngeal wall. In order to train the proposed FCN we make sure that all three contours are closed.

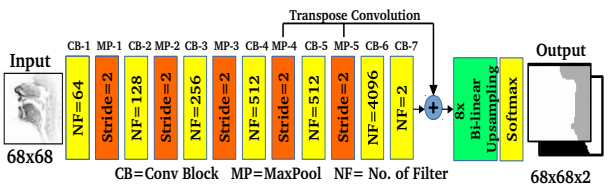

Figure 3: *FCN architecture used in this work*

3. Proposed FCN based segmentation

The FCN based semantic segmentation approach for ATB segmentation in the rtMRI video frames is explained using a block diagram shown in Figure 2. The rtMRI images are passed through FCN based semantic segmentation step for each of the three contours separately. For every contour, this step generates a binary image where one class corresponds to pixels within the contour and the second class for outside the contour. These three binary images in each video frame are passed through post processing stages followed by contour prediction step. The post processing stages include morphological filling and moving average filtering. Three complete predicted contours in every rtMRI video frame are pruned to obtain the ATB within the vocal tract. The detailed description for each block is given below.

3.1. FCN based Image Segmentation

The understanding of an image at pixel level by assigning a class label to each pixel is called as semantic segmentation. For example, labeling the pixels of an image based on the object class at that particular pixel location. ATB segmentation of an rtMRI video frame is equivalent to labeling each pixel as belonging to tissue class or air-cavity class. Hence, we can use semantic segmentation to understand the rtMRI image at the pixel level. In this work, we incorporate FCN [26] based semantic segmentation over U-Net [30] and SegNet [31] due to their state-of-the-art performance and the memory versus accuracy trade-off involved in achieving good segmentation performance in semantic segmentation. The idea behind using FCN is to preserve the spatial information, that is the network takes input of size 68×68 and produces output image of same size. The state of the art performance was achieved for the semantic segmentation task by using VGG-16 architecture. Hence we incorporate a similar architecture for our segmentation task as shown in Figure 3. In our experiment, we have only two class labels: 1) region inside of the contour polygon is labeled as class-1 and 2) region outside of the contour polygon is labeled as class-0. We train three different FCNs as shown in Figure 2, for each contour polygon as defined in section-2 and as depicted using three masks $Mask_1$, $Mask_2$ and $Mask_3$ in Figure 1(b). Each FCN outputs a binary mask image of 1s at all the pixel locations inside the contour polygon and 0s elsewhere.

3.2. Image Enhancement

Image enhancement consists of post processing steps on the binary mask image obtained from the FCN based image segmentation block. Image enhancement involves two steps: 1) Image

filling - to obtain uniform class value inside the contour polygon 2) Smoothing - to remove sharp edges in the binary mask image obtained from FCN. In this work, image filling comprises two functions: a) morphological dilation/filling [32] and b) Logical OR between the FCN output and the morphologically dilated image as shown in Fig 2. The morphological filling operation requires a structural element. In this work, we selected a disk shaped structural element of radius 5-pixels, decided based on its performance on the development set. This dilation operation at a pixel location is controlled by the surrounding pixels (in the shape of a disk). Hence, this operation fills the holes inside contour polygon but it also dilates the boundary regions, ultimately corrupting the shape of the binary mask. In order to preserve the shape, the Logical OR operation, after morphological filling, is done. Figure 4(b), 4(e) show the output for $Mask_1$ and $Mask_2$ respectively after image filling (step-1). Following

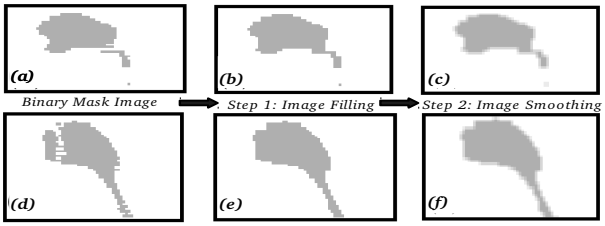

Figure 4: *(a), (b), (c): Illustration of the steps in Image enhancement of $Mask_1$ and (d), (e), (f): Illustration of the steps in Image enhancement of $Mask_2$*

image filling, in order to obtain smooth and realistic contours, we aim to remove sharp edges in the morphologically filled binary mask image. Hence the final post processing step involves image smoothing which is done using moving average filter of size 2×2. Figure 4(c), 4(f) show the output of moving average filtering on $Mask_1$ and $Mask_2$ respectively.

3.3. Contour Prediction

The contour prediction for all three post-processed FCN outputs works based on the principle of edge detection. In this work, we use the canny edge detection algorithm. This section concentrates on predicting a contour that connects the boundary edge points to form a closed contour. In order to predict a closed contour that fits all the edge points with a minimum area enclosed we use concave hull algorithm for 2-dimension [33], which works based on N nearest neighbors. The threshold N in the concave hull algorithm describes the smoothness level of the computed hull on the edge points. In this work, we fix the parameter N to be 3, decided based on its performance on the development set. Figure 5(a) illustrates the full contour prediction on $Mask_1$, $Mask_3$ and Figure 5(c) illustrates the full contour prediction on $Mask_2$. The predicted contours corresponding to three masks are denoted by $\hat{C}_1, \hat{C}_2, \hat{C}_3$.

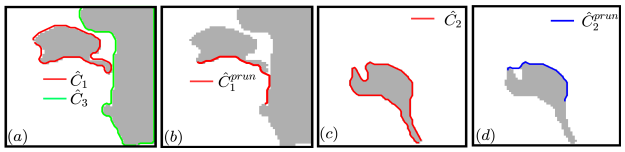

Figure 5: *(a) Illustration of the contour prediction for $Mask_1$, $Mask_3$ (b) Illustration of the contour pruning for upper ATB (c) Illustration of the contour prediction for $Mask_2$ (d) Illustration of the contour pruning for lower ATB*

3.4. Contour Pruning

The predicted ATBs cover regions both inside and outside the vocal tract as illustrated in section 3.3. In order to obtain the ATBs within the vocal tract, we follow different procedures for upper ($\hat{C}_1$) and lower ($\hat{C}_2$) ATBs. For pruning the upper contour ($\hat{C}_1$), at first, the VEL point is spotted using the inflection point in contour $\hat{C}_1$. Then, $\hat{C}_1$ is segmented from UL point till the VEL tip to form $\hat{C}_{11}$. Likewise, $\hat{C}_3$ is also segmented from point closest to VEL till GLTB to form $\hat{C}_{12}$. Eventually, $\hat{C}_{11}$ and $\hat{C}_{12}$ are stitched together to form upper ATB $\hat{C}_1^{prun}$. Figure 5(b) shows the $\hat{C}_1^{prun}$ obtained from $\hat{C}_1$ and $\hat{C}_3$.

Similarly, $\hat{C}_2$ is pruned from LL to GLTB. However, the portion of contour near the tongue base (due to the presence of lower teeth) is not a part of the vocal tract. In order to obtain a smooth contour near the tongue base, contour smoothing technique is followed. Firstly, we segment the $\hat{C}_2$ starting from the point $\hat{C}_2^{start}$ with lower row index (typically close to LL) to point $\hat{C}_2^{end}$ which is identified as the next point with the similar row index. Lets denote this segment (around the AVR point) of length P as $C_{seg} = \{(x_i, y_i), 1 \leq i \leq P\}$. This C_{seg} is replaced with $C_{sm} = \{(x_i, y_i^{sm}), 1 \leq i \leq P\}$, where $y_i^{sm} \triangleq b_0 + b_1 x_i + b_2 (x_i)^2$, ensuring a smooth contour near the AVR. The coefficients of the polynomial are obtained by minimizing the MSE between the data points and the polynomial function. After replacing the portion of C_{seg} in $\hat{C}_2$ with C_{sm}, we obtain the lower ATB $\hat{C}_2^{prun}$. Figure 5(d) shows $\hat{C}_2^{prun}$ after pruning and smoothing $\hat{C}_2$.

4. Experiments and Results

4.1. Experimental Setup

In this work, for the estimation of ATBs we consider 16 videos of rtMRI data from 4-subjects (F1, F2, M1, M2). The FCN model was trained and evaluated using 4-fold cross validation by choosing the videos in a round robin fashion. In each fold, the test set consists of four videos from each of the four subjects, i.e., a total of 16 videos. The FCN model is trained with eight videos from each subject resulting a total of 32 videos. Remaining sixteen videos, four from each subject, are used as the development set. As described in section 3.1, each FCN model is trained for a particular contour. Each fold on an average consists of ∼2900 training images, ∼1443 images in both development and test sets. The FCN model is trained for a maximum of 120 epochs with early stopping condition imposed based on the validation loss. In this work, we use the development set for selecting 1) the structural element and its size for image filling, 2) size of moving average filter, 3) parameter N in contour prediction (sec-3.3).

4.2. Evaluation metric

For evaluation, we use two metrics 1) DTW Distance and 2) pixel accuracy. DTW distance measures the closeness of the estimated contour to the ground truth contour [34]. In addition to DTW distance, we also provide the pixel accuracy score, similar to evaluating performance of a semantic segmentation method [26], [30].

1) DTW distance: Lets denote the ground truth contour C_g of length M_g as $C_g(i) = \{(x_i^g, y_i^g) | 1 \leq i \leq M_g\}$ and predicted contour C_p of length M_p as $C_p(i) = \{(x_i^p, y_i^p) | 1 \leq i \leq M_p\}$.

The DTW distance between C_g and C_P is defined as:

$$DTW(C_g, C_p) = \underset{\substack{1\leq o_g^{'}\leq M_g \\ 1\leq o_p^{'}\leq M_p}}{\arg\min} \frac{1}{L}\sum_{l=1}^{L} ||C_g(o_g^{'}(l)) - C_p(o_p^{'}(l))||_2 \quad (1)$$

The DTW scores have a unit of pixel. Smaller value of DTW(C_g, C_p) indicates more similarity between C_g and C_p, which means that they are located closely. The DTW scores are reported separately for 1) The pruned contours $\hat{C_1}^{prun}, \hat{C_2}^{prun}$ where MG based approach [15] is used as a baseline 2) full contours $(\hat{C_1}, \hat{C_2}, \hat{C_3})$. In order to carry out the evaluation similar to the baseline MG, we prune the ground truth following the steps outlined in section 3.4. The pruned ground truth contours are denoted as (C_1^{prun}, C_2^{prun}).

2) Pixel accuracy: Let p_{ij} be the number of pixel of class i predicted to class j and T_i is total number of pixels in class i, i.e., $T_i = \sum_j p_{ij}$, where $i, j \in \{0, 1\}$. In order to show the performance of the FCN model used for the semantic segmentation we provide the $Pixel\_accuracy$ corresponding to each mask which is defined as $\frac{\sum_i p_{ii}}{\sum_i T_i}$.

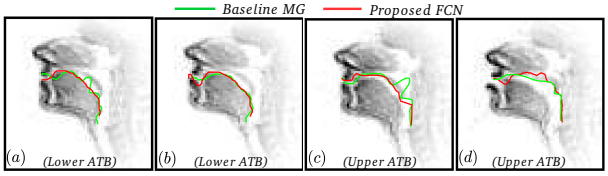

Figure 6: *Illustration of the ATBs within the vocal tract using using MG and FCN schemes*

4.3. Results and Discussion

The mean ± standard deviation of DTW$(C_2^{prun}, \hat{C_2}^{prun})$ and DTW$(C_1^{prun}, \hat{C_1}^{prun})$ are shown in Table 1 for the baseline method MG and the proposed FCN based method. The DTW distance, when averaged across subjects is found to be 8.87% lesser for the lower ATBs and 9.65% lesser for the upper ATBs, when the FCN method is compared to the baseline scheme.

	Lower ATB $(\hat{C_2}^{prun})$		Upper ATB $(\hat{C_1}^{prun})$	
SUB	MG	FCN	MG	FCN
F1	1.21±0.21	1.00±0.25	1.02±0.19	0.91±0.21
F2	1.28±0.27	1.13±0.31	1.24±0.29	1.08±0.19
M1	1.26±0.60	1.17±0.25	1.10±0.20	1.02±0.20
M2	1.35±0.30	1.21±0.23	1.19±0.24	1.09±0.21
Average	1.24±0.35	1.13±0.26	1.14±0.23	1.03±0.20

Table 1: *DTW distance of the predicted ATBs within the vocal tract.*

Figure 6(a) and Figure 6(c) illustrate example where the proposed FCN scheme performs better compared to the MG method. The improved performance compared to the baseline MG can be associated with the following reasons: 1) spatial characteristics captured in the semantic segmentation approach 2) individual-FCN model for each contour and 3) image enhancement techniques, applied as the post processing step. These altogether prevent the proposed method from estimating jagged contours. The Figure 6(b) and Figure 6(d) illustrate examples where the predicted contours using FCN are not as precise as the ones obtained using MG scheme. The poor performance of FCN based method could be due to the low resolution of the image. Hence the model can not differentiate the contact region of velum and tongue dorsal well in the case of Figure 6(b). Similarly, in Figure 6(d) the model can not precisely locate the velum and pharyngeal wall.

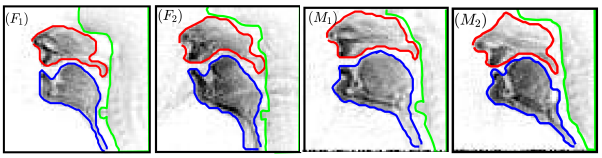

Figure 7: *Illustration of the complete ATBs predicted using the FCN based method for four subjects.*

SUB	C_1	C_2	C_3
F1	0.89±0.11	1.05±0.19	0.83±0.11
F2	1.02±0.17	1.12±0.24	0.80±0.10
M1	1.03±0.21	1.37±0.35	0.80±0.09
M2	0.98±0.09	1.01±0.17	0.85±0.10

Table 2: *DTW distance of the full contours*

In addition to the pruned ATBs, we also present the complete predicted contours for each subject $(\hat{C_1}, \hat{C_2}, \hat{C_3})$ as shown in Figure 7. We supplement these examples with mean ± standard deviation DTW distance between the predicted and the ground truth complete contours in Table 2 and mean $Pixel\_accuracy$ corresponding to each mask in Table 3.

SUB	$Mask_1$	$Mask_2$	$Mask_3$
F1	99.39	98.34	99.73
F2	99.20	98.14	99.75
M1	99.28	97.97	99.75
M2	99.32	98.09	99.70

Table 3: *Average Pixel accuracy (in %) for different mask*

On an average $\sim$ 1% pixels are being misclassified. This 1% accounts for 49 pixels out of the total 4624 (68×68) pixels in the image. These misclassified pixels predominantly lie in the boundary region where the model cannot differentiate the contact between the upper and lower ATB. This is mainly because of the low resolution of the image. To an extent, the image enhancement step helps to correct the misclassified pixels, but this step may also lead to a shift in the boundaries' position, ultimately predicting an un-reliable contour. Hence, the error corresponding to the proposed FCN based method (Table 1 and Table 2) can be strongly associated with pixel accuracy of the FCN model. Even though the proposed technique has minor disadvantages, the results clearly show that the proposed FCN based method predicts a reliable contour both inside and outside the vocal tract.

5. Conclusions

In this paper, we proposed a Deep learning based semantic segmentation approach for prediction of the Air tissue boundaries in the midsagittal rtMRI video frames. The proposed FCN model learns the shapes across all the subjects from the training data. The robust performance of the proposed method is associated to the model trained for individual contour prediction which avoids uncertainty in semantic segmentation/contour prediction. Future works include developing completely automated boundary detection schemes by modifying the FCN architecture.

6. Acknowledgements

The authors thank Pratiksha Trust for their support. Special credits to Anisha Banerjee, Vijitha Periyasamy and Advait Koparkar for their vital contribution in manual annotation of the rtMRI video frames.

7. References

[1] E. Bresch, Y. C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," in *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 123–132, May 2008.

[2] D. Maurer, B. Grne, T. Landis, G. Hoch, and P. W. Schnle, "Reexamination of the relation between the vocal tract and the vowel sound with electromagnetic articulography in vocalizations," in *Clinical Linguistics & Phonetics*, vol. 7, no. 2, pp. 129–143, 1993.

[3] K. L. Watkin and J. M. Rubin, "Pseudothreedimensional reconstruction of ultrasonic images of the tongue," in *The Journal of the Acoustical Society of America*, vol. 85, no. 1, pp. 496–499, 1989.

[4] D. C. Wold, "Generation of vocaltract shapes from formant frequencies," in *The Journal of the Acoustical Society of America*, vol. 78, no. S1, pp. S54–S55, 1985.

[5] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," in *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, 2004.

[6] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," in *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 537–554, 1996.

[7] A. Toutios, T. Sorensen, K. Somandapelli, R. Alexander, and S. S. Narayanan, "Articulatory synthesis based on real-time magnetic resonance imaging data," in *Interspeech*, 2016.

[8] N. Patil, T. Greer, R. Blaylock, and S. Narayanan, "Comparison of basic beatboxing articulations between expert and novice artists using real-time magnetic resonance imaging," in *Interspeech*, 2017.

[9] A. Lammert, M. Proctor, and S. Narayanan, "Interspeaker variability in hard palate morphology and vowel production," in *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 6, pp. 1924–1933, 2013.

[10] V. Ramanarayanan, L. Goldstein, D. Byrd, and S. S. Narayanan, "An investigation of articulatory setting using real-time magnetic resonance imaging," in *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 510–519, 2013.

[11] B. Parrell and S. Narayanan, "Interaction between general prosodic factors and languagespecific articulatory patterns underlies divergent outcomes of coronal stop reduction," in *International Seminar on Speech Production (ISSP) Cologne, Germany*, 2014.

[12] F.-Y. Hsieh, L. Goldstein, D. Byrd, and S. Narayanan, "Pharyngeal constriction in english diphthong production," in *Interspeech*, vol. 19, no. 1, pp. 968–972, 2013.

[13] A. Prasad, V. Periyasamy, and P. K. Ghosh, "Estimation of the invariant and variant characteristics in speech articulation and its application to speaker identification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4265–4269, 2015.

[14] M. Li, J. Kim, A. Lammert, P. K. Ghosh, V. Ramanarayanan, and S. Narayanan, "Speaker verification based on the fusion of speech acoustics and inverted articulatory signals," in *Computer Speech and Language*, vol. 36, pp. 196 – 211, 2016.

[15] J. Kim, N. Kumar, S. Lee, and S. S. Narayanan, "Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data," in *International Seminar on Speech Production (ISSP)*, pp. 222 – 225, 2014.

[16] S. E. Öhman, "Numerical model of coarticulation," in *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 310–320, 1967.

[17] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modelling*. Springer, pp. 131–149, 1990.

[18] M. I. Proctor, D. Bone, A. Katsamanis, and S. S. Narayanan, "Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis," in *Proceedings of the International Conference on Speech Communication and Technology*, 2010.

[19] A. C. Lammert, V. Ramanarayanan, M. I. Proctor, S. Narayanan *et al.*, "Vocal tract cross-distance estimation from real-time mri using region-of-interest analysis." in *INTERSPEECH*, pp. 959–962, 2013.

[20] A. C. Lammert, M. I. Proctor, and S. S. Narayanan, "Data-driven analysis of realtime vocal tract mri using correlated image regions," in *Interspeech*, 2010.

[21] S. Asadiabadi and E. Erzin, "Vocal tract airway tissue boundary tracking for rtmri using shape and appearance priors," in *Interspeech*, pp. 636–640, 2017.

[22] K. Somandepalli, A. Toutios, and S. S. Narayanan, "Semantic edge detection for tracking vocal tract air-tissue boundaries in real-time magnetic resonance images," *in Interspeech 2017*, pp. 631–635, 2017.

[23] A. Toutios and S. S. Narayanan, "Factor analysis of vocaltract outlines derived from real-time magnetic resonance imaging data," in *International Congress of Phonetic Sciences (ICPhS), Glasgow, UK*, 2015.

[24] T. Sorensen, A. Toutios, L. Goldstein, and S. Narayanan, "Characterizing vocal tract dynamics with real-time mri," in *15th Conference on Laboratory Phonology, Ithaca, NY*, 2016.

[25] D. Zhang, M. Yang, J. Tao, Y. Wang, B. Liu, and D. Bukhari, "Extraction of tongue contour in real-time magnetic resonance imaging sequences," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 937–941, 2016.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[27] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)," in *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

[28] A. A. Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *5th Seminar of Speech Production*, pp. 305–308, 2000.

[29] A. K. Pattem, A. Illa, A. Afshan, and P. K. Ghosh, "Optimal sensor placement in electromagnetic articulography recording for speech production study," in *Computer Speech & Language*, vol. 47, pp. 157–174, 2018.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, 2015.

[31] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," in *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[32] R. Van Den Boomgaard and R. Van Balen, "Methods for fast morphological image transforms using bitmapped binary images," in *CVGIP: Graphical Models and Image Processing*, vol. 54, no. 3, pp. 252–258, 1992.

[33] J.-S. Park and S.-J. Oh, "A new concave hull algorithm and concaveness measure for n-dimensional datasets," in *Journal of Information science and engineering*, vol. 28, no. 3, pp. 587–600, 2012.

[34] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *The 3rd International Conference on Knowledge Discovery and Data Mining*, pp. 359–370, 1994.