AN IMPROVED AIR TISSUE BOUNDARY SEGMENTATION TECHNIQUE FOR REAL TIME MAGNETIC RESONANCE IMAGING VIDEO USING SEGNET

Valliappan CA[1]*, Avinash Kumar*[2]*, Renuka Mannem*[1]*, Karthik GR*[1]*, Prasanta Kumar Ghosh*[1*]

[1]Electrical Engineering Department, Indian Institute of Science, Bangalore.
[2]Electrical and Electronics Engineering, National Institute of Technology, Surathkal.
valliappanc@iisc.ac.in, mannemrenuka@iisc.ac.in, prasantg@iisc.ac.in

ABSTRACT

This paper presents an improved methodology for the segmentation of the Air-Tissue boundaries (ATBs) in the upper airway of the human vocal tract using Real-Time Magnetic Resonance Imaging (rtMRI) videos. Semantic segmentation is deployed in the proposed approach using a Deep learning architecture called SegNet. The network processes an input image to produce a binary output image of the same dimensions having classified each pixel as air cavity or tissue, following which contours are predicted. A Multi-dimensional least square smoothing technique is applied to smoothen the contours. To quantify the precision of predicted contours, Dynamic Time Warping (DTW) distance is calculated between the predicted contours and the manually annotated ground truth contour. Four fold experiments are conducted with four subjects from the USC-TIMIT corpus, which demonstrates that the proposed approach achieves a lower DTW distance of 1.02 and 1.09 for the upper and lower ATB compared to the best baseline scheme. The proposed SegNet based approach has an average pixel classification accuracy of 99.3% across all the subjects with only 2 rtMRI videos ($\sim$180 frames) per subject for training.

Index Terms— Real-Time Magnetic Resonance Imaging, Air-Tissue Boundary Segmentation, SegNet.

1. INTRODUCTION

The advancements in the field of speech have inspired researchers to understand the speech production using various tools like Electromagnetic articulograph [1], Ultrasound [2], X-Ray [3] and real-time magnetic resonance imaging (rtMRI) are used. The rtMRI has an edge over the other techniques in capturing the complete vocal tract (in the midsagittal plane) in a non-invasive manner [4], making it an indispensable tool for studying speech production mechanism. The rtMRI videos provide the temporal information of speech articulators that help in understanding the mechanism of speech production [5], creating augmented video for spoken language training [6] and several other applications related to speech [7]. However, in order to accurately estimate the positions of different articulators or implement a realistic augmented video, it is essential to precisely estimate the ATBs. Studies that deal with rtMRI videos to understand the morphological shape of the vocal tract [8] and vocal tract movements [9] use the ATB segmentation as a pre-processing step. Work by Toutis et al. [10], on text-to-speech using the rtMRI videos, uses the estimated ATBs for articulatory synthesis. Similarly, Patil et al.[11] used the ATBs to compare the articulatory controls of beatboxers to understand the uses of articulators in achieving acoustic goals. Hence, estimating precise ATBs from the rtMRI video is very essential for the study of vocal tract dynamics and different articulators [12, 13, 14, 15].

Many research works in the past have addressed the problems of ATB segmentation from rtMRI videos using different techniques. These include the use of composite analysis grid-line superimposition on each rtMRI frame [16, 17, 18], the pixel intensity based data driven approach [19], region of interest (ROI) based technique [20], statistical methods that use the shape and appearance model of vocal tract [20], factor analysis to predict the compact outline of the vocal tract [21, 22] and contrast measure based Fisher discriminant method (FDM) [23]. Somandepalli et al. [24] initiated the use of deep learning based semantic ATB prediction. This was followed by the use of fully convolutional network (FCN) by Valliappan et al. [25]. Though, FCN and FDM based approaches perform better than their counterparts [16, 18, 19, 26] (unsupervised algorithms), they use a large number of training frames to learn and predict reliable ATBs. Hence, in this paper, we propose an approach that can estimate precise ATBs using a fewer number of training frames.

The problem of ATB segmentation is to differentiate the high intensity pixels (tissue region) from the low intensity pixels (airway cavity) by estimating a precise boundary. The proposed technique uses an encoder-decoder type image segmentation architecture called SegNet [27]. The SegNet architecture is better than the existing FDM and FCN approaches due to the following reasons 1) deep decoder convolution layers of SegNet helps the network reconstruct better masks for precise contour prediction and 2) fewer number of frames used for training. Unlike FCN [25], the proposed SegNet ar-

*Authors thank Pratiksha Trust for their support.

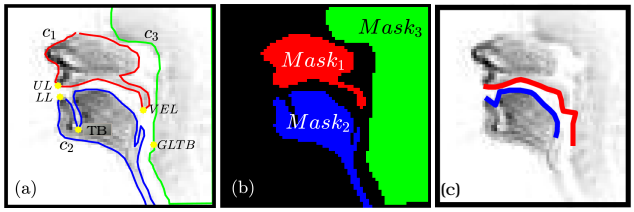

Fig. 1. (a) Illustration of the 3-major contours (C_1, C_2, C_3) (b) Closed contour Mask [25], (c) Upper and lower contour within vocal tract

chitecture does not require any complex post processing steps. SegNet is trained to differentiate the airway cavity from the tissue region in the rtMRI images. Ultimately, the aim of this work is to obtain precise ATBs within the vocal tract as shown in Fig. 1(c).

The performance of the proposed technique is evaluated using two metrics 1) Pixel classification accuracy and 2) DTW distance. The pixel classification accuracy, indicates the correctness with which the network classifies the pixels as belonging to the tissue region or the cavity region. The DTW distance is used to measure the closeness of the predicted contours with the ground truth contour obtained via manual annotation. The SegNet based method is capable of achieving an average pixel accuracy of 99.3% with only 2 training videos ($\sim$180 frames) per subject. This is $\sim$0.25% more than that of FCN which, however, was trained using 8 videos, indicating a significant reduction in the number of videos needed for training without compromising on the pixel accuracy. The DTW distance of the upper and lower ATB is compared with the Maeda grid (MG) [16], semantic segmentation network (FCN) [25] and Fisher Discriminant method (FDM) [23]. The DTW distance for the lower contour, for the SegNet is 3.5%, 2.8% and 13.8% less than that using FDM, FCN and MG respectively. Similarly, for the upper contour the DTW distance for SegNet is 5.6%, 0.7% and 10.7% less than that using FDM, FCN and MG respectively. In addition, this technique also estimates the ATBs outside the vocal tract.

2. DATASET

The USC-TIMIT [28] corpus is rich in rtMRI videos collected from five male and five female subjects. These rtMRI videos captured the upper airway in the midsagittal plane of the subjects speaking the 460 MOCHA-TIMIT [29] sentences. In this work, we considered a small subset of the USC-TIMIT corpus, that consists of two male (M1, M2) and two female (F1, F2) with 16 videos per subject. The total number of frames for each of the subjects M1, M2, F1 and F2 are 1642, 1399, 1270 and 1462 respectively. The video was recorded at 23.18 fps with a spacial resolution of 68×68 pixels with individual pixel dimension of $2.9mm \times 2.9mm$.

In order to train a network that estimates the ATBs from the rtMRI frames we need to have the ground truth ATBs. To obtain the ground truth ATBs, manual annotations were done using a MATLAB-GUI [30]. The manual annotation includes three major contours (C_1, C_2, C_3) representing the complete ATBs along with the five landmarks points upper lip (UL), lower lip (LL), tongue base (TB), velum tip (VEL) and glottis begin (GLTB) as shown in the Fig. 1(a). C_1 is a closed contour that connects upper lip (UL) and velum (VEL) through hard palate. Similarly, the C_2 contour runs through the jawline and connects the lower lip (LL), tongue base (TB) and extends along the tongue blade and below the epiglottis. Finally, Contour C_3 marks the pharyngeal wall. These closed contours are converted to masks as shown in Fig. 1(b), in order to train a semantic segmentation network.

3. PROPOSED SEGNET BASED SEGMENTATION

The proposed approach for ATB prediction using SegNet is illustrated in Fig. 2. The rtMRI frames are used as the input to the network. A separate SegNet is trained for each of the contours (C_1, C_2, C_3) . Output in each case is a binary mask consisting of two pixel values (0, 1), pixels corresponding to class-1 belongs to the region enclosed by the contour while those belonging to class-0 correspond to the region lying outside the contour. The network is made to predict these masks on the test set, following which the contours are predicted. These predicted contours are pruned in order to obtain the ATBs within the vocal tract.

3.1. SegNet based segmentation

The representation learning of an image in which we associate pixels with a class value is called as semantic segmentation. To obtain ATBs using the proposed technique, we need to segment the image into two regions, one representing the tissue region and the other representing the air-cavity. Therefore, semantic segmentation can be employed here by assigning pixel values to these regions. In this paper, SegNet [27] is used here for segmentation of the rt-MRI video frames. The weights of the architecture are learned from the rtMRI frames. The SegNet architecture has a multiple deconvolution layers compared to FCN, which helps the network to learn better on the rtMRI video frames. The architecture of SegNet is depicted in Fig. 3. As shown in Fig. 3, the input and the output have the same dimensions. Hence the spatial information pertaining to the ATBs are preserved. Three SegNets are trained, each corresponding to one of the contours (C_1, C_2, C_3). The input in each case is an rt-MRI frame of dimensions 68×68 and the output is a contour mask corresponding to one of the contours (C_1, C_2, C_3), as depicted in Fig. 1(b). These binary masks are images which contain two classes, one is the collection of pixels inside each contour (class-1) and other lying outside the contour (class-0).

3.2. Contour prediction

In this section, edge detection is performed on the binary mask output images to predict the contours. To avoid sharp edges in the binary mask image, a moving average filter of dimension 2×2 is applied. Canny edge detector is used for edge detection. In this step, we focus on estimating a closed contour that fits in all the edge points and also occupies minimum

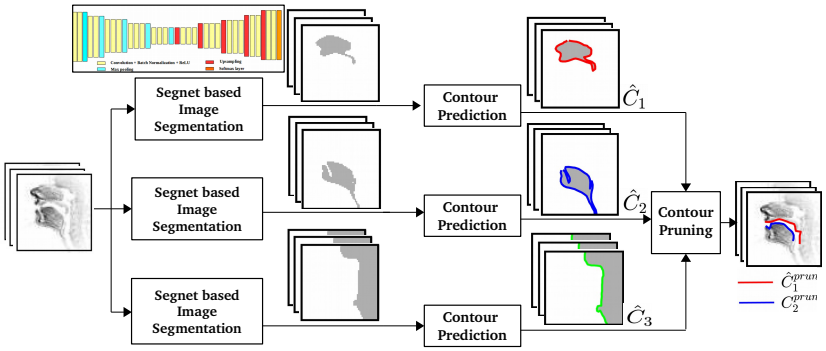

Fig. 2. Illustration the steps involved in the SegNet based ATB segmentation

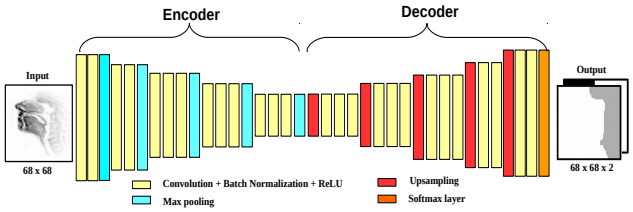

Fig. 3. SegNet architecture used in this work

area. In order to perform this, we make use of the concave hull algorithm for 2-dimension [31] which works on finding the K nearest neighbours. K describes the smoothness of the hull which is detected on the edges. Here, we choose K to be 3 based on the performance on the validation set.

3.3. Contour-pruning

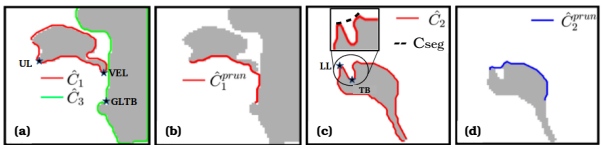

Fig. 4. (a, b) Illustrates the upper contour pruning, (c, d) Illustrates the lower contour pruning

The predicted contours have to be pruned in order to obtain the ATBs within the vocal tract as shown in Fig. 1(c). Separate procedures are followed for the pruning of the upper contour and the lower contour. The construction of the upper contour requires information from both $Mask_1$ and $Mask_3$, while the lower contour only requires information from $Mask_2$. The pruning of upper contour($\hat{C}_1$) begins by locating the velum point (VEL), which is found using the point of inflection property. Now, ($\hat{C}_1$) is segmented from the UL to the VEL. Then ($\hat{C}_3$) is segmented from the point closest to the VEL until the GLTB. Following this, the segmented portions of C_1 and C_3 are stitched together to form the pruned upper contour $\hat{C}_1^{prun}$, as shown in Fig. 4(a,b). On the other hand, the pruning of the lower contour begins by removing the tongue base (TB) ridge, as it is not a part of the vocal tract. In order to remove the TB ridge, we fit a 2^{nd} order polynomial for the set of points starting ($\hat{C}_2^s$) a point in ($\hat{C}_2$) with lower row index to the left of TB) to ($\hat{C}_2^e$) (a point with identical row index to the right of TB). The portion of ($\hat{C}_2^s$) to ($\hat{C}_2^e$) is replaced with smooth contour ($\hat{C_{seg}}$). Following this, $\hat{C}_2$ is pruned from the lower LL to the GLTB to form $\hat{C}_2^{prun}$, as shown in Fig. 4(c,d).

After pruning of the upper and the lower contours, in order to have a smooth and realistic contour, we apply the multidimensional least square smoothing using orthogonal polynomials [32].

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

In this work, we have considered two kinds of experiments: 1) for ATBs estimation and 2) for calculation of the minimum number of rtMRI videos required for the achieving saturating pixel classification accuracy (on the test videos). In the first experiment, we consider a four fold setup, having 8 videos for training, 4 videos for both testing and validation from each subject. Here, the number of rtMRI frames used for training, validation and testing are ∼720, ∼360 and ∼360 per subject respectively. In the second experiment, we train 8 models (SegNet, FCN) by varying the training set, like the i^{th} model has i training videos from across the subjects, where $i \in \{1, 2, ...8\}$. So the $model_{i+1}$ is trained with one video more than that for $model_i$. In this experiment, each video has ∼90 frames, hence i^{th} model has ∼90×i training rtMRI frames per subject. The validation and test frames for all the $model_i$ remained fixed, which was ∼360 frames per subject. The validation set is considered for 1) early stopping in order to avoiding the model over fitting, 2) shape of the moving average filter and 3) K parameter in Concave hull algorithm.

4.2. Evaluation metric

The paper uses two evaluation metric 1) DTW distance 2) Pixel classification accuracy, similar to previous technique [25]. DTW distance is used to measure the closeness between the predicted and the ground truth contour (manually annotated). The DTW score has the units of pixel. Lower the DTW score better is the prediction. The DTW scores are presented for 1) ATBs within the vocal tract (pruned contours) and 2) the complete contours (C_1, C_2, C_3). The pixel classification accuracy is one of the standard metric used in seman-

	Upper ATB				Lower ATB			
SUB	MG	FCN	$SegNet$	FDM	MG	FCN	$SegNet$	FDM
F1	1.02±0.19	0.91±0.21	**0.83±0.11**	0.94±0.17	1.21±0.21	1.00±0.25	**0.92±0.17**	0.99±0.23
F2	1.24±0.29	1.08±0.19	**0.96±0.15**	1.16±0.19	1.28±0.27	1.13±0.31	**1.12±0.29**	1.24±0.25
M1	1.10±0.20	**1.02±0.20**	1.15±0.16	1.11±0.20	1.26±0.60	1.17±0.25	**1.16±0.26**	1.17±0.26
M2	1.19±0.24	**1.09±0.21**	1.10±0.19	1.10±0.23	1.35±0.30	1.21±0.23	1.18±0.24	**1.16±0.41**
AVG:	1.13±0.22	1.02±0.20	**1.02±0.15**	1.08±0.19	1.27±0.35	1.13±0.26	**1.09±0.23**	1.14±0.29

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

	C_1		C_2		C_3	
SUB	SegNet	FCN	SegNet	FCN	SegNet	FCN
F1	**0.88**	0.89	**0.85**	1.05	**0.80**	0.83
F2	**0.98**	1.02	1.15	**1.12**	0.81	**0.80**
M1	1.03	1.03	**0.94**	1.37	**0.79**	0.80
M2	1.03	**0.89**	1.03	**1.01**	**0.83**	0.85

Table 2. DTW distance of the complete ATBs using SegNet and FCN

SUB	1	2	3	4	5	6	7	8
$Mask_1^{seg}$	88.70	**99.54**	99.53	99.57	99.54	99.54	99.55	99.57
$Mask_2^{seg}$	85.89	**98.64**	98.65	98.61	98.65	98.60	98.64	98.68
$Mask_3^{seg}$	90.30	**99.78**	99.77	99.77	99.76	99.76	99.78	99.77
$Mask_1^{fcn}$	85.68	90.89	94.47	96.09	98.14	**99.17**	99.24	99.28
$Mask_2^{fcn}$	84.12	88.14	93.88	95.51	97.77	**98.09**	98.08	98.14
$Mask_3^{fcn}$	89.45	93.45	95.80	98.80	**99.60**	99.71	99.73	99.72

Table 3. Pixel classification accuracy averaged across all subjects (on test set) for each mask vs number of training videos for SegNet, FCN. (**Bold** indicating the saturation point)

tic segmentation algorithm, which indicates the accuracy with which pixels are correctly classified [27]. For example, let P_{ij} be the number of pixels belonging to class-i but classified as class-j. T_i is the total number of pixels in class-i. $T_i = \Sigma_j P_{ij}$, where $i, j \in \{0, 1\}$. Pixel classification accuracy $= \frac{\Sigma_i P_{ii}}{\Sigma_i T_i}$.

4.3. Results

Two sets of results are presented in this work 1) using the experimental setup-1, where we compare the DTW distance of the pruned contours produced by SegNet with the MG baseline, FDM and FCN schemes and 2) using the experimental setup-2, where we present the number of training videos vs pixel classification accuracy (computed on the test videos).

Table 1 presents the DTW distance of the pruned contours $\hat{C}_1^{prun}$ and $\hat{C}_2^{prun}$ for the proposed technique (SegNet) in comparison with the other techniques, namely, MG baseline[16], FDM [23] and FCN [25]. The average DTW distance for the lower contour, using the SegNet is 3.5%, 2.8% and 13.8% lower than that using FDM, FCN and MG respectively. Similarly, for the upper contour, the DTW distance using SegNet is 5.6%, 0.7% and 10.7% less than that using FDM, FCN and MG respectively. These results are also supplemented with the DTW distance of the complete ATBs (as in Fig. 1(a)) in Table 2.

In addition to the ATBs, we also find out the minimum number of videos required for attaining the saturating pixel classification accuracy following the experimental setup-2. Table 3 presents the average pixel classification accuracy (averaged across all the subjects) on the test set vs number of training videos used to train the SegNet and FCN. From the Table 3 we observe that the pixel classification accuracy saturates beyond two training videos for SegNet. On the other hand, in FCN, it takes almost 6 videos to attain the saturating pixel accuracy. This shows that SegNet can learn patterns and shapes from fewer rtMRI compared to FCN. The FCN and SegNet has a similar encoder structure (VGG architecture), but differs in the decoding structure. Unlike FCN [25], where one single deconvolution block up-scales the encoder image many folds to form output mask, the SegNet uses multiple decoder blocks which gradually up-scales the encoded image to form an output mask. Thus capturing the details in rtMRI frames better than FCN. The proposed technique misclassifies on an average $\sim$0.70% pixels (unlike 1% for FCN) which is equivalent to having 32 pixels incorrectly classified out of $4624(68\times68)$. The reason behind the mis-classification can be associated with the low resolution of the input images. The mis-classification is predominantly along the boundary of the tissue region, specially in the frames where the there is a contact between the upper contour and the lower contour. Additionally, the manual annotations are marked with a precision of 2 decimal places, whereas the semantic segmentation based approach is capable of predicting boundary only to the precision of a pixel. This undermines the DTW distance of the predicted ATBs. Although the improvement produced by SegNet is only 0.30% (in terms of pixel classification accuracy) compared to FCN, the proposed technique can be achieved with a minimum of only two training videos per subject.

5. CONCLUSION

This paper presents a deep architecture for semantic segmentation called as SegNet for the Air tissue boundary segmentation in the rtMRI video frames. The proposed technique is capable of producing accurate contours compared to the best supervised baseline scheme. The improved performance of the model can be associated with the deep encoder-decoder architecture of SegNet. In addition, the model is capable of learning the contour from fewer number of training frames per subject. This makes the task of automatic ATBs segmentation on new subjects simple, as it requires fewer manually annotated frames to learn the boundary patterns.

6. REFERENCES

[1] D. Maurer, B. Grne, T. Landis, G. Hoch, and P. W. Schnle, "Re-examination 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.

[2] Kenneth L. Watkin and Jonathan 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.

[3] Donald 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.

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

[5] Shrikanth Narayanan, Krishna Nayak, Sungbok Lee, Abhinav Sethy, and Dani 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] Chandana S, Chiranjeevi Yarra, Ritu Aggarwal, Sanjeev Kumar Mittal, Kausthubha N K, Raseena K T, Astha Singh, and Prasanta Kumar Ghosh, "Automatic visual augmentation for concatenation based synthesized articulatory videos from real-time mri data for spoken language training," in *Proc. Interspeech*, 2018.

[7] Brad H. Story, Ingo R. Titze, and Eric 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.

[8] Adam Lammert, Michael Proctor, and Shrikanth 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.

[9] Vikram Ramanarayanan, Louis Goldstein, Dani Byrd, and Shrikanth 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.

[10] Asterios Toutios, Tanner Sorensen, Krishna Somandapelli, Rachel Alexander, and Shrikanth S. Narayanan, "Articulatory synthesis based on real-time magnetic resonance imaging data," in *Interspeech*, 2016.

[11] Nimisha Patil, Timothy Greer, Reed Blaylock, and Shrikanth Narayanan, "Comparison of basic beatboxing articulations between expert and novice artists using real-time magnetic resonance imaging," in *Interspeech*, 2017.

[12] Benjamin Parrell and Shrikanth 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.

[13] Fang-Ying Hsieh, Louis Goldstein, Dani Byrd, and Shrikanth Narayanan, "Pharyngeal constriction in english diphthong production," *in Interspeech*, vol. 19, no. 1, pp. 968–972, 2013.

[14] 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)*, 2015.

[15] Ming Li, Jangwon Kim, Adam Lammert, Prasanta Kumar Ghosh, Vikram Ramanarayanan, and Shrikanth 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.

[16] Jangwon Kim, Naveen Kumar, Sungbok Lee, and Shrikanth S. Narayanan, "Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data," in *International Seminar on Speech Production (ISSP)*, 2014.

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

[18] Michael I Proctor, Daniel Bone, Athanasios Katsamanis, and Shrikanth S Narayanan, "Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

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

[20] Adam C Lammert, Vikram Ramanarayanan, Michael I Proctor, Shrikanth Narayanan, et al., "Vocal tract cross-distance estimation from real-time mri using region-of-interest analysis.," in *INTERSPEECH*, 2013.

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

[22] Tanner Sorensen, Asterios Toutios, Louis Goldstein, and SS Narayanan, "Characterizing vocal tract dynamics with real-time mri," in *15th Conference on Laboratory Phonology, Ithaca, NY*, 2016.

[23] Advait Koparkar and Prasanta Kumar Ghosh, "A supervised air-tissue boundary segmentation technique in real-time magnetic resonance imaging video using a novel measure of contrast and dynamic programming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[24] Krishna Somandepalli, Asterios Toutios, and Shrikanth 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.

[25] Valliappan CA, Renuka Mannem, and Prasanta Kumar Ghosh, "Air-tissue boundary segmentation in real-time magnetic resonance imaging video using semantic segmentation with fully convolutional networks," in *Proc. Interspeech*, 2018.

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

[27] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *in IEEE transactions on pattern analysis and machine intelligence*, 2017.

[28] Shrikanth Narayanan, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee, Krishna Nayak, Yoon-Chul Kim, Yinghua Zhu, Louis Goldstein, et al., "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)," *The Journal of the Acoustical Society of America*, 2014.

[29] Alan A Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *5th Seminar of Speech Production*, 2000.

[30] Ashok Kumar Pattem, Aravind Illa, Amber Afshan, and Prasanta Kumar Ghosh, "Optimal sensor placement in electromagnetic articulography recording for speech production study," *in Computer Speech & Language*, vol. 47, pp. 157–174, 2018.

[31] Jin-Seo Park and Se-Jong 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.

[32] John E Kuo, Hai Wang, and Stephen Pickup, "Multidimensional least-squares smoothing using orthogonal polynomials," *Analytical chemistry*, vol. 63, no. 6, pp. 630–635, 1991.