Air-tissue boundary segmentation in real time Magnetic Resonance Imaging video using 3-D convolutional neural network

Renuka Mannem [1], *Navaneetha Gaddam* [2], *Prasanta Kumar Ghosh* [1].

[1]Electrical Engineering, Indian Institute of Science, Bangalore-560012, India.
[2]Electronics and Communications Engineering, Rajiv Gandhi University of Knowledge Technologies, Kadapa-516330, India.

mannemrenuka@iisc.ac.in, snnavaneetha95@gmail.com, prasantg@iisc.ac.in

Abstract

The real-time Magnetic Resonance Imaging (rtMRI) is often used for speech production research as it captures the complete view of the vocal tract during speech. Air-tissue boundaries (ATBs) are the contours that trace the transition between high-intensity tissue region and low-intensity airway cavity region in an rtMRI video. The ATBs are used in several speech related applications. However, the ATB segmentation is a challenging task as the rtMRI frames have low resolution and low signal-to-noise ratio. Several works have been proposed in the past for ATB segmentation. Among these, the supervised algorithms have been shown to perform well compared to the unsupervised algorithms. However, the supervised algorithms have limited generalizability towards subjects not involved in training. In this work, we propose a 3-dimensional convolutional neural network (3D-CNN) which utilizes both spatial and temporal information from the rtMRI video for accurate ATB segmentation. The 3D-CNN model captures the vocal tract dynamics in an rtMRI video independent of the morphology of the subject leading to an accurate ATB segmentation for unseen subjects. In a leave-one-subject-out experimental setup, it is observed that the proposed approach provides $\sim$32% relative improvement in the performance compared to the best (SegNet based) baseline approach.

Index Terms: real-time Magnetic Resonance Imaging video, Air-Tissue Boundary segmentation, 3-dimensional convolutional neural network, temporal information.

1. Introduction

The real-time Magnetic Resonance Imaging (rtMRI) is extensively used in speech science and linguistic studies to understand the dynamics of speech production across languages, and health conditions [1]. The rtMRI video captures the vocal tract in the midsagittal plane during speech in a safe and non-invasive manner [2]. However, the vocal tract dynamics can also be captured using X-ray [3], Electromagnetic articulography [4] and Ultrasound [5]. In particular, the rtMRI has an advantage of capturing a complete view of the vocal tract including pharyngeal structures [2]. A common pre-processing step for using rtMRI video is Air-Tissue Boundary (ATB) segmentation to identify different speech articulators in every rtMRI frame of the video. ATB segmentation has been used in numerous speech applications including text-to-speech synthesis [6], speaker identification [7]. The ATBs have been used in the studies that involve morphological structures of vocal tract and analysis of vocal tract movement [8, 9]. In [10], the ATBs were utilized to determine optimal sensor placement in electromagnetic articulography recording. Likewise, a number of speech applications [11, 12] used ATBs in the upper airway of the vocal tract. Hence, it is essential to have an accurate ATB segmentation in the rtMRI video.

The problem of ATB segmentation in an rtMRI frame has been addressed by several works in the past using various supervised and unsupervised approaches. For example, Asadiabadi et al. presented a statistical method using an appearance and shape model of the vocal tract [13]. Lammert et al. proposed a region of interest based technique [14] using pixel intensity for the ATB segmentation. A factor analysis approach was used by Toutios et al. [15] and Sorensen et al. [16] to predict the compact outline of the vocal tract. Zhang et al. [17] used multi-directional Sobel operators in order to construct boundary intensity maps in the rtMRI video frames. A robust ATB segmentation technique has also been proposed using a composite analysis grid line superimposed on each rtMRI frame [18]. The Maeda Grid (MG) [18] based technique is advantageous due to the image enhancement of the rtMRI frames. However, the accuracy of the predicted ATBs from the unsupervised approaches is limited as they consider the low-level gradients which may not always correspond to the ATB points. Hence, more accurate ATBs are predicted using the supervised approaches [19, 20, 21, 22, 23, 24, 25]. Somandepalli et al. [20] and Ashwin et al. [25] proposed semantic edge detection based techniques for tracking the vocal tract boundaries. However, these works [20, 25] formulate ATB segmentation as a 14-class classification problem where each pixel in an rtMRI image is assigned to one of the 14 classes corresponding to different articulators. In contrast to this, in our work, we generate the ATBs as series of 2D points which trace the vocal tract boundaries precisely similar to the works presented in [18, 19, 21, 22, 23, 24] as the applications involving the vocal tract boundaries [7, 26, 27] use the precise ATBs, instead of using the entire rtMRI image with pixel classification. For example, Advait et al. [19] proposed a Fisher Discriminant Measure (FDM) based approach in which the ATBs for a test rtMRI image are predicted as a combination of ATBs from the training set that maximize the FDM based objective function. Thus, the predicted ATB is not smooth and its dynamics are limited by the ATBs from the training set. Avoiding these limitations, Valliappan et al. [23] proposed a semantic segmentation based ATB prediction technique using a 2-dimensional deep convolutional encoder-decoder network (SegNet). The SegNet based approach has been shown to provide better performance compared to the works presented in [21, 22, 24] which also use 2-dimensional deep convolutional neural networks (2D-CNN). Likewise, several works have been presented in the literature for ATB segmentation.

Although, the supervised algorithms have been shown to

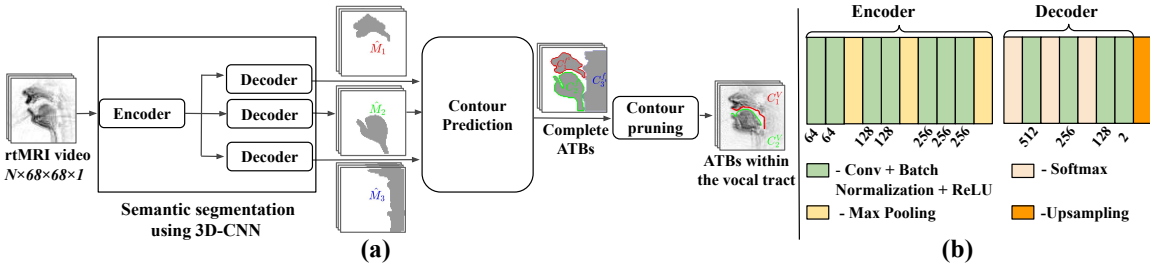

Figure 1: *(a) Illustration of the steps in the proposed 3D-CNN based approach and (b) architectures of encoder and decoders in 3D-CNN.*

provide accurate ATBs in the seen subject condition, the generalizability of these approaches for unseen subjects is a major challenge due to variability in the morphology of the subjects. In this work, we propose a 3-dimensional deep convolutional neural network (3D-CNN) for ATB segmentation which utilizes the temporal continuity of the rtMRI frames. The 3-dimensional convolutional layers have been shown to provide better performance for human action recognition [28, 29]. Thus, the 3D-CNN could help in accurate ATB segmentation for rtMRI videos which captures the smoothly varying vocal tract dynamics. The temporal continuity criterion ensures that the ATBs do not vary drastically in successive frames of an rtMRI video as the articulatory movements do not vary rapidly while a person speaks. The temporal continuity across the rtMRI video frames helps in capturing the smoothly varying vocal tract dynamics independent of the morphology of the subject and the spatial information from the rtMRI frame helps in understanding the vocal tract shape. Thus, in unseen subject condition, using both temporal and spatial information from the rtMRI video could help in better ATB prediction compared to a frame level prediction. However, to utilize both spatial and temporal information, we need to have suitable network architecture and objective function which is optimized for accurate ATB prediction. For example, the MG and FDM approaches also utilize the temporal continuity criterion in their respective objective functions. However, the accuracy of the predicted ATBs from these approaches is limited as they do not utilize the spatial information well due to the consideration of the first order pixel intensity differences. On the other hand, the 2D-CNN based ATB segmentation approaches [21, 23, 24, 20] perform frame level predictions using spatial information without utilizing the temporal continuity. However, the 3D-CNN model uses 3-dimensional convolutional filters which extract the spatial and temporal features using the given rtMRI video frames for accurate ATB prediction. In this work, we follow steps for ATB prediction similar to the ones described in [23]. We use the 3D-CNN model for semantic segmentation of the rtMRI images which are further post-processed to obtain the ATBs using contour prediction approach. In semantic segmentation, each pixel in an rtMRI image is classified into one of the pre-defined classes. In our work, we classify each pixel in an rtMRI image to tissue class or air cavity class using 3D-CNN. We use the MG, FDM and SegNet approaches as baselines. The proposed approach provides ∼32.9%, ∼61.2%, and ∼32.3% relative improvements in performance compared to the MG, FDM and SegNet approaches in a leave-one-subject-out experimental setup.

2. Dataset

In this work, we use USC-TIMIT corpus [30] which consists of rtMRI videos of the upper airway in the mid-sagittal plane. The database contains 5 female (F1, F2, F3, F4, F5) and 5 male (M1, M2, M3, M4, M5) subjects. The rtMRI videos are recorded at a

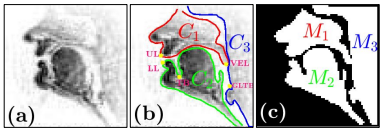

Figure 2: *Illustration of (a) an rtMRI frame (b) the respective complete ground truth air-tissue boundaries (c) and the ground truth binary mask image where the white pixels correspond to class-1 and black pixels correspond to class-0.*

frame rate of 23.18 frames/sec while a subject speaks 460 sentences from MOCHA-TIMIT database [31]. Each rtMRI frame has a spatial resolution of 68×68 (with a pixel dimension of 2.9mm×2.9mm). For the experiments in this work, 11 videos from each subject, a total of 110 videos are considered. Each subject, on an average, contains ∼974 number of frames. For the 110 videos, the manual annotation of ATB is carried out using a Matlab Graphical User Interface [10]. The manual annotation is done for three ATBs (C_1,C_2,C_3) and five points which indicate upper lip (UL), lower lip (LL), tongue base (TB), velum (VEL) and glottis begin (GLTB). As shown in Figure 2(b), C_1 is a closed contour that starts from UL, traverses through the hard palate, joins VEL and goes around the fixed nasal tract. C_2 is a closed contour that covers the jawline, LL, TB and extends below the epiglottis. C_3 contour marks the pharyngeal wall. Figure 2(b) and (c) illustrate the ground truth manually annotated ATBs and the corresponding ground truth binary mask image, respectively. In the ground truth binary mask image, the white pixels (lie inside the complete ATB) correspond to tissue with pixel value as 1 and the black pixels (lie outside the complete ATB) correspond to airway cavity with pixel value as 0. Likewise, there are three masks M_1, M_2 and M_3 corresponding to three complete ATBs C_1, C_2, and C_3 respectively.

3. Methodology

Figure 1 illustrates the steps followed in the proposed 3D-CNN based approach. The trained 3D-CNN model generates three semantically segmented images. The predicted binary masks are further used to estimate the three complete ATBs using a contour prediction approach. Then, contour pruning is done to obtain the ATBs within the vocal tract from the complete ATBs.

3.1. Semantic Segmentation using 3D-CNN

The 3D-CNN consists of one encoder and three decoders as shown in Figure 1. In SegNet based approach [23], three SegNet models are trained separately for the three binary masks which avoid overlapping of the masks in the constriction regions. This ensures that, in the contour prediction step, we will get the three precise complete ATBs corresponding to the three masks. However, due to the usage of three SegNets, the computational time and complexity increase by a factor of three. Thus, in this work, we use three decoders corresponding to the three binary masks. However, we reduce the number of layers in the three decoders

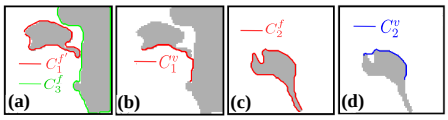

Figure 3: *Illustration of (a) contour prediction using $\hat{M}_1$ and $\hat{M}_3$ (b) contour pruning to obtain the upper ATB (C_1^v) (c) contour prediction using $\hat{M}_2$ (d) contour pruning to obtain the lower ATB (C_2^v).*

compared to the encoder to reduce the complexity of the network.

Figure 1(b) illustrates the architectures of the encoder and decoder used in the 3D-CNN model. The number of filters in each convolutional layer (which indicates the depth of the layer) is indicated below the layer. To learn the high-level features from the low-level features, in encoder, depth is doubled from the input layer to the last layer and in decoder, depth is halved in the first three layers [32, 23]. In decoder, the output of the last convolutional layer is fed to softmax which generates a 2 channel image of probabilities. From the softmax output, at each pixel, we consider the class with maximum probability to generate the final output binary mask image with classes correspond to tissue and airway cavity. The 3D convolutional filters have a dimension of $2 \times 3 \times 3$ and the max-pooling layer has a pooling dimension of $2 \times 2 \times 2$ which reduces the input dimension by half; In decoder, the up-sampling is done by a factor of 2 to get the original input dimension. Thus, the 3D-CNN takes any arbitrary size input and generates an output of corresponding size. The receptive field of the encoder is 22 in the temporal axis, which indicates that the encoder output feature map considers 22 number of frames (∼one second temporal context) for each pixel. The 3D-CNN takes an input image sequence with a dimension of $N \times 68 \times 68 \times 1$ where N is the number of frames in a given input rtMRI video. The corresponding output image sequence has a dimension of $N \times 68 \times 68 \times 3$. Thus, for each frame in a given input rtMRI video, 3 outputs are generated from the 3 decoders. Likewise, the 3D-CNN model is trained to generate the three binary masks (M_1, M_2, M_3) as target outputs corresponding to the three complete ATBs (C_1, C_2, C_3) respectively which are illustrated in Figure 2(c). For training, the ground truth binary masks are generated as explained in section 2. Each binary mask consists of two classes: class-1 corresponds to the pixels inside the complete ATB (tissue region) and class-0 corresponds to the pixels outside the complete ATB (air-cavity region). In this way, the semantic segmentation of the rtMRI image is formulated as a binary classification problem. Hence, binary cross entropy loss function is optimized to train 3D-CNN. The binary cross entropy losses corresponding to the three outputs of 3D-CNN are added and used to optimize the weights of the encoder and three decoders during training. Likewise, the 3D-CNN model is trained for semantic segmentation of the rtMRI images.

3.2. Contour prediction and contour pruning

Given the input rtMRI video, the trained 3D-CNN generates three binary masks corresponding to the three complete ATBs as shown in Figure 1. The predicted binary masks are denoted as $\hat{M}_1$, $\hat{M}_2$, $\hat{M}_3$. Figure 3 illustrates the contour prediction and contour pruning steps. The contour prediction step uses canny edge detection algorithm to predict the complete ATBs using the binary mask images. The predicted complete ATBs are denoted as $C_1^{f'}$, C_2^f, and C_3^f. In contour pruning step, $C_1^{f'}$, C_2^f, and C_3^f are pruned to obtain the ATBs within the vocal tract which are denoted as C_1^v, C_2^v, and C_3^v respectively. The contour

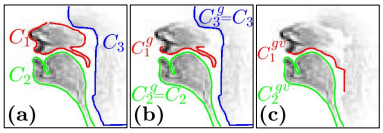

Figure 4: *Illustration of (a) manually annotated ground truth ATBs (b) complete ground truth ATBs (c) ground truth ATBs within the vocal tract after contour pruning.*

prediction and contour pruning techniques are described in [23].

4. Experimental Setup

In this work, we analyze the performance of the proposed 3D-CNN based approach and baseline FDM and SegNet approaches in unseen subject condition. Thus, we use a leave-one-subject-out (LOSO) experimental setup. Since, the baseline MG approach is unsupervised, the ATBs are predicted for all the rtMRI videos and the performance is evaluated across all the subjects. The LOSO setup consists of 10 folds as we consider a total of 10 subjects. 11 videos are considered for every subject. In each fold, the train and validation sets consist of 9 subjects and test set contains 1 subject. The train set contains 10 videos from the 9 subjects (total 90 videos) and for validation set, we consider the remaining one video from the 9 subjects (total 9 videos) and the test set contains 11 videos from the remaining subject. Likewise, we choose the 10 subjects in a round-robin fashion which forms a 10 fold cross validation setup. For training and validation of the FDM and SegNet based approaches, the corresponding frames of the rtMRI videos from the train, validation and test sets are considered. For training and validation of 3D-CNN, each rtMRI video from the train and validation sets is divided into one-second duration chunks with an overlap of 0.5 seconds. Thus, the input to the 3D-CNN has a dimension of $24 \times 68 \times 68 \times 1$ which is a stack of 24 frames corresponding to the one-second duration chunk. However, for testing, the image sequence from the entire rtMRI video is used as input since the 3D-CNN model can take arbitrary sized input. Each fold, on an average, consists of ∼685, ∼68 one-second duration video chunks in train and validation sets respectively. The 3D-CNN is trained for a maximum of 30 epochs with early stopping criterion based on the validation loss.

Evaluation metric: To evaluate the proposed approaches, we use two metrics: 1) Dynamic Time Warping (DTW) distance is used to measure the alignment between the predicted and ground truth ATBs [33]. The DTW scores have a unit of pixel. The DTW distance is less if the predicted and ground truth ATBs have a similar shape and located close to each other. 2) Pixel classification accuracy is used to evaluate the performance of 3D-CNN and SegNet architectures' performance in semantic segmentation [23]. Pixel classification accuracy indicates the fraction of the pixels that are correctly classified in the predicted image compared to the ground truth image. In this work, two types of performance evaluations are done using DTW distance: (1) evaluation of the complete predicted ATBs $C_1^{f'}$, C_2^f, and C_3^f. (2) evaluation of the predicted ATBs within the vocal tract C_1^v, C_2^v. To evaluate the predicted complete upper ATB, the ground truth C_1^g contour is obtained considering the non-fixed points from the manually annotated complete ATB (C_1) as shown in Figure 4(b). In a similar way, the complete predicted ATB C_1^f is obtained considering the non-fixed points from the predicted complete ATB ($C_1^{f'}$). To evaluate C_2^f and C_3^f, manually annotated complete lower ATB ($C_2 = C_2^g$) and the contour corresponding to pharyngeal wall ($C_3 = C_3^g$) are used, respectively, which are shown in Figure 4(a) and (b).

To evaluate the predicted ATBs within the vocal tract, the upper and lower ground truth complete ATBs (C_1^g, C_2^g) are pruned to obtain the ATBs within the vocal tract using the contour pruning approach as described in section 3. The pruned upper and lower ground truth ATBs within the vocal tract are represented as C_1^{gv}, C_2^{gv} respectively and they are illustrated in Figure 4(c).

5. Results and Discussions

Table 1 shows the average ± standard deviation (std) of the DTW distance for complete ATBs predicted from FDM, SegNet, and 3D-CNN approaches. The MG approach does not predict complete ATBs and the FDM approach does not predict C_3^f. Thus, the corresponding results are not provided in Table 1. It is observed that the 3D-CNN provides better performance compared to the baseline FDM and SegNet approaches for all the complete ATBs. The average DTW distance for the complete ATBs predicted using 3D-CNN is 56.1% and 53.7% lower than those using FDM and SegNet respectively. Figure 5 illustrates the complete ground truth ATBs and predicted ATBs from FDM, SegNet, and 3D-CNN. It is observed that the predicted ATBs from 3D-CNN are more accurate compared to the baseline FDM and SegNet approaches.

Table 1: *Average (± standard deviation) of DTW distances (in pixels) for the predicted complete ATBs obtained using FDM, SegNet, and 3D-CNN (bold indicates least DTW distances).*

ATB	FDM	SegNet	3D-CNN
C_f^1	2.86 ± 0.46	1.91 ± 0.42	**0.94 ± 0.16**
C_f^2	2.51 ± 1.06	1.73 ± 0.97	**1.38 ± 0.86**
C_f^3	-	2.4 ± 0.34	**0.75 ± 1.22**

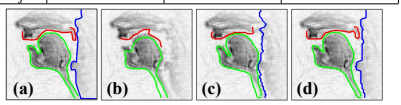

Figure 5: *Illustration of the complete (a) ground truth ATBs (C_1^g, C_2^g, C_3) and predicted ATBs (b) (C_1^f, C_2^f) from FDM, (C_1^f, C_2^f, C_3^f) from (c) SegNet (d) 3D-CNN approaches.*

Table 2 shows the average ± std of the DTW distance for ATBs within the vocal tract predicted from MG, FDM, SegNet, and 3D-CNN. It is observed that the 3D-CNN provides better performance compared to the baselines MG, FDM, SegNet for both upper and lower ATBs. The average DTW distance for the ATBs within the vocal tract from 3D-CNN is 32.9%, 61.2%, and 32.3% lower than that using the MG, FDM, and SegNet approaches, respectively. Interestingly, the supervised FDM approach does not perform better than the unsupervised MG approach for both upper and lower ATBs. The SegNet also does not perform better than the baseline MG approach for upper ATB. However, in [19] and [23], it has been shown that both FDM and SegNet approaches perform better than the MG approach in seen subject condition. Thus, the supervised algorithms perform well in seen subject condition and they do not perform well for the unseen subjects. Thus, utilizing both temporal and spatial information using suitable architecture helps in accurate ATB prediction independent of the morphology of the subject. Figure 6 illustrates the ground truth ATBs within the vocal tract and corresponding predicted ATBs obtained from MG, FDM, SegNet, and 3D-CNN. It is observed that the 3D-CNN predicts more accurate ATBs compared to the baselines. The FDM approach predicts the test ATBs as a combination of train ATBs which optimize the FDM based objective function. In unseen subject condition, due to the mismatch in morphology of train and test subjects, the predicted ATBs do not capture the vocal tract shape leading to erroneous ATBs.

Table 2: *Average (± standard deviation) of DTW distances (in pixels) for the predicted ATBs within vocal tract obtained using MG, FDM, SegNet, and 3D-CNN (bold indicates least DTW distances).*

ATB	MG	FDM	SegNet	3D-CNN
C_v^1	1.53 ± 0.29	2.19 ± 0.29	1.86 ± 0.45	**1.08 ± 0.21**
C_v^2	1.65 ± 0.38	3.71 ± 0.80	1.33 ± 0.38	**1.05 ± 0.27**

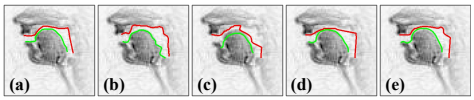

Figure 6: *Illustration of (a) ground truth ATBs within vocal tract (C_1^{gv}, C_2^{gv}) and the corresponding predicted ATBs (C_1^v, C_2^v) from (b) MG (c) FDM (d) SegNet (e) 3D-CNN approaches.*

In SegNet based approach, the 2-D CNN filters are applied for each rtMRI frame independently without considering the temporal information. Thus, the filters could learn to understand the morphology of the subject rather than capturing smoothly varying vocal tract dynamics. Hence, for an unseen subject, the SegNet model predicts erroneous ATBs. Table 3 shows the average pixel classification accuracy for the predicted binary masks ($\hat{M}_1$, $\hat{M}_2$, $\hat{M}_3$) obtained from SegNet and 3D-CNN for the validation and test data. For 3D-CNN, it is observed that the pixel classification accuracy for the validation and test data is very high and almost the same. However, for SegNet, the pixel accuracy for the test data is less compared to the validation data. As explained in Section 4, the validation data consists of unseen sentences from the subjects which are used for training and the test data consists of unseen subject's data which is not used in training. Thus, the SegNet provides accurate semantic segmentation for seen subjects and it fails for unseen subjects due to the morphology mismatch compared to the subjects which are used for training. However, 3D-CNN provides accurate semantic segmentation for unseen subjects independent of the morphology.

Table 3: *Average pixel classification accuracy on test and validation data for SegNet and 3D-CNN*

Method	$\hat{M}_1$		$\hat{M}_2$		$\hat{M}_3$	
	Test	Validation	Test	Validation	Test	Validation
SegNet	96.62	99.83	97.71	99.72	95.64	99.78
3D-CNN	99.43	99.98	99.18	99.96	99.35	99.97

6. Conclusion

In this work, we proposed a 3D-CNN model which uses both spatial and temporal information of the rtMRI images and predicts accurate ATBs for arbitrary length of the rtMRI videos. The 3D-CNN utilizes the temporal continuity of rtMRI frames to capture the smoothly varying vocal tract dynamics. Experiments with LOSO cross-validation setup reveal that the 3D-CNN based approach provides better performance in terms of the DTW distance compared to the baseline MG, FDM and SegNet approaches which indicates that the proposed approach has better generalizability for unseen subjects. The 3D-CNN model provides high pixel classification accuracy for semantic segmentation compared to SegNet. In our future work, we will analyze the minimum number of subjects and videos from each subject that are required to train the 3D-CNN for obtaining the saturating pixel classification accuracy. We will also exploit the proposed 3D-CNN based approach for the 3-dimensional rtMRI images.

7. Acknowledgement

Authors thank the Department of Science and Technology (DST), Government of India for their support in this work.

8. References

[1] C. Hagedorn, T. Sorensen, A. Lammert, A. Toutios, L. Goldstein, D. Byrd, and S. Narayanan, "Engineering innovation in speech science: Data and technologies," *Perspectives of the ASHA Special Interest Groups*, vol. 4, no. 2, pp. 411–420, Apr 2019.

[2] 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.

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

[4] D. Maurer, B. Gröne, T. Landis, G. Hoch, and P. W. Schönle, "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, Jan 1993.

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

[6] A. Toutios, T. Sorensen, K. Somandepalli, R. Alexander, and S. S. Narayanan, "Articulatory synthesis based on real-time magnetic resonance imaging data," in *INTERSPEECH*, Sep 2016, pp. 1492–1496.

[7] 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)*, Apr 2015, pp. 4265–4269.

[8] 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.

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

[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)*, 2014, pp. 308–311.

[12] F.-Y. Hsieh, L. Goldstein, D. Byrd, and S. Narayanan, "Pharyngeal constriction in English diphthong production," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, p. 060271, 2013.

[13] S. Asadiabadi and E. Erzin, "Vocal tract airway tissue boundary tracking for rtMRI using shape and appearance priors," Aug 2017, pp. 636–640.

[14] A. Lammert, V. Ramanarayanan, M. Proctor, and S. Narayanan, "Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis," *INTERSPEECH*, pp. 959–962, Jan 2013.

[15] A. Toutios and S. Narayanan, "Factor analysis of vocal-tract outlines derived from real-time magnetic resonance imaging data," in *18th International Congress of Phonetic Sciences (ICPhS)*, Aug 2015.

[16] T. Sorensen, A. Toutios, L. Goldstein, and S. S. Narayanan, "Characterizing vocal tract dynamics across speakers using Real-Time MRI," in *INTERSPEECH*, Sep 2016, pp. 465–469.

[17] 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)*, Mar 2016, pp. 937–941.

[18] 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)*, May 2014, pp. 222 – 225.

[19] A. Koparkar and P. K. 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5004–5008.

[20] 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*, Aug 2017, pp. 631–635.

[21] V. CA, R. Mannem, and P. K. Ghosh, "Air-tissue boundary segmentation in real-time magnetic resonance imaging video using semantic segmentation with fully convolutional networks," in *INTERSPEECH*, Sep 2018, pp. 3132–3136.

[22] R. Mannem, V. Ca, and P. K. Ghosh, "A segnet based image enhancement technique for air-tissue boundary segmentation in real-time magnetic resonance imaging video," in *National Conference on Communications (NCC)*, Feb 2019, pp. 1–6.

[23] C. Valliappan, A. Kumar, R. Mannem, G. Karthik, and P. K. Ghosh, "An improved air tissue boundary segmentation technique for real time magnetic resonance imaging video using SegNet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5921–5925.

[24] R. Mannem and P. K. Ghosh, "Air-tissue boundary segmentation in real time magnetic resonance imaging video using a convolutional encoder-decoder network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5941–5945.

[25] S. A. Hebbar, R. Sharma, K. Somandepalli, A. Toutios, and S. Narayanan, "Vocal tract articulatory contour detection in real-time magnetic resonance images using spatio-temporal context," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 7354–7358.

[26] 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, Mar 2016.

[27] S. Chandana, C. Yarra, R. Aggarwal, S. Mittal, K. N K, R. K T, A. Singh, and P. Kumar Ghosh, "Automatic visual augmentation for concatenation based synthesized articulatory videos from real-time MRI data for spoken language training," Sep 2018, pp. 3127–3131.

[28] J. Tu, M. Liu, and H. Liu, "Skeleton-based human action recognition using spatial temporal 3D convolutional neural networks," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, Jul 2018, pp. 1–6.

[29] W. Yang, Y. Chen, C. Huang, and M. Gao, "Video-based human action recognition using spatial pyramid pooling and 3D densely convolutional networks," *Future Internet*, vol. 10, no. 12, p. 115, Dec 2018.

[30] 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, Sep 2014.

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

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv 1409.1556*, Sep 2014.

[33] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," *in KDD workshop*, vol. 10, no. 16, pp. 359–370, Jul 1994.