Accepted at INTERSPEECH 2022

Air tissue boundary segmentation using regional loss in real-time Magnetic Resonance Imaging video for speech production

Anwesha Roy, Varun Belagali, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India

anwesharoy@iisc.ac.in, varunbelagali98@gmail.com, prasantg@iisc.ac.in

Abstract

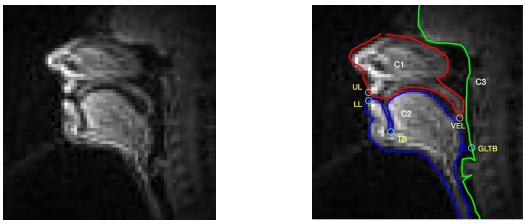
The SegNet model has been shown to provide the best performance in air-tissue boundary (ATB) segmentation in real-time Magnetic Resonance Imaging (rtMRI) videos in seen subject conditions. The SegNet model uses overall binary cross entropy as the loss function. However, such a global loss function does not give enough emphasis on regions which are more prone to errors. In this work, together with global loss, we explore the use of regional loss functions which focus on areas of the contours which have been analysed as error prone in the past. Evaluation is done using global Dynamic Time Warping (DTW) distance as well as regional metrics. The regional metrics used are EVEL and VELrDTW for contour1, and ETB and TBrDTW for contour2. We show that using such combinations of regional and global losses improves the regional, as well as global, evaluation metrics. For the best combination of losses, the two regional metrics show an improvement of 37.2% and 25.3% for contour1 and 23.9% and 28.4% for contour2, over a baseline model which uses only global loss. Global DTW distance, on the other hand, improves by 11.2% for contour1 and 5.6% for contour2.

Index Terms: real-time Magnetic Resonance Imaging video, air-tissue boundary segmentation, SegNet, tongue base, velum, dice loss, binary cross entropy loss

1. Introduction

Real-time Magnetic Resonance Imaging (rtMRI) is a tool used exhaustively in speech science and linguistics to understand the dynamics of the speech production process across languages and health conditions [1]. rtMRI has two advantages over other methods like X-ray [2], Ultrasound [3] and Electromagnetic articulography [4] - it is non-invasive, and it captures a complete view of the vocal tract including pharyngeal structures [5]. The rtMRI video provides the spatio-temporal information of speech articulators, which helps in modelling speech production. For this purpose, a common step is to obtain the air-tissue boundary (ATB) segmentation in all frames of the rtMRI video. The accurate estimation of ATBs of the upper airway of the vocal tract is essential for many speech processing applications [6, 7]. Many works have also used segmented ATBs for different applications like speaker verification [8], text-to-speech synthesis [9], visual augmentation for synthesized articulatory videos [10], and analysis of vocal tract movement [11, 12]. Thus, it is necessary to have a proper and accurate air-tissue boundary segmentation in every frame of the rtMRI videos.

Many works in the past have addressed the problem of ATB segmentation in rtMRI frames using several supervised and unsupervised approaches. Accuracy of ATBs predicted by supervised approaches is higher as unsupervised approaches consider low-level gradients which may not correspond properly to ATB points. The works presented in [13] and [14] formulate ATB segmentation as a multi-class classification problem where each pixel of an rtMRI image is assigned to one of 14 classes corresponding to different articulators. On the other hand, works illustrated in [15, 16, 17, 18, 19] generate ATBs as series of 2D points tracing the vocal tract boundaries, since the applications involving vocal tract boundaries require exact ATBs instead of an rtMRI image with pixel classification. Advait el al. [16] proposed an approach based on the Fisher Discriminant Measure (FDM) 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. Here, the dynamics of the predicted ATBs are limited by the ATBs from the training set. To overcome these limitations, Valliappan et al. [17] proposed a semantic segmentation based ATB prediction technique using a 2-dimensional deep convolutional encoder-decoder network (SegNet). The SegNet based approach used in [17] and [18] has been shown to provide best performance in seen subject conditions.



(a) (b)

Figure 1: *Illustration of (a) a rtMRI frame, (b) manually annotated Air Tissue Boundaries in it*

Figure 1 illustrates the three manually annotated contours, contour1 (C1), contour2 (C2) and contour3 (C3) in an rtMRI frame. Careful analysis of predicted contours in [20] reveals errors in regions like the velum part of contour1 and tongue base section of contour2, which are not captured by a global evaluation metric like DTW distance. They propose analysis, error detection and error correction schemes based on new regional evaluation metrics. The velum and tongue base regions tend to be error-prone because these sections have higher motion across the video and tend to have more variation compared to other regions. Another reason for errors in TB regions is that the gap between the tongue and lower lip region becomes very narrow when the tongue moves closer to the lip. Using a global binary cross entropy, like in previous works, does not always penalize the erroneous predictions in velum and TB regions as they form a small part of masks used in loss computations. Designing losses that focus on such error-prone regions can make the networks more robust and significantly improve the ATB segmentation. In this work, we propose the introduction of regional

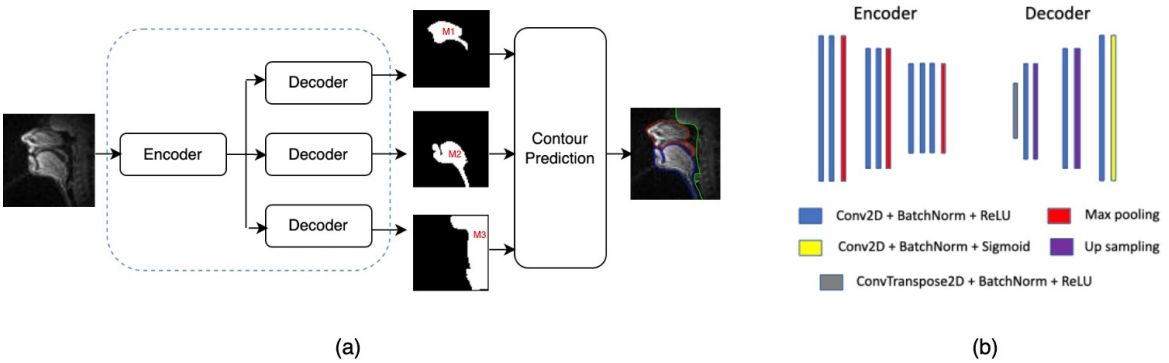


Figure 2: *(a) Illustration of the steps in the segmentation process and (b) architectures of encoder and decoders in SegNet*

losses in the model training process to overcome region-specific (velum and TB) errors. We show that introducing such losses, in addition to global loss, reduces the regional errors to a great extent.

Two different losses are considered in the regions around VEL and TB - binary cross entropy (BCE) loss and dice loss. It is observed that dice-loss based models perform better than their BCE loss based counterparts. Best performance is observed for the regional dice loss based model where both C1 and C2 regional losses are considered simultaneously, and the weights used for C1 and C2 regional dice loss are 0.8 and 0.6 respectively. This model shows an improvement of 37.2% and 25.3% in EVEL and VELrDTW over the baseline model and ETB and TBrDTW improve by 23.9% and 28.4% respectively.

2. Dataset

In this work, we use the USC-TIMIT corpus [21]. The USC-TIMIT database consists of rtMRI videos of the upper airway in the midsagittal plane, of five female (F1, F2, F3, F4, F5) and five male (M1, M2, M3, M4, M5) subjects speaking 460 sentences from the MOCHA-TIMIT database [22]. The videos have a frame rate of 23.18 frames/second and each frame of the video has a spatial resolution of 68 × 68 pixels (2.9mm × 2.9mm).

10 rtMRI videos (one for each sentence) are used from each of the 10 subjects. The selected videos have 3043 frames for the five female subjects and 3145 frames for the five male subjects. 80 videos are used for training the model, 10 videos for the validation set and 10 videos are used as the test set. ATBs are drawn manually in each rtMRI frame. A MATLAB based graphical user interface (GUI) is used for manual annotation of the three contours representing the complete ATB in a typical rtMRI frame, as shown in Figure 1 . These manually annotated ground truth contours are denoted as C1, C2 and C3. The details of the manual annotation are available in [23]. Along with the contours, upper lip (UL), lower lip (LL), tongue base (TB), velum tip (VEL) and glottis begin (GLTB) were also marked for each frame using the GUI. C1 is a closed contour that starts from UL, goes through the hard palate till VEL and goes around the fixed nasal tract. C2 is a closed contour that covers the jawline, LL, tongue blade and extends below the epiglottis. The C3 contour marks the pharyngeal wall. Ground truth binary masks are generated from these manual annotations, where class-1 corresponds to tissue and class-0 is for the air cavity.

3. Proposed Method

The trained SegNet model [24] generates three semantically segmented images, one for each contour. The predicted binary masks are further used to estimate the three complete ATBs using a contour prediction approach as used in previous works. The complete method is illustrated in Figure 2 (a).

3.1. Model Architecture

The encoder and decoder networks of the SegNet is implemented using the multi-convolutional layered VGG-16 architecture [25], shown in Figure 2 (b). For each frame in a given input rtMRI video, 3 outputs are generated from the 3 decoders. The model generates three binary masks (M1,M2,M3) as outputs, where class-1 corresponds to tissue and class-0 is for the air cavity. On these masks, contour prediction gives corresponding ATBs (C1, C2,C3), which are illustrated in Figure 1 (b). In previous works, the cumulative BCE loss of the three predicted masks of the model (with respect to ground truth binary masks) has been used to optimize the weights of the encoder and three decoders during training. In this work, regional losses are used in addition to the BCE loss during optimization of weights in the training process.

3.2. Loss functions

Global and regional loss functions used in this work are explained in detail in the following subsections.

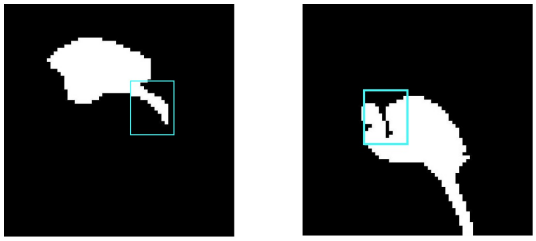
3.2.1. Global loss

Global loss for each contour is estimated on the mask of the entire contour. Similar to previous works, we have used binary cross entropy (BCE) as the global loss for training our models. Equation 1 indicates the BCE loss computed between sigmoid output P (corresponding to the complete contour) and ground truth Y.

$$Loss_{BCE} = -Y * \log P - (1 - Y) * \log(1 - P) \quad (1)$$

3.2.2. Regional loss

Regional loss focuses on specific regions of the masks. There are regions in the video frame like tongue base in C2 and velum in C1 which are prone to erroneous segmentation. Using regional losses (which focus on error prone regions) along with global loss improves the quality of segmentation in these regions, reduces C1 VEL errors and C2 TB errors reported in [20] to a great extent.



(a) (b)

Figure 3: *Regions considered for regional loss for contour 1(a) and contour2 (b)*

The two regions considered for regional losses are - C1 around the VEL region, and C2 around the TB region, as shown in Figure 3. The exact dimensions of the regions are selected based on metrics evaluated on the validation set, as explained in the next section. Once the region is fixed, both BCE loss and dice loss [26] are used as regional losses. Equation 2 indicates the dice score calculated between sigmoid output *P* (corresponding to the regions selected for each contour) and ground truth *Y*. Dice loss is considered in our experiments because it has been seen in literature that dice loss performs well in boundary regions for segmentation problems.

$$Loss_{Dice} = 1 - 2\frac{sum(P * Y)}{sum(P) + sum(Y)} \quad (2)$$

4. Experiments and Results

The experimental setup for this work, the evaluation metrics used, and results are illustrated in the following subsections.

4.1. Experimental setup

In this work, a total of 100 rtMRI videos (10 videos from each subject) are used for ATB segmentation. The experiments are carried out in a four-fold cross-validation set-up. In each fold, for each subject, a random permutation is done using the video numbers. First 8 videos of the permutation are used for training, 9^{th} for validation, and 10^{th} for test. Thus, in each fold, the training set has 80 videos, and the validation and test set both have 10 videos each. To ensure that the permutations are unique in each of the folds, random permutation is given seed value equal to fold number.

The SegNet model is trained for a maximum of 500 epochs with an early stopping condition imposed based on the validation loss with a patience of 5 epochs.

4.1.1. Evaluation metrics

Global DTW distance has been used in previous works as an evaluation metric to analyse the overall quality of ATB segmentation. Regional metrics proposed in [20] are evaluated to explicitly observe the effect of the proposed method in error-prone regions. Regional metrics for C1 include regional DTW (VELrDTW) and EVEL, whereas regional DTW (TBrDTW) and ETB are the metrics for C2. VELrDTW distance is computed around the VEL region, taking 30% of the total number of points of C1 from the pharyngeal wall end. TBrDTW for C2 is found by calculating the DTW distance between annotated and predicted C2 in the region between LL and the uppermost point on the tongue. EVEL, for C1, is the Euclidean distance between predicted and ground truth VEL position and ETB refers to the Euclidean distance between the predicted and ground truth TB position (ETB).

4.1.2. Model Loss

BCE is used as the global loss for optimization. The total global loss (gBCE) is the summation of the BCE loss for each contour (C1BCE + C2BCE + C3BCE). To find regional losses for C1 and C2, rectangular patches are selected around velum and TB regions respectively, based on the validation dataset. Then the losses are estimated in these regions.

Patch selection

Let (v_1 , v_2) and (t_1 , t_2) be the positions of VEL and TB respectively, in the 68x68 image space, where (0,0) lies at the top left corner. The lower lip point is denoted as (l_1 , l_2). The two rectangular patches considered for C1 and their corresponding vertices are:

1. **c1_patch1:** Vertices at (v_1 - 10 , v_2 - 10), (v_1 - 10 , v_2), (v_1 , v_2), and (v_1 , v_2 - 10)
2. **c1_patch2:** Vertices at (v_1 - 10 , v_2 - 10), (v_1 - 10 , v_2 + 5), (v_1 + 5 , v_2 + 5), and (v_1 + 5 , v_2 - 10)

These two patches are used for running the Seg_c1dice model (explained in detail in the next subsections) and the regional metrics (EVEL and VELrDTW) are observed (on the validation set) to be better for the first setup.

For C2, previous works have observed that a recurring error is the improper prediction of the TB dip between lip and tongue. So, for C2, the following three patches are experimented with:

1. **c2_patch1:** Vertices at (l_1 + 2 , l_2 + 2), (l_1 + 2 , t_2 - 2), (t_1 - 2 , t_2 - 2), and (t_1 - 2 , l_2 + 2)
2. **c2_patch2:** Vertices at (l_1 , l_2), (l_1 , t_2), (t_1 , t_2), and (t_1 , l_2)
3. **c2_patch3:** Vertices at (l_1 - 5 , l_2 - 5), (l_1 - 5 , t_2 + 5), (t_1 + 5 , t_2 + 5), and (t_1 + 5 , l_2 - 5)

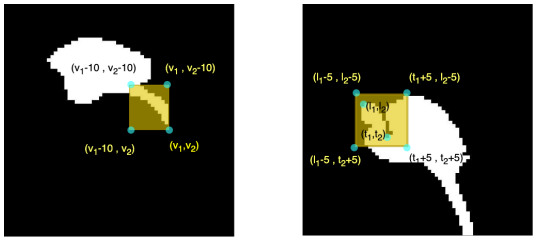
These three patches are used to run the Seg_c2dice model and the regional metrics (ETB and TBrDTW) are observed to be better for the third patch.

All regional metrics, calculated on validation set, are tabulated in Table 1 and the final selected patches are illustrated in Figure 4.

Table 1: *Mean ± standard deviation (in pixels) of global and regional evaluation metrics for different patches considered for C1 and C2*

Metric	c1_patch1	c1_patch2
C1DTW	1.02 ± 0.16	1.02 ± 0.15
EVEL	1.51 ± 0.83	1.91 ± 0.89
VELrDTW	1.18 ± 0.37	1.28 ± 0.42

Metric	c2_patch1	c2_patch2	c2_patch3
C2DTW	1.24 ± 0.18	1.23 ± 0.19	1.19 ± 0.18
ETB	3.36 ± 3.41	3.09 ± 3.05	2.49 ± 1.48
TBrDTW	1.55 ± 0.92	1.45 ± 0.77	1.34 ± 0.74



(a) (b)

Figure 4: *Selected patches for regional loss estimation for contour 1 (a) and contour2 (b) respectively*

Table 2: *Mean ± standard deviation of global and regional metrics (in pixels) for different models. Seg_c1BCE, Seg_c2BCE and Seg_c12BCE are based on regional BCE loss, whereas Seg_c1dice, Seg_c2dice and Seg_c12dice are regional dice loss based models.*

Model name		**SegNet**	**Seg_c1BCE**	**Seg_c2BCE**	**Seg_c12BCE**	**Seg_c1dice**	**Seg_c2dice**	**Seg_c12dice**
		Global loss	Global loss + Regional BCE			Global loss + Regional dice		
Loss		gBCE	gBCE + VELrBCE	gBCE + TBrBCE	gBCE + VELrBCE + TBrBCE	gBCE + VELrdice	gBCE + TBrdice	gBCE + VELrdice + TBrdice
C1DTW	Global	1.07 ± 0.81	1.01 ± 0.17	1.04 ± 0.16	1.02 ± 0.17	1.02 ± 0.16	1.04 ± 0.16	1.02 ± 0.15
C2DTW	metrics	1.25 ± 0.73	1.22 ± 0.19	1.19 ± 0.19	1.18 ± 0.18	1.21 ± 0.19	1.18 ± 0.19	1.18 ± 0.19
EVEL		2.34 ± 1.76	1.84 ± 1.03	2.13 ± 1.04	1.95 ± 1.06	1.56 ± 0.90	2.07 ± 1.05	$\mathbf{1.54 \pm 0.97}$
VELrDTW	Regional	1.50 ± 1.27	1.19 ± 0.39	1.37 ± 0.40	1.22 ± 0.40	1.13 ± 0.35	1.28 ± 0.41	$\mathbf{1.12 \pm 0.36}$
ETB	metrics	2.84 ± 3.18	2.82 ± 2.37	2.60 ± 2.49	2.56 ± 1.96	2.67 ± 2.54	2.56 ± 2.51	$\mathbf{2.43 \pm 2.35}$
TBrDTW		1.76 ± 0.95	1.58 ± 1.04	1.32 ± 0.65	1.29 ± 0.38	1.40 ± 0.64	1.29 ± 0.62	$\mathbf{1.27 \pm 0.60}$

Loss estimation

Let C1 regional loss be denoted as VELrloss and C2 regional loss as TBrloss. BCE loss in patch region for C1 and C2 is denoted as VELrBCE and TBrBCE respectively. Similarly, regional dice loss in these regions is denoted as VELrdice and TBrdice. For both the types of regional losses (dice and BCE), three models are considered. The first model considers only VELrloss along with the global loss, and the second one considers sum of global loss and TBrloss. The last model combines both regional losses with the global loss. The combinations of loss for each model are illustrated in Table 2.

4.2. Results

As explained in the previous subsections, six different models have been used for experiments (three for regional BCE loss and three for regional dice loss). Seg_c1BCE and Seg_c1dice models consider the sum of gBCE and VELrloss for optimization. Similarly, the sum of gBCE and TBrloss is considered for models Seg_c2BCE and Seg_c2dice. The combined models (Seg_c12BCE and Seg_c12dice) use both VELrloss and TBrloss along with gBCE. The average ± standard deviation values of global metrics (C1DTW and C2DTW), as well as, the regional metrics (EVEL, VELrDTW, ETB and TBrDTW) have been reported in Table 2 for the baseline model (which does not consider any regional loss) as well as, the six proposed models. It is observed that C3DTW values, for all models, have a range similar to the baseline model.

All models show significant improvement in metrics over the baseline model. Regional dice loss based models perform better than the corresponding BCE loss based models. The Seg_c12dice model gives the best performance for both C1 and C2, based on regional metrics. Global DTW also improves for this model. On the test set, EVEL and VELrDTW improve by 34.2% and 25.3% respectively, as illustrated in Table 2. ETB decreases by 14.4% and TBrDTW improves by 27.8%.

Adding weights to regional loss

As the best performance is observed for a dice loss based model that considers both C1 and C2 regional losses simultaneously, we use a similar setup for our next experiments. Instead of directly adding regional loss to the global loss, we add weighted regional losses:

C1loss = C1BCE + w1*VELrdice, w1 = 0.4, 0.6, 0.8, 1, 1.2
C2loss = C2BCE + w2*TBrdice, w2 = 0.4, 0.6, 0.8, 1, 1.2
Total loss = C1loss + C2loss + C3BCE

A total of 25 models are trained with these combinations of weights. The regional metrics are evaluated on the validation data for all models to compare them. It is seen that models show better performance when the weight for regional loss is lower than 1. This is probably because if regional loss is given more weight than global loss, predictions in other regions of the contour (which are not considered in the patch) may be affected. The best regional metrics are observed for both C1 and C2 when weight w1 is 0.8 and w2 is 0.6. For this model, EVEL is observed to improve by 37.2% and VELrDTW by 25.3 % over the baseline SegNet model. ETB and TBrDTW show an improvement of 23.9% and 28.4% respectively. Global C1DTW also decreases by 11.2% and C2DTW by 5.6% for this model.

Figure 5 shows improvement of predictions in the C1 VEL region using the best model (b and d) over the predictions using baseline model on the respective frames (a and c). Improvement is also seen in C2 TB region using this model in Figure 5 (h) over baseline (g). 5 (e) shows that the baseline model predicted contour sometimes goes into the tissue in the TB region. This problem is not observed in the new predictions (f).

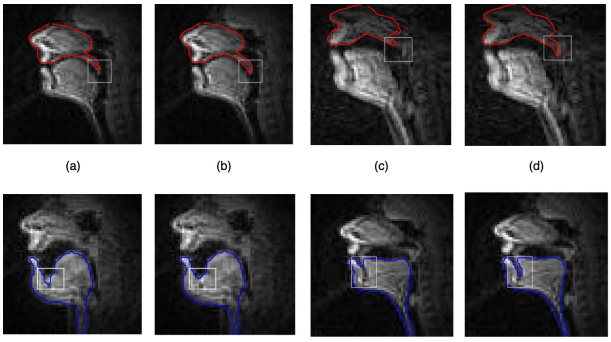


Figure 5: *Global loss based model predictions (a, c, e, g) and corresponding predictions using proposed model (b, d, f, h)*

5. Conclusion

In this work, we propose an air tissue boundary (ATB) segmentation scheme using regional loss for real-time MRI videos. Previous works on ATB segmentation have always been based on a global binary cross entropy (BCE) loss. But it has been seen in [20] that errors in predictions occur largely in two particular regions - C1 velum region and C2 tongue base region. So this work proposes the introduction of regional losses along with the global loss to rectify such regional errors in prediction. The proposed methods show improvement in regional and global evaluation metrics, and also show refinement in prediction upon visual inspection. In our future work, we will try to further improve ATB segmentation by using phonetic information regarding the frames. We will also experiment on semi-supervised approaches for ATB prediction. Further, we will observe the performance of improved ATBs, predicted using our proposed method, in different applications.

6. 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, 2019.

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

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

[4] D. Maurer, B. Gröne, T. Landis, G. Hoch, and P. Schönle, "Reexamination of the relation between the vocal tract and the vowel sound with electromagnetic articulography (ema) in vocalizations," *Clinical linguistics & phonetics*, vol. 7, no. 2, pp. 129–143, 1993.

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

[6] 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*. Citeseer, 2014, pp. 308–311.

[7] F.-Y. Hsieh, L. Goldstein, D. Byrd, and S. Narayanan, "Pharyngeal constriction in english diphthong production," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. Acoustical Society of America, 2013, p. 060271.

[8] 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 *IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4265–4269.

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

[10] S. Chandana, C. Yarra, R. Aggarwal, S. K. Mittal, N. Kausthubha, K. Raseena, A. Singh, and P. K. Ghosh, "Automatic visual augmentation for concatenation based synthesized articulatory videos from real-time mri data for spoken language training," in *Proc. Interspeech*, 2018.

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

[12] A. Lammert, M. Proctor, and S. Narayanan, "Interspeaker variability in hard palate morphology and vowel production," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 6, pp. S1924–S1924, 2013.

[13] 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.

[14] 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 *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7354–7358.

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

[16] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5004–5008.

[17] C. Valliappan, 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*, 2018, pp. 3132–3136.

[18] 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 *2019 National Conference on Communications (NCC)*. IEEE, pp. 1–6.

[19] R. Mannem, N. Gaddam, and P. K. Ghosh, "Air-tissue boundary segmentation in real time magnetic resonance imaging video using 3-d convolutional neural network." in *INTERSPEECH*, 2020, pp. 1396–1400.

[20] A. Roy, V. Belagali, and P. K. Ghosh, "An error correction scheme for improved air-tissue boundary in real-time mri video for speech production," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 8247–8251.

[21] 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)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

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

[23] 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.

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

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

[26] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 240–248.