AN ERROR CORRECTION SCHEME FOR IMPROVED AIR-TISSUE BOUNDARY IN REAL-TIME MRI VIDEO FOR SPEECH PRODUCTION

Anwesha Roy, Varun Belagali, Prasanta Kumar Ghosh

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

ABSTRACT

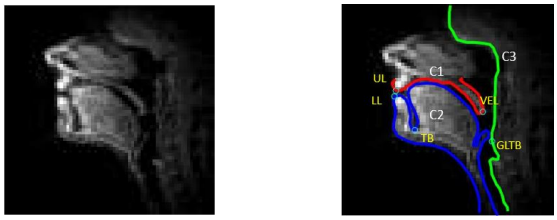
The best performance in Air-tissue boundary (ATB) segmentation of real-time Magnetic Resonance Imaging (rtMRI) videos in speech production is known to be achieved by a 3-dimensional convolutional neural network (3D-CNN) model. However, the evaluation for this model, as well as other ATB segmentation techniques reported in the literature, is done using Dynamic Time Warping (DTW) distance between the entire original and predicted contours. Such an evaluation measure may not capture local errors in the entire contour. Careful analysis of predicted contours reveals errors in regions like the velum part of contour1 and tongue base section of contour2, which are not captured in a global evaluation metric like DTW distance. In this work, we automatically detect such errors and propose a correction scheme for the same. We also propose two new evaluation metrics for ATB segmentation separately in contour1 and contour2 to explicitly capture two types of errors in these contours. The proposed detection and correction strategies result in the improvement of these two evaluation metrics by 61.8% and 61.4% for contour1 and by 67.8% and 28.4% for contour2. Traditional DTW distance, on the other hand, improves by 44.6% for contour1 and 4.0% for contour2.

***Index Terms*—** real-time Magnetic Resonance Imaging video, Air-Tissue Boundary segmentation, 3-dimensional convolutional neural network, tongue base, velum

1. INTRODUCTION

Real-time Magnetic Resonance Imaging (rtMRI) is a tool used exhaustively in speech science and linguistics to understand the speech production process in depth across languages and health conditions [1]. rtMRI has two advantages over other methods like X-ray [2], Electromagnetic articulography [3] and Ultrasound [4] - it is safe and non-invasive, and it captures a complete picture of the vocal tract including pharyngeal structures [5]. A common step before using these rtMRI videos is obtaining the Air-Tissue Boundary (ATB) segmentation in every frame. Many works have used ATBs for different applications like text-to-speech synthesis [6], speaker verification [7], visual augmentation for synthesized articulatory videos [8], and analysis of vocal tract movement [9, 10]. The accurate estimation of ATBs of the upper airway of the vocal tract is also essential for many other speech processing applications [11, 12]. Hence, it is necessary to have an accurate and proper ATB 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 [13, 14] and unsupervised approaches [15]. Although the supervised algorithms have been shown to provide accurate ATBs in the seen subject condition, a major challenge is that there is high variability in the morphology of different subjects and it is difficult to generalize ATBs in unseen subject conditions. The best performance in ATB segmentation is observed in the work by Renuka et al. [14], where the authors tackle this problem by using a 3-dimensional deep convolutional neural network (3D-CNN) for ATB segmentation. Although the temporal continuity criterion of 3D-CNN ensures that the ATBs do not vary drastically in successive frames of a rtMRI video, in the event that there is an error in one frame, it also propagates to surrounding frames. Figure 1 illustrates the three manually annotated contours, contour1 (C1), contour2 (C2) and contour3 (C3) in an rtMRI frame.



(a) (b)

Fig. 1. Illustration of (a) a rtMRI frame, (b) manually annotated Air Tissue Boundaries in it

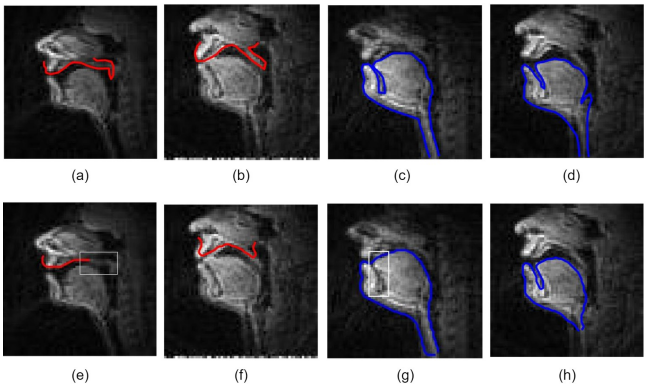


Fig. 2. Manual annotations (a,b,c,d) and corresponding erroneous predictions for C1 incomplete (e), C1 frame (f), C2 TB (g) and C2 frame (h) errors

We have observed that errors in contour1 (C1) can be categorized into two classes as shown in Figure 2 (e and f) - incomplete contours (where velum (VEL) portion is missing) and frame errors where the entire C1 has defects. For contour2 (C2), we again observe two types of errors: error in the tongue base (TB) region which is illustrated in Figure 2 (g), where TB dip is not predicted properly and

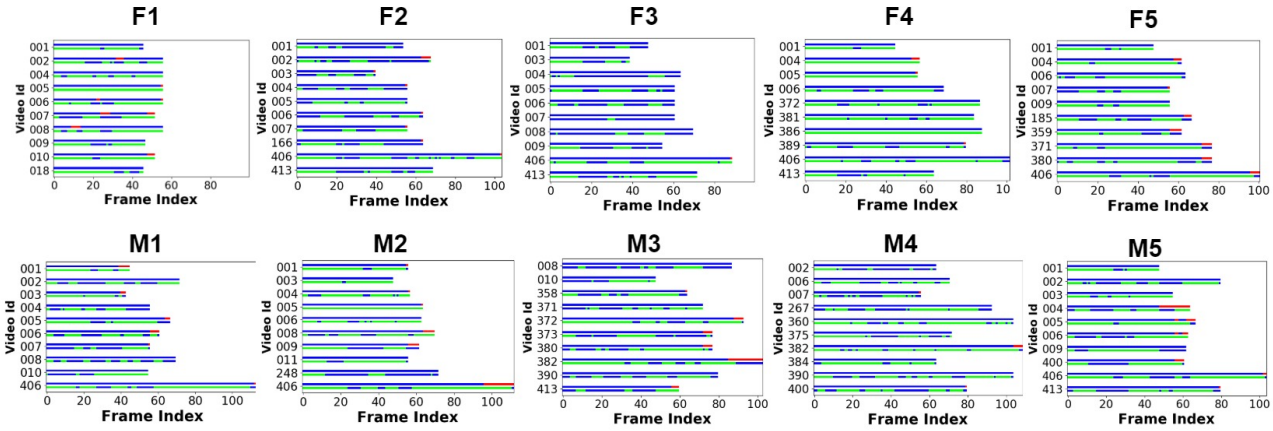


Fig. 3. Illustration of position of error frames in each video for all subjects. There are two line plots for each video index, where each line shows correct frames in blue. C1 errors are shown in red in first line and C2 errors in green in second line.

frame error, where entire frame is wrongly predicted (Figure 2 (h)). The TB dip errors occur mainly because of the low number of pixels with low intensity present in the dip region.

In cases like this, a global measure like overall DTW distance between annotation and prediction, which has been used as an evaluation metric in previous works, does not reflect which region has what kind of error. So it is essential to conduct a region specific analysis and define evaluation metrics that focus on defects in particular regions like VEL or TB.

We have analyzed the different types of errors in the ATBs in a set of 100 videos in a subjective manner. We have proposed methods to detect and correct each of these errors. Post correction, we have developed region specific metrics for the evaluation of the quality of ATB. In order to evaluate the accuracy of the predicted C2 in the region around TB, the Euclidean distance between TB in annotation and prediction (ETB), and regional DTW distance in the TB dip area (TBrDTW) are proposed as metrics. For C1, as the problem most often lies in the VEL region, we propose Euclidean distance of VEL point between model output and ground truth (EVEL), and regional DTW distance in the VEL section (VELrDTW) as the evaluation metrics. It is seen that using regional metrics helps in identifying specific problems in frames which might not be possible using global measures like DTW distance which evaluate the entire contour as a whole. For C1, after correction, the average EVEL reduces from 8.10 to 3.09 pixels, and the mean VELrDTW across frames goes down from 4.12 to 1.59 pixels. Overall DTW distance also improves from 2.04 to 1.13 pixels. For the case of C2, overall DTW distance goes down slightly (2.05 to 1.98 pixels) post-correction, but ETB reduces substantially from 11.31 to 3.64 pixels and TBrDTW from 4.26 to 3.05 pixels.

2. DATASET

USC-TIMIT corpus [16] consisting of rtMRI videos of the upper airway in the mid-sagittal plane is used in this work. There are 5 female (F1, F2, F3, F4, F5) and 5 male (M1, M2, M3, M4, M5) subjects, each of them speaking 460 sentences from MOCHA-TIMIT database [17]. The videos are recorded at a frame rate of 23.18 frames/sec. Each rtMRI frame has a spatial resolution of 68×68 (pixel dimension of 2.9mm $\times$ 2.9mm).

For this work, 3D-CNN is trained on 90 videos (9 videos from each of the 10 subjects) using the model in [14] and the ATBs are predicted on a set of 100 videos (10 videos from each subject) not seen in training. There are 33 unique sentences in this randomly chosen list of 100 videos. For these 100 videos (6738 frames), the ATBs are manually annotated using a MATLAB Graphical User Interface [18]. The manual annotation is done for three ATBs - contour1 (C1), contour2 (C2), and contour3 (C3), as well as five points that indicate upper lip (UL), lower lip (LL), tongue base (TB), velum (VEL) and glottis begin (GLTB) as illustrated in Figure 1. 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.

3. ERROR ANALYSIS

A graphical user interface using MATLAB is developed to observe both annotation and prediction in each frame. If prediction deviates a lot in any region from the annotation, the frame is declared erroneous and the contour where error is observed is noted (C1 or C2). Predicted C3 is not found to have any observable defects. The observations are then cross-checked by an unbiased viewer. Erroneous frames selected based on subjective criteria are considered as ground truth for error classification.

In Figure 3, for each video for all subjects, two line-plots are constructed. The vertical axis shows the video (sentence) ID and horizontal axis shows frame index. The first line plot illustrates correct frames in blue and C1 errors in red. The second line plot shows correct in blue and C2 error frames in green. It is observed that C2 (especially TB) errors occur recurrently throughout the video. C1 errors, on the other hand, are mostly observed at the end of videos. This may be due to the end frame padding done in 3D-CNN model during prediction.

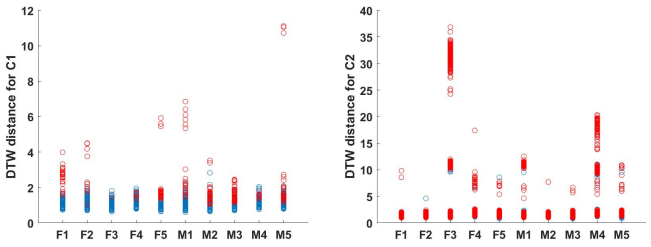
3.1. Contour 1

In a frame, if the VEL part of the contour is not predicted properly or there is noticeable deviation of entire contour from annotated contour or C1 is incomplete, then the frame is declared to have C1 error. As per the subjective analysis illustrated in Figure 3, 207 frames (3.07 %) have C1 error. In both types of C1 error illustrated in Figure 2, the VEL point is not correct.

Table 1. Summary of proposed metric, detection and correction schemes

	Error type	Evaluation metric	Detection method	Correction method
Contour1	Incomplete	EVEL, VELrDTW	Deviation from mean VEL, VEL to pharyngeal wall distance	Interpolation + Appending
	Frame			Interpolation
Contour2	TB	ETB, TBrDTW	LL to TB slope, LL to TB distance, Combined	Otsu thresholding + Contour warping
	Frame		No. of points	Interpolation

Analysis of these frames also shows that over all subjects, mean $\pm$ standard deviation (std) DTW distance between annotated C1 and predicted C1 for the frames declared erroneous subjectively is 2.15 $\pm$ 1.48 pixels compared to 1.11 $\pm$ 0.18 pixels for frames declared correct (Figure 4 (a)). But the range of DTW distance of the erroneous and correct frames overlap.



(a) (b)

Fig. 4. DTW distance analysis plots for C1 (a) and C2 (b) over all frames, where red bubbles represent erroneous frames and blue are correct ones

3.2. Contour 2

The subjective analysis shows that in about 325 frames (4.82 %), the predicted contour2 has shapes very different from expected shape (frame error) or has very low number of points. Moreover, in about 65.73 % of the frames, the TB dip region is not predicted properly.

The DTW distance analysis shown in Figure 4 (b) for C2 illustrates that erroneous and correct frames have highly overlapping range of DTW distance values. Mean $\pm$ std DTW distance for error frames is 3.61 $\pm$ 3.49 pixels and for correct frames is 1.92 $\pm$ 1.89 pixels. There is high variability in the global DTW distance value and it does not reflect C2, especially TB dip, errors.

4. IMPROVED ATB USING PROPOSED ERROR CORRECTION

The proposed evaluation metric, error detection and correction methods undertaken in this work are summarized in Table 1. The methods are explained in details in the following subsections.

4.1. Experimental Setup

The experiments are carried out in a 4 fold cross-validation set-up where, in each fold, 3 randomly chosen subjects are taken for validation and the rest of the 7 subjects for test. The results are reported on the test set.

We predict the VEL point by finding the dip in the row index of the contour in the region around velum in C1 and also find the TB in C2 by locating the dip in the region between LL and uppermost point on tongue.

*4.1.1. **Evaluation metric***

F-score [19] is used as an evaluation metric for error detection methods because of data imbalance in the set of videos considered, where only 3.07% frames have C1 errors. For error correction of C1, VELrDTW distance is computed around the VEL region, taking 30% of total number of points of C1 on the pharyngeal wall side. TBrDTW for C2 is found by calculating the DTW distance of C2 in the region between LL and uppermost tongue point. Further, for C1, the Euclidean distance between predicted and ground truth VEL position (EVEL) is also used for assessment. Similarly for C2, the Euclidean distance between the predicted and ground truth TB position (ETB) is also used as an evaluation metric. These metrics capture region specific errors observed during analysis. Additionally, global DTW is also reported for both contours.

4.2. Error Detection

We propose different methods of error detection for C1 based on relative position of VEL point. Similarly, error detection is done for C2 on the basis of position of TB and number of points. The hyperparameters used by the methods in each fold are selected based on F-score achieved on validation set.

*4.2.1. **Contour 1***

4.2.1.1. Deviation from mean velum

For each video, Euclidean distance between velum point in a frame and mean velum point across the video is used as a measure to detect erroneous frame. This method is based on the observation in Figure 3 that the C1 errors occur at the end of videos in most cases and are limited in number in a video. The variation in velum point coordinates across the rest of the video is low. Any frame with a significantly large distance between velum point and the mean velum point is likely to have a C1 error. A threshold is applied on this distance to classify the frame as erroneous. The thresholds selected based on validation data are 3.5, 4, 4, 4 for four folds. F-score of 0.86 ($\pm$ 0.01) is achieved on the test set.

4.2.1.2. Distance of velum from pharyngeal wall

If a video has many C1 errors then the mean velum can be affected by a cluster of erroneous velum points and can itself be wrong. Hence, instead of using mean velum point as reference, in this method we use the nearest point on C3 from the velum. This nearest point is found out and fixed for each subject from the manual annotations used while training the 3D-CNN model. We threshold the distance between this nearest point on C3 and predicted VEL to detect erroneous frames. The thresholds selected based on validation data are 8, 8, 8, 7.5 for four folds. This method achieves F-score of 0.85 ($\pm$ 0.02) on the test set.

4.2.1.3. Combined

In this method a frame is declared erroneous if it satisfies either one of the two aforementioned error criteria. The combined method achieves an F-score of 0.86 ($\pm$ 0.02) on the test set.

*4.2.2. **Contour 2***

4.2.2.1. Number of points

For the small number of frames (4.82%) where the predictions are very poor, it is observed that number of points of predicted C2 is

low. So, we find the average number of points in C2 over all videos in validation set and threshold the number of points at 65% of it to find erroneous frames. This method of thresholding gives a detection F-score of 0.941 ($\pm$ 0.02) on test set.

4.2.2.2. LL to TB slope

TB error occurs when the groove between the lower lip and tongue is not predicted in C2. In such cases, the slope of the line joining TB and lower lip is observed to be low. We threshold this slope to detect the erroneous frames. The thresholds (0.7, 0.7, 0.8, 1) are selected based on validation data for each of the four folds. Using this method, the F-score of 0.85 ($\pm$ 0.02) is achieved on the test set.

4.2.2.3. LL to TB distance

When a TB error is observed, it is seen that the distance from the lower lip to TB is short. A threshold on this distance is selected based on F-score achieved on validation data and any frame with distance higher than it are declared erroneous. The thresholds selected are 8, 7, 10, and 10 for four folds. The F-score on the test set is 0.88 ($\pm$ 0.02) for this method. The distance thresholding performs better than slope.

4.2.2.4. Combined

In this method a frame is declared erroneous if it satisfies either one of the three aforementioned error criteria. F-score of 0.90 ($\pm$ 0.02) is achieved on the test set.

4.3. Error Correction

The error correction methods proposed for contour 1 and contour 2 are detailed in the following section.

*4.3.1. **Contour 1***

For all erroneous frames detected using the combined method mentioned in section 4.2.1.3, contour1 is generated by linear interpolation using neighbouring frame contours as the frame to frame variance is low because of temporal continuity. In case of the incomplete C1 errors, it is observed through section wise DTW distance analysis that the incomplete part of the predicted contour is actually correct and the rest is missing. So, for them, we find the end point of original contour on interpolated one and append the rest of the interpolated contour to the existing contour so that the lower lip part of existing C1 is not affected by interpolation. For the frame errors, the interpolated contour is taken completely.

The EVEL decreases by 61.8% after correction as shown in Table 2, whereas VELrDTW improves by 61.4% for these frames. We also observe that for all frames, the DTW distance over entire C1 improves by 44.6% after correction, which is not as significant as the change in VELrDTW. An example of a frame with incomplete C1 error before and after correction is illustrated in Figure 5 (b) and (c) respectively.

*4.3.2. **Contour 2***

For all frame errors detected in section 4.2.2.1, we generate the entire contour by linear interpolation using neighbouring frame contours. Next, these frames along with the error frames detected in section 4.2.2.4 are considered for C2 TB correction. To correct C2 near TB, we first correct the position of TB. The TB is observed to be within a 15 $\times$ 20 patch of the frame in the low intensity air cavity region between lip and tongue. We apply Ostu thresholding [20] to this 15 $\times$ 20 patch in the rtMRI video frame to find darker pixels that lie within C2. The result of Ostu thresholding is a binary image, where class-1 corresponds to tissue and class-0 is for the air cavity. The lowest point in the class-0 region of the binary image within C2 is selected and marked as corrected TB. Next, we adjust C2 in the vicinity of the 3D-CNN predicted TB location by warping the existing contour. We find the shift in TB point from prediction to correction and map the shift of the neighbouring points accordingly in a gradient based fashion to find corrected C2 in TB dip region.

The evaluation metrics - ETB and TBrDTW, improve by 67.8% and 28.4% respectively, after correction as shown in Table 2. The global DTW distance, on the other hand, does not show any significant improvement. A frame with C2 TB error is shown before and after correction in Figure 5 (e) and (f) respectively.

Table 2. Mean $\pm$ standard deviation of evaluation metrics (in pixels) before and after correction for C1 and C2

	Evaluation Metric	**Pre-correction**	**Post-correction**
C1	EVEL	8.10 $\pm$ 2.33	3.09 $\pm$ 1.34
	VELrDTW	4.12 $\pm$ 1.56	1.59 $\pm$ 0.43
	DTW	2.04 $\pm$ 1.19	1.13 $\pm$ 0.19
C2	ETB	11.31 $\pm$ 3.40	3.64 $\pm$ 2.71
	TBrDTW	4.26 $\pm$ 1.26	3.05 $\pm$ 1.06
	DTW	2.06 $\pm$ 1.22	1.98 $\pm$ 1.32

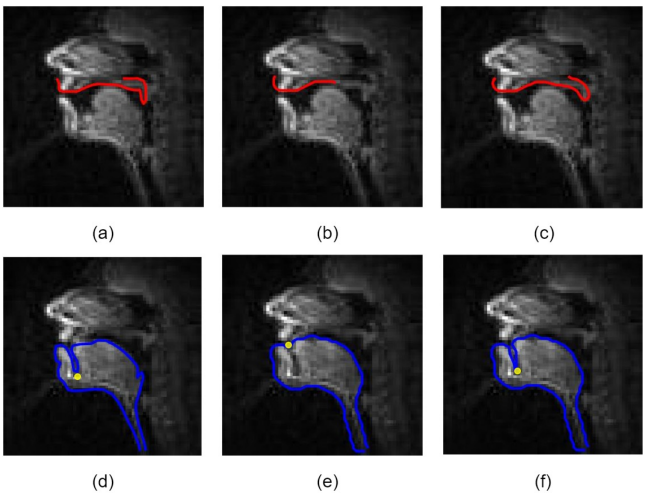


Fig. 5. Annotations (a,d), erroneous predictions (b,e) and corrected contours (c,f) for VEL and TB errors respectively, where yellow point represents TB

5. CONCLUSION

In this work, we propose an error correction scheme for improving predicted air tissue boundaries (ATBs) in real time MRI video. Careful analysis reveals different types of errors present in the results of 3D-CNN model like VEL or TB region defects. Automatic methods are proposed to detect and correct such observed errors. Further, region specific metrics are proposed for evaluation of the quality of the predicted and corrected contours. The proposed methods show observable refinement in air tissue boundaries, which is reflected in the improvement in proposed metrics. In our future work, we will explore Active Appearance Models (AAM) [21] for correction of regional errors like TB or VEL. Further, we will try to explore robust neural network approaches using region specific loss functions, which target specific problems in particular contour regions.

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] D. Maurer, B. Gröne, T. Landis, G. Hoch, and P. Schönle, "Re-examination 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.

[4] 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.

[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] 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.

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

[8] 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.

[9] 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.

[10] A. Lammert, M. Proctor, and S. Narayanan, "Interspeaker variability in hard palate morphology and vowel production," 2013.

[11] B. Parrell and S. Narayanan, "Interaction between general prosodic factors and languagespecific articulatory patterns underlies divergent outcomes of coronal stop reduction," in *International Seminar on Speech Production (ISSP) Cologne, Germany*. Citeseer, 2014, pp. 308–311.

[12] 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.

[13] 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.

[14] 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.

[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] 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.

[17] 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.

[18] 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.

[19] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation," in *Australasian joint conference on artificial intelligence*. Springer, 2006, pp. 1015–1021.

[20] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[21] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *European conference on computer vision*. Springer, 1998, pp. 484–498.