AIR-TISSUE BOUNDARY SEGMENTATION IN REAL TIME MAGNETIC RESONANCE IMAGING VIDEO USING A CONVOLUTIONAL ENCODER-DECODER NETWORK

Renuka Mannem, Prasanta Kumar Ghosh

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

ABSTRACT

In this paper, we propose a convolutional encoder-decoder network (CEDN) based approach for upper and lower Air-Tissue Boundary (ATB) segmentation within vocal tract in real-time magnetic resonance imaging (rtMRI) video frames. The output images from CEDN are processed using perimeter and moving average filters to generate smooth contours representing ATBs. Experiments are performed in both seen subject and unseen subject conditions to examine the generalizability of the CEDN based approach. The performance of the segmented ATBs is evaluated using Dynamic Time Warping distance between the ground truth contours and predicted contours. The proposed approach is compared with three baseline schemes - one grid-based unsupervised and two supervised schemes. Experiments with 5779 rtMRI images from four subjects show that the CEDN based approach performs better than the unsupervised baseline scheme by 8.5% for seen subjects case whereas it does better than the supervised baseline schemes only for lower ATB. For unseen subjects case, the proposed approach performs better than the supervised baseline schemes by 63.96%, 22.9% respectively whereas it performs worse than the unsupervised baseline scheme. However, the proposed approach outperforms the unsupervised baseline scheme when a minimum of 30 images from unseen subjects are used to adapt the trained CEDN model.

***Index Terms*—** air-tissue boundary segmentation, real-time magnetic resonance imaging video, convolutional encoder-decoder network, dynamic time warping distance.

1. INTRODUCTION

The real-time magnetic resonance imaging (rtMRI) video of the upper airway in the mid-sagittal plane during speech is an important emerging tool for speech production research. While the vocal tract movement can also be investigated using other methods like Electromagnetic articulography [2] Ultrasound [3] and X-ray [4], rtMRI has an advantage of capturing a complete view of the entire vocal tract including the pharyngeal structures in a safe and noninvasive manner. The rtMRI video provides a spatio-temporal information of speech articulators which is essential for modelling speech production. Thus, the rtMRI video is important for analyzing the dynamics of the vocal tract. The rtMRI data was used for understanding the usage of articulators in achieving acoustic goals by comparing the articulatory control of beatboxers [5]. A text-to-speech synthesis system was developed by Toutios [6] using the predicted air-tissue boundaries (ATBs) from the rtMRI video. The ATBs are the contours separating the high pixel intensity tissue region and low pixel intensity airway cavity region in the vocal tract. The ATB segmentation is required to study the time evolution of the vocal tract cross-sectional area [7] which forms the basis for many speech processing applications. The rtMRI video has been used in the studies that involve morphological structures of vocal tract [8] and analysis of vocal tract movement [9]. These studies use ATB segmentation as a pre-processing step. Hence, it is essential to have an accurate ATB segmentation of rtMRI videos to study the articulators and dynamics of the vocal tract [10, 11, 12, 13].

Several works in the past have addressed the problem of ATB segmentation. For example, Toutios et al.[14] and Sorensen et al. [15] used a factor analysis approach to predict the compact outline of the vocal tract. Somandepalli et al. [16] proposed a semantic edge detection based algorithm for ATB segmentation. A statistical method was presented by Asadiabadi et al. using the appearance and shape model of the vocal tract [17]. A data-driven approach using pixel intensity [18] and a region of interest (ROI) based method [19] for the ATB segmentation were proposed by Lammert et al. A boundary intensity map was constructed using the multi-directional Sobel operators in the rtMRI video frames in [20]. Many ATB segmentation methods were proposed using the composite analysis grid line superimposed on each rtMRI frame [21, 22, 23, 24]. The Maeda Grid approach (MG) [21] achieved the best performance among all unsupervised approaches. Advaith et al. proposed a Fisher discriminant measure based supervised approach (SFDM) for ATB prediction [25]. Valliappan et al. [26] used a semantic segmentation approach with fully convolutional networks (SFCN).

In this work, we consider a supervised approach and propose a deep learning based contour detection method for ATB segmentation within the vocal tract. The proposed approach uses a convolutional encoder-decoder network (CEDN) [27] which provides state-of-the-art performance compared to the other approaches such as DeepEdge [28] and DeepContour [29]. Due to its supervised nature, the proposed approach is robust to imaging artifacts and grainy noise which pose challenges for naive low-level gradient-based approaches. The CEDN model treats the ATB segmentation as an image labelling problem where a pixel is classified as one if the ATB traces through that pixel otherwise the pixel is classified as zero. Due to high contrast (tissue to air cavity) on the ATBs, the CEDN model learns the intensity variation from tissue to the airway cavity region and labels the pixels accordingly. The output images of the CEDN are further processed to get smooth ATBs. The performance of the proposed approach is evaluated using Dynamic Time Warping (DTW) [30] distance between the predicted ATBs and the manually annotated ground truth ATBs. Lower DTW distance indicates a better performance. In this work, both seen and unseen subject experiments are done to analyze the performance of the proposed approach compared to the baseline schemes: SFCN, SFDM and MG. In seen subject case, test data consists of the trained subject's images whereas in unseen subject case, test data consists of images from a new subject not included in training. In the seen subject experiments, the average DTW score of the predicted contours using the proposed approach is found to be 8.5% lesser than

that using the MG scheme whereas it performs better than SFDM and SFCN approaches only for lower ATBs. Due to the supervised nature, the SFCN and SFDM approaches also perform better than the MG approach in seen subject experiments. However, they do not yield satisfactory performance in unseen subject experiments due to the mismatch in the vocal tract morphology of the training and test subjects. Interestingly, in the unseen subject experiments, the average DTW distance using CEDN based approach is 63.96%, 22.9% less than the SFCN and SFDM approaches respectively, although it fails to perform better than MG scheme. However, when the trained CEDN model is adapted using only 30 images from the unseen subject, the proposed scheme outperforms the baseline MG scheme. From the unseen subject experiment, it can be inferred that the CEDN based approach has better generalizability compared to SFDM and SFCN approaches.

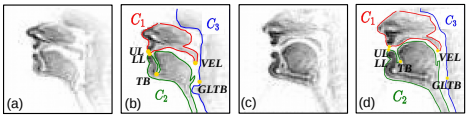

Fig. 1. (a,c) Sample rtMRI Images (b,d) Corresponding ATBs

2. DATASET

For all experiments in this paper, USC-TIMIT [31] corpus is used. The USC-TIMIT is a rich database consisting of rtMRI videos of the upper airway in the mid-sagittal plane. The rtMRI videos are recorded at 23.18 frames/sec. The database contains the videos of five female and five male subjects speaking 460 sentences from MOCHA-TIMIT database [32]. The rtMRI video frames have a spatial resolution of 68×68 pixels (each pixel having dimension of $2.9mm \times 2.9mm$). A total of four subjects - two female (F1, F2) and two male (M1, M2) - are used for the experiment in this work. 16 videos (one for each sentence) from each of these subjects are considered. The chosen 16 videos have 1463, 1272, 1642, 1402 frames from F1, F2, M1, M2 subjects respectively. The ATBs were drawn manually in each rtMRI frame using a MATLAB based graphical user interface (GUI). Using the GUI, three contours were manually annotated in each rtMRI frame. Two such illustrative frames with corresponding ATBs are shown in Figure 1. 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. The three manually annotated complete ground truth contours are denoted as C_1, C_2 and C_3. As shown in Figure 1, C_1 is a closed contour starting from upper lip (UL), which after passing through hard palate, joins velum (VEL) and goes around the fixed nasal tract. C_2 is also a closed contour covering the jawline, lower lip (LL), tongue blade and extends below the epiglottis. C_3 covers the pharyngeal wall.

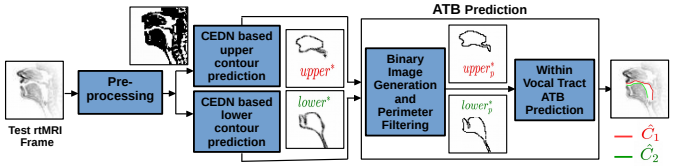

Fig. 2. Illustration of steps in the proposed CEDN based ATB segmentation

3. PROPOSED CEDN BASED ATB SEGMENTATION

The steps of the proposed method of ATB segmentation are explained in Figure 2. The test rtMRI image is preprocessed before ATB prediction. For upper and lower contours, two different CEDN models are trained. The output probability images from the CEDN models are processed to obtain binary images on which perimeter filtering is applied. From the filtered binary images, ATBs within vocal tract are predicted.

3.1. Preprocessing

The rtMRI frames are enhanced using the image processing technique used in the work by Kim et al. [21] to reduce the image artifacts for better performance of the ATB segmentation. Figure 2 shows example of an enhanced rtMRI frame after the preprocessing block.

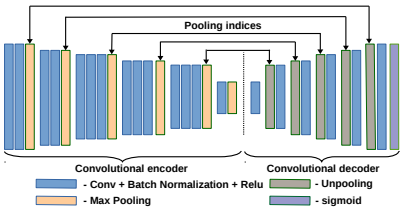

Fig. 3. Block diagram of CEDN architecture

3.2. CEDN based contour prediction

In this paper, we have used CEDN which is a deep convolutional neural network with asymmetric encoder-decoder architecture [27]. The CEDN model provides state-of-the-art performance for object contour detection compared to DeepEdge [28] and DeepContour [29]. The CEDN model generalizes better to unseen object classes than the DeepEdge and DeepContour. Due to having less number of deconvolutional layers in the decoder network, CEDN produces accurate label boundaries with limited number of training images. And CEDN preserves the spatial information by generating the output image of dimension identical to that of the input image dimension. Figure 3 shows the encoder-decoder architecture of the CEDN network used in this work. Typically in CEDN [27], the encoder weights are initialized with VGG-16 weights [33] and are fixed during training. The rtMRI images are different from the image dataset used in VGG-16 training. Thus, in this work, both encoder and decoder weights of the CEDN model are learned during training without utilizing VGG-16 weights as encoder weights. Two CEDN models are trained for upper and lower contours C_1, C_2 separately using the preprocessed input images and the ground truth binary images which are obtained from the manually annotated ground truth contours (as in Figure 4 b, d). The trained CEDN model generates a probability image for a test rtMRI image. The last layer of the CEDN network is a sigmoid layer which generates the probability image in which the pixel value ranges from zero to one. A pixel value of one indicates the most probable ATB pixel and zero indicates the least probable ATB pixel. Figure 4 c, e show the upper and lower probability images ($upper^*$ and $lower^*$) for a test rtMRI image.

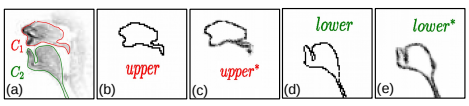

Fig. 4. (a) Manually annotated ground truth contours (b,d) Ground truth binary images (c,e) Probability images from CEDN model for upper and lower contour respectively

3.3. ATB Prediction

This step generates the upper and lower ATBs within the vocal tract from $upper^*$ and $lower^*$ images respectively.

3.3.1. Binary Image Generation and Perimeter Filtering

The $upper^*$ and $lower^*$ images are thresholded to obtain the binary images as well as a fixed contour representing predicted C_3. The best threshold value is decided based on the performance on the validation data. The images obtained by thresholding $upper^*$ and $lower^*$ are denoted as $upper_b^*$ and $lower_b^*$ respectively. The $upper_b^*$ and $lower_b^*$ images are then passed through a perimeter filter which generates filtered binary images that contain only the perimeter pixels of the detected closed ATB in the input binary images. A pixel is considered to be part of the perimeter if the pixel is non-zero and it is connected to at least one zero-valued pixel with the given connectivity. In this work, 4-connectivity is considered which means that two adjoining pixels are part of the ATB if both the pixels are connected along the horizontal and vertical directions and their values are one. The output images of the perimeter filter are denoted as $upper_p^*$ and $lower_p^*$ for the given input binary images $upper_b^*$ and $lower_b^*$ respectively. It should be noted that, as contour C_3 does not change across rtMRI frames, a fixed contour is used as predicted C_3 in all frames. Figure 5 shows $upper_b^*$, $lower_b^*$, $upper_p^*$ and $lower_p^*$ images for a sample rtMRI image.

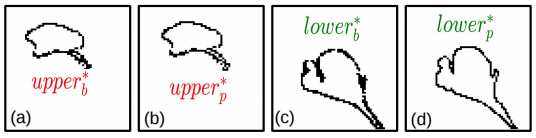

Fig. 5. (a,c) Binary images after thresholding (b,d) Output images from perimeter filter for upper and lower contours respectively

3.3.2. Within Vocal Tract ATB Prediction

This step predicts ATBs within vocal tract from $upper_p^*$ and $lower_p^*$ images as well as a fixed contour representing predicted C_3. For obtaining the contour coordinates from $upper_p^*$ and $lower_p^*$, the indices of the pixels with value one are considered and are sorted in the clockwise direction. The obtained ATBs are pruned to predict the ATBs within the vocal tract following the contour pruning method as described in the SFDM approach [25]. The pruned ATBs are jagged because of the binary thresholding and perimeter filtering. To obtain the smooth contours, the pruned ATBs are passed through a moving average filter with size $q \times q$ (optimum value of q ranges from 3 to 9). The value of q is decided based on the performance on the validation data set. The contours obtained after moving average smoothing are denoted as $\hat{C}_1^p$ and $\hat{C}_2^p$ respectively.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

In this work, experiments are performed in two conditions: 1) seen subject condition 2) unseen subject condition. The unseen subject experiments are done to analyze the generalizability of the CEDN, SFCN and SFDM approaches. The ground truth binary images for both experiments are generated from the manually annotated ground truth contours. In the ground truth binary image, a pixel is labelled as one if the manually annotated contour traverses through that pixel otherwise the pixel is labelled as zero. Figure 4(b) and 4(d) show the upper and lower ground truth binary images ($upper$ and $lower$) for a typical rtMRI image respectively. Experimental setups for the seen and unseen subject conditions are explained below.

Seen subject: In this experiment, a CEDN model is trained and validated using four-fold cross-validation by choosing the videos in a round robin fashion. In each fold, for training the CEDN model, a total of 32 videos comprising 8 videos from each subject (F1, F2, M1, and M2) are considered. For validating, a total of 16 videos comprising 4 videos from each subject are considered. And for testing, a total of 16 videos comprising 4 videos from each subject are considered. Each fold, on average, contains $\sim$2900 training images, $\sim$1443 images in both validation and test sets from all four subjects. The CEDN model is trained for a maximum of 30 epochs by imposing early stopping condition based on the validation loss.

Unseen subject: In this experiment, the CEDN model is trained and tested using four-fold cross-validation by choosing the subjects (F1, F2, M1, M2) in a round robin fashion. In each fold, for training the CEDN model, a total of 48 videos from 3 subjects are considered. For testing, 16 videos from one subject are considered. Each fold, on average, consists of $\sim$4334 and $\sim$1444 images in training and test sets respectively. The CEDN model is trained for 50 epochs.

Adaptation using unseen subject's data: To find out the minimum number of images from the unseen subject required to achieve better performance than the MG scheme, in each fold, the trained model is adapted with P many frames of the unseen subject. P is varied over the following values - 0, 10, 20, 30. The weights of the last five deconvolutional layers of the CEDN model are only learned during adaptation while other layers' weights are kept fixed. To select the adaptation images, the frames from the unseen subject are split into two halves. From the first part, the adaptation images are considered from a set of 100 images and remaining images are used as validation data. The second part is used as the test data.

Performance metric: The performance of the proposed approach is evaluated by finding the DTW distance [30] between the predicted contours and the ground truth contours. The DTW distance has a unit of pixel. In order to obtain the evaluations for $\hat{C}_1^p$ and $\hat{C}_2^p$ contours, the complete ground truth contours C_1 and C_2 are pruned using a method followed in SFDM approach [25] which are denoted as C_1^p and C_2^p respectively. The DTW distance between C_1^p and $\hat{C}_1^p$ ($\mathcal{D}(C_1^p, \hat{C}_1^p)$) and between C_2^p and $\hat{C}_2^p$ ($\mathcal{D}(C_2^p, \hat{C}_2^p)$) are computed following the definition used in the work by Advaith et al. [25].

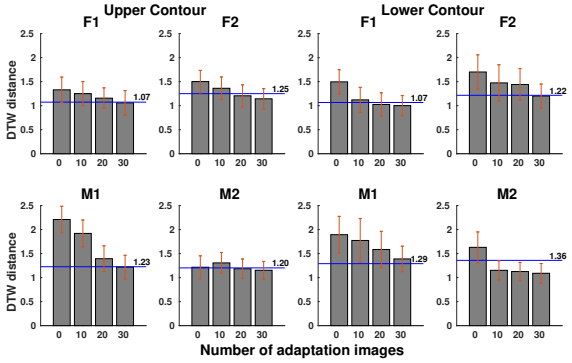

Fig. 6. Bar plot showing DTW distance using CEDN approach for varying number of adaptation images. Errorbar indicates the std. Blue horizontal line indicates the DTW distance using MG scheme.

4.2. Results and Discussions

Table 1 shows the average ($\pm$ standard deviation (std)) of $\mathcal{D}(C_1^p, \hat{C}_1^p)$ and $\mathcal{D}(C_2^p, \hat{C}_2^p)$ using the MG, SFDM, SFCN and CEDN approaches for the seen subject experiments. Table 2 shows simi-

Table 1. DTW distance in pixel (average $\pm$ std) of upper contours and lower contours using MG, SFDM, SFCN and CEDN for the seen subject experiment (blue colour indicates the least DTW distance)

Sub	Upper Contour				Lower Contour			
	MG	SFDM	SFCN	CEDN	MG	SFDM	SFCN	CEDN
F1	1.02 ± 0.19	0.94 ± 0.17	$\mathbf{0.91 \pm 0.21}$	0.95 ± 0.16	1.21 ± 0.21	1.10 ± 0.24	1.00 ± 0.25	$\mathbf{0.94 \pm 0.22}$
F2	1.24 ± 0.29	1.16 ± 0.19	$\mathbf{1.08 \pm 0.19}$	1.13 ± 0.20	1.28 ± 0.27	1.24 ± 0.25	1.13 ± 0.31	$\mathbf{1.12 \pm 0.24}$
M1	1.10 ± 0.20	1.11 ± 0.20	$\mathbf{1.02 \pm 0.20}$	1.14 ± 0.23	1.26 ± 0.60	1.17 ± 0.26	1.17 ± 0.25	$\mathbf{1.16 \pm 0.23}$
M2	1.19 ± 0.24	1.11 ± 0.23	$\mathbf{1.09 \pm 0.21}$	1.17 ± 0.22	1.35 ± 0.30	1.16 ± 0.41	1.21 ± 0.23	$\mathbf{1.14 \pm 0.27}$
Avg	1.13 ± 0.23	1.08 ± 0.20	$\mathbf{1.03 \pm 0.20}$	1.10 ± 0.20	1.27 ± 0.36	1.14 ± 0.29	1.13 ± 0.26	$\mathbf{1.09 \pm 0.24}$

Table 2. DTW distance in pixel (average $\pm$ std) of upper contours and lower contours using MG, SFDM, SFCN and CEDN for the unseen subject experiment (blue and green colours indicate first and second least DTW distances respectively)

Sub	Upper Contour				Lower Contour			
	MG	SFDM	SFCN	CEDN	MG	SFDM	SFCN	CEDN
F1	$\mathbf{1.02 \pm 0.19}$	1.54 ± 0.58	3.32 ± 0.40	$\mathbf{1.28 \pm 0.29}$	$\mathbf{1.21 \pm 0.21}$	1.78 ± 0.43	15.9 ± 0.98	$\mathbf{1.53 \pm 0.26}$
F2	$\mathbf{1.24 \pm 0.29}$	2.28 ± 0.42	1.85 ± 0.31	$\mathbf{1.69 \pm 0.24}$	$\mathbf{1.28 \pm 0.27}$	2.60 ± 0.68	15.4 ± 1.31	$\mathbf{1.76 \pm 0.37}$
M1	$\mathbf{1.10 \pm 0.20}$	$\mathbf{1.68 \pm 0.48}$	4.01 ± 0.41	2.32 ± 0.47	$\mathbf{1.26 \pm 0.60}$	2.17 ± 0.51	9.85 ± 0.83	$\mathbf{1.99 \pm 0.40}$
M2	$\mathbf{1.19 \pm 0.24}$	4.00 ± 0.39	1.67 ± 0.27	$\mathbf{1.20 \pm 0.19}$	$\mathbf{1.35 \pm 0.30}$	1.75 ± 1.36	12.7 ± 0.85	$\mathbf{1.59 \pm 0.23}$
Avg	$\mathbf{1.13 \pm 0.23}$	2.34 ± 0.47	2.79 ± 0.35	$\mathbf{1.65 \pm 0.30}$	$\mathbf{1.27 \pm 0.36}$	2.06 ± 0.78	13.3 ± 0.98	$\mathbf{1.72 \pm 0.32}$

lar results for the unseen subject experiments. From Table 1, it is observed that the CEDN approach results in less average DTW distance compared to MG, SFDM and SFCN approaches for lower contour whereas, for the upper contour, the CEDN approach results in less average DTW distance compared to MG scheme only but doesn't perform better than the SFDM and SFCN approaches. The average DTW distance using CEDN approach is 8.5% lesser than the MG scheme for both upper and lower contours. From Table 2, it is observed that the average DTW distance using CEDN approach is 63.96%, 22.9% less than the SFCN and SFDM approaches but doesn't perform better than MG scheme.

Table 3. DTW distance in pixel (average $\pm$ std) using MG and CEDN (in unseen subject experiments with 30 adaptation images) for test data (blue colour indicates the least DTW distance)

Sub	Upper Contour		Lower Contour	
	MG	CEDN	MG	CEDN
F1	1.03 ± 0.27	$\mathbf{1.02 \pm 0.20}$	1.04 ± 0.21	$\mathbf{1.00 \pm 0.21}$
F2	1.20 ± 0.24	$\mathbf{1.16 \pm 0.22}$	1.32 ± 0.25	$\mathbf{1.21 \pm 0.25}$
M1	1.23 ± 0.19	$\mathbf{1.21 \pm 0.24}$	$\mathbf{1.19 \pm 0.53}$	1.44 ± 0.26
M2	1.20 ± 0.24	$\mathbf{1.18 \pm 0.20}$	1.30 ± 0.26	$\mathbf{1.00 \pm 0.14}$
Avg	1.17 ± 0.23	$\mathbf{1.14 \pm 0.20}$	1.21 ± 1.21	$\mathbf{1.17 \pm 0.21}$

During the CEDN model adaptation using unseen subject's data, DTW distance on the validation data (as explained in 4.1) is computed using 0, 10, 20, and 30 adaptation images. These DTW distances are shown in Fig. 6 for all four subjects separately for upper and lower ATBs. For upper contour, the average ($\pm$ std) DTW scores for validation data across all subjects using 20 and 30 adaptation images are 1.24 ± 0.23 and 1.14 ± 0.22 respectively. Similarly for lower contour, the average ($\pm$ std) DTW scores for validation data across all subjects using 20 and 30 adaptation images are 1.30 ± 0.29 and 1.18 ± 0.23 respectively. Using MG scheme, the average ($\pm$ std) DTW distances are 1.19 ± 0.26 and 1.23 ± 0.34 for upper and lower contours respectively. As the CEDN models yield better validation data performance with 30 adaptation images, the CEDN models adapted with 30 unseen images are considered for evaluation on the test data. The average ($\pm$std) DTW scores on the test data are provided in Table 3. From Table 3, it is clear that the adapted CEDN models perform better than MG scheme even on the test set of the unseen subject. The average ($\pm$ std) DTW distances on test data using SFDM approach with 30 adaptation images are 1.15 ± 0.22 and 1.81 ± 1.00 for upper and lower contours respectively. Thus, the SFDM approach with 30 adaptation images performs better than the MG scheme for upper contour (while still worse than CEDN) but it fails to perform better than the MG scheme for lower contour. Unlike CEDN, we observe that SFCN with 30 adaptation images does not perform better than MG scheme.

The superior performance of the proposed CEDN based approach could be due to the following reasons: 1) Because of its supervised nature, the CEDN approach predicts the reliable contours by overcoming the imaging artifacts and grainy noise. 2) The light decoder (less number of deconvolutional layers) of the CEDN and learning both encoder and decoder weights during training help in predicting accurate boundaries from the limited number of training images in both unseen and seen subject experiments. 3) The perimeter filter used in the post-processing helps in getting precise boundary pixels from the binary thresholded image. From the experimental results, it is observed that the CEDN approach does not perform better in some cases for upper contour predictions. The CEDN model is often found to predict a cluster of points at the velum region instead of predicting a smooth boundary. Due to this, in the contour pruning step, velum point is not detected properly leading to inaccurate upper ATBs in some cases.

5. CONCLUSION

In this paper, a supervised approach using CEDN model is proposed for the ATB segmentation. The proposed approach performs better than the baseline MG scheme in the seen subject experiment. In the unseen subject condition, the approach achieves better performance than the MG scheme with only 30 unseen subject's images as adaptation data. The proposed method has better generalization ability compared to the other supervised approaches like SFCN and SFDM. The performance of the proposed approach can be improved further by using adaptive thresholding to generate binary images from CEDN output probability images. In adaptive thresholding, the threshold value for a pixel in an image is decided by statistically examining the intensity values of the local neighbourhood of that pixel.

Acknowledgement: We thank the Pratiksha Trust for their support.

6. REFERENCES

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

[2] D. Maurer, B. Grne, T. Landis, G. Hoch, and P. W. Schnle, "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.

[3] 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.

[4] 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.

[5] Nimisha Patil, Timothy Greer, Reed Blaylock, and Shrikanth S. Narayanan, "Comparison of basic beatboxing articulations between expert and novice artists using real-time magnetic resonance imaging," in *Proc. Interspeech 2017*, pp. 2277–2281.

[6] Asterios Toutios, Tanner Sorensen, Krishna Somandepalli, Rachel Alexander, and Shrikanth S. Narayanan, "Articulatory synthesis based on real-time magnetic resonance imaging data," in *Interspeech 2016*, pp. 1492–1496.

[7] Brad H Story, Ingo R Titze, and Eric A Hoffman, "Vocal tract area functions from magnetic resonance imaging," *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] 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.

[11] Fang-Ying Hsieh, Louis Goldstein, Dani Byrd, and Shrikanth Narayanan, "Pharyngeal constriction in English diphthong production," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, pp. 060271, 2013.

[12] 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, pp. 4265–4269.

[13] 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.

[14] Asterios Toutios and Shrikanth Narayanan, "Factor analysis of vocaltract outlines derived from real-time magnetic resonance imaging data," in *18th International Congress of Phonetic Sciences, ICPhS 2015, Glasgow, UK, August 10-14, 2015*, 2015.

[15] Tanner Sorensen, Asterios Toutios, Louis Goldstein, and Shrikanth S. Narayanan, "Characterizing vocal tract dynamics across speakers using real-time MRI," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 2016, pp. 465–469.

[16] 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.

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

[18] 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, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 1572–1575.

[19] 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, pp. 959–962.

[20] 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)*, 2016, pp. 937–941.

[21] 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, pp. 222 – 225.

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

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

[24] 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 *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pp. 1576–1579.

[25] 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)*, April 2018, pp. 5004–5008.

[26] 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*, 2018, pp. 3132–3136.

[27] Jimei Yang, Brian Price, Scott Cohen, Honglak Lee, and Ming-Hsuan Yang, "Object contour detection with a fully convolutional encoder-decoder network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 193–202.

[28] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani, "Deepedge: A multi-scale bifurcated deep network for top-down contour detection," *CoRR*, vol. abs/1412.1123, 2014.

[29] Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, and Zhijiang Zhang, "Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[30] Donald J. Berndt and James Clifford, "Using dynamic time warping to find patterns in time series," in *in The 3rd International Conference on Knowledge Discovery and Data Mining*, 1994, AAAIWS'94, pp. 359–370.

[31] 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*, vol. 136, no. 3, pp. 1307–1311, 2014.

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

[33] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.