Airway segmentation in speech MRI using the U-net architecture

Subin Erattakulangara[1], *Sajan Goud Lingala*[1,2]
[1]Roy J Carver Department of Biomedical Engineering
[2]Department of Radiology
The University of Iowa, Iowa city

ABSTRACT

We develop a fully automated airway segmentation method to segment the vocal tract airway from surrounding soft tissue in speech MRI. We train a U-net architecture to learn the end to end mapping between a mid-sagittal image (at the input), and the manually segmented airway (at the output). We base our training on the open source University of Southern California's (USC) speech morphology MRI database consisting of speakers producing a variety of sustained vowel and consonant sounds. Once trained, our model performs fast airway segmentations on unseen images at the order of 210 ms/slice on a modern CPU with 12 cores. Using manual segmentation as a reference, we evaluate the performances of the proposed U-net airway segmentation, against existing seed-growing segmentation, and manual segmentation from a different user. We demonstrate improved DICE similarity with U-net compared to seed-growing, and minor differences in DICE similarity of U-net compared to manual segmentation from the second user.

1. INTRODUCTION

Magnetic resonance imaging of vocal tract shaping is emerging as a powerful tool to non-invasively assess speech production. Its utility is growing in both basic speech science and clinical applications. These include understanding phonetics, providing new insights into language production, modeling speech, assessing movement disorders, assessing speech, pre and post oral cancer treatment [1]–[3]. Several researchers have recently applied sparse sampling and constrained reconstruction schemes to significantly improve speech MRI. These include enabling rapid dynamic 2D MRI of free-running speech (eg. at frame rates of ~100 frames/sec [4]), rapid volumetric 3D scans of sustained speech (scan time of ~6 sec) [5] and dynamic volumetric 3D scans of free-running speech at ~20 frames/sec [6].

While high-speed MRI has dramatically improved the richness of speech MRI datasets, they have also introduced challenges with large scale segmentation, where manual segmentation would be timeconsuming, and not practical. Segmentation in speech MRI range from (a) segmenting the airway from the surrounding tissue, (b) segmenting the air-tissue interfaces or (c) segmenting the articulators themselves (eg. the tongue, velum, pharyngeal wall). Various segmentation methods have been proposed for these tasks. For instance, the classical seed growing algorithm has been applied to semi-automatically segment the airway in [7], and the tongue in [8]. A subjectspecific anatomical template-based approach was proposed in [9], where contours of moving air-tissue interfaces of various articulators were segmented. Recently, these contours have been segmented using automated network-driven segmentation tools (eg. [10], [11]).

In this paper, we develop an airway segmentation network based on the original U-net architecture[12]. We leverage speech MRI datasets from the publicly available University of Southern California's (USC) speech MRI morphology database [13]. We manually annotate/segment the airway in mid-sagittal cross sections of speech MR images obtained from several speakers producing a variety of sustained vowel and consonant sounds. The U-net architecture is trained to learn the end to end mapping between the mid-sagittal image (at the input) and the segmented airway (at the output). Once trained, the model performs fast airway segmentations on unseen images at the order of 210ms/slice on a CPU with 12 cores. Using manual segmentation as the reference, we compare the performances of the proposed U-net airway segmentation against existing seed-growing based airway segmentation, and a second manual segmentation from a different user. We demonstrate U-net to provide considerably improved DICE similarity compared to seed-growing segmentation and has minor differences in DICE similarity when compared to manual segmentation from a second user.

2. METHODS

2.1 Speech MRI dataset specifications:

We used the USC speech morphology database which contained volumetric scans of the airway during sustained production of vowel and consonant sounds[13]. Imaging was performed with a head coil, and the parameters were FOV: 20x20x10 cm^3; resolution: 1.25mm^3; flip angle: 5 degrees; scan time: 6 sec; acceleration factor of 6. Images were reconstructed via a spatial total variation based constrained reconstruction. Here, we considered the 2D problem of segmenting the airway from the mid-sagittal section of these volumetric scans. A total of 100 images from 10 speakers were used in this study with a split of 75 images for training; 5 images for validation; and 20 images for testing. Few images with significant total variation staircase blurring artifacts across the air-tissue boundary were omitted from the original database. We considered images where manual segmentation of the airway was feasible with a single connected mask having a diameter of atleast 2 pixels. This was done so that the network is robust to learning disconnected false-positive airway masks such as airspace behind the velum, airspace in the nose, dark pixels representing the teeth. With this selection criterion, the pruned database contained the following stimuli of "bit, bait, bet, bat, pot, but, bought, boat, boot, put, bird, abbot, afa, ava, atha, aha, ama, ana", where the underlined text represents the sustained sounds. Figure 1 shows representative speech sounds across all the 10 speakers used in this study.

2.2 Pre-processing:

The reconstructed mid-sagittal images had a considerable smooth intensity variation across the FOV (largely due to sensitivity variation of the head coil elements across the large FOV). The images were corrected for this intensity bias by accounting for the smooth bias field, which was estimated via a low pass operation on the uncorrected image. The low pass filter employed a Gaussian filter with a kernel size of 50 pixels. The bias correction leads to noise amplification in low intensity regions. To account for this, we performed de-noising using a spatial total variation regularizer. The intensity corrected de-noised images were then cropped to zoom into a pre-defined region of interest containing the vocal tract airway and the neighboring articulators. Finally, the cropped images were resized to a size of 256 px × 256 px using the imresize command in MATLAB with default bi cubic interpolation.

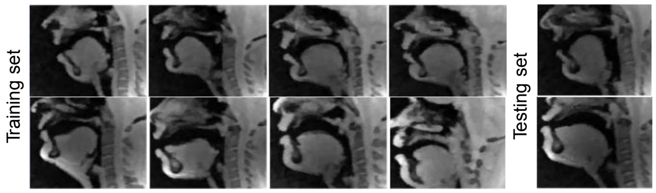

Figure 1: Datasets used in this study were from the USC-speech morphology database. Shown here are example vocal tract poses from 8 speakersin training set, and 2 speakers in testing set.

2.3 U-net architecture and training:

We used the original U-net architecture[12] to learn the mapping of the pre-processed mid-sagittal image at the input and the airway segmentation (represented as a binary mask) at the output (see Figure 2). The architecture was implemented in Keras with TensorFlow backend. All our experiments were performed on an Intel Core-i7 8700CK, 3.70 GHz 12 core CPU machine. The number of base feature maps per convolutional layer in the first resolution scale was 64, which was doubled and halved in the next resolution scale in the contracting and expanding U-net paths respectively.

We used the following binary cross-entropy loss function to train the U-net model:

$$\mathrm{L_{BCE}} = -\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y_{ref}(i)}.\log(\hat{\mathrm{y}}(\mathrm{i})) + (1-\mathrm{y_{ref}(i)}).\log(\hat{\mathrm{y}}(\mathrm{i}))); (1)$$

Where $\mathrm{y_{ref}}(\mathrm{i})$ and $\hat{y}(i)$ are the intensities of the reference and the predicted segmented masks at the $(\mathrm{i}, \mathrm{j})^{\mathrm{th}}$ pixel and N is the total number of pixels in each image. The reference mask was obtained from manual segmentation. As described in section 2.1, the number of training image pairs $\{\mathrm{y_{ref}}, \hat{y}\}$ across speakers producing a variety of consonant and vowel sounds were 75. We performed data augmentation on these images to increase the training set by 4-fold using random operations of rotation, scaling, and shifting. Other relevant training parameters were number of epochs = 100, batch size = 2, learning rate =0.0001, dropout rate = 0.5 at the fourth and fifth convolution operations and a choice of the adaptive moment estimation (ADAM) optimizer. The total training time was 20 hours.

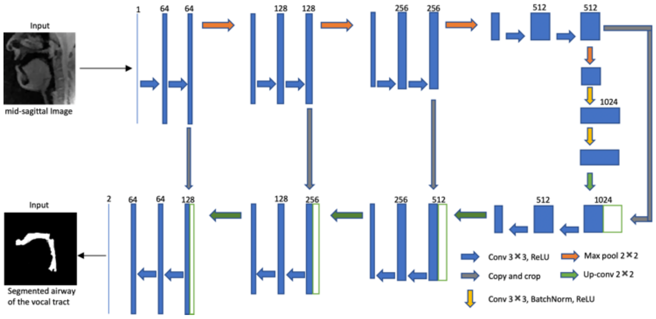

Figure 2: U-net architecture. Each blue box represents a multi-channel feature map with number of features represented on top of the box. White boxes correspond to copied feature maps. Different color arrows denote different operations as listed in the legends on the bottom right.

2.4 Evaluation:

For evaluating the performance of U-net against seed-growing and manual segmentation, we used a total of 20 sustained vowel and consonant sounds from the two speakers in our test set. Manual segmentation obtained from user-1 was considered as the reference mask. The performance of the proposed automatic U-net airway segmentation, an existing seed-growing airway segmentation, and a second manual segmentation from a different user (termed as user-2) were compared. The performances were evaluated in terms of DICE similarity *(D)* of the estimated segmentation mask ($\hat{y}$) with the reference mask (y_{ref}):

$$D = \frac{2(|y_{ref}| \cap |\hat{y}|)}{|y_{ref}| + |\hat{y}|}$$

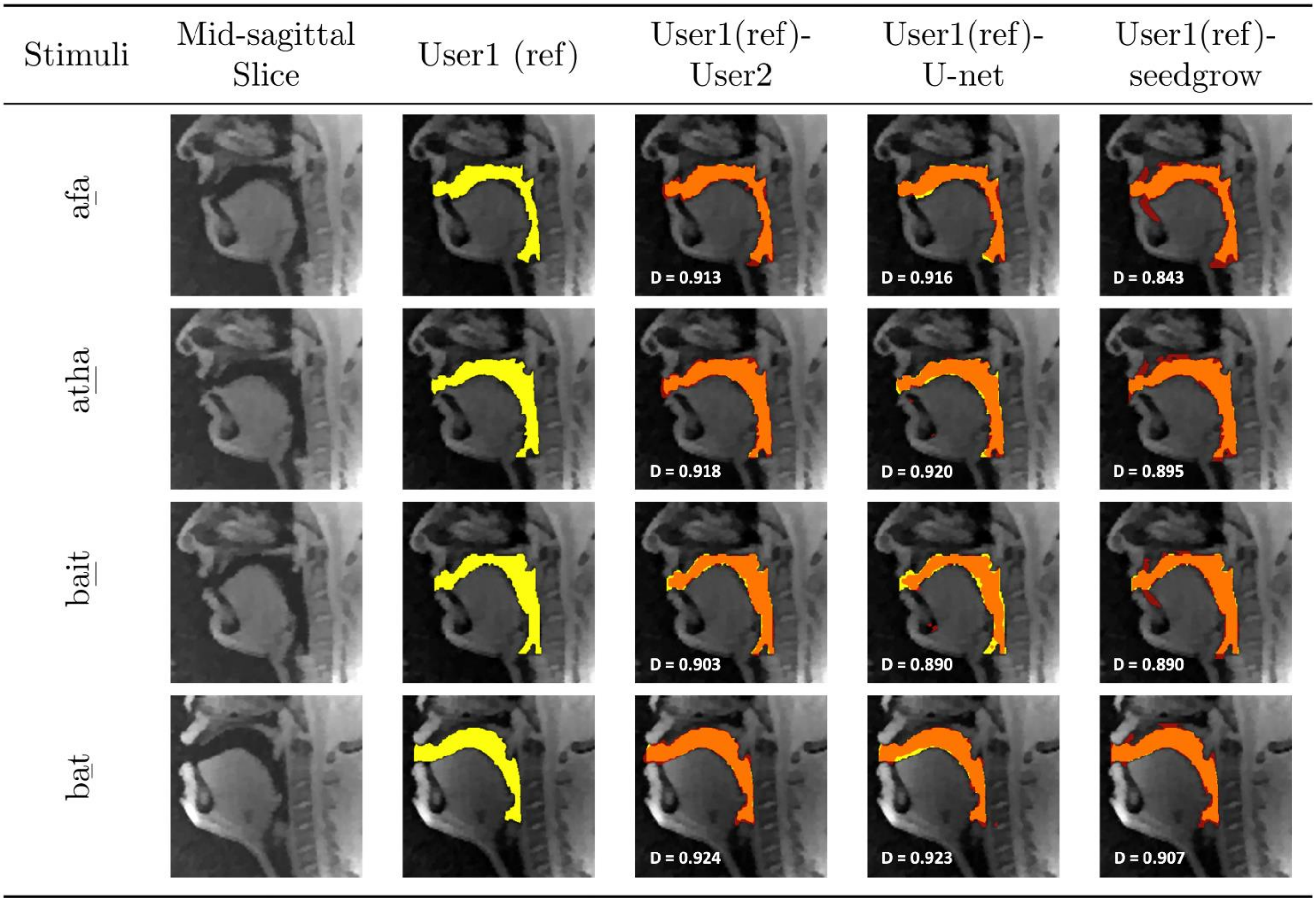

Figure 3: Vocal tract airway segmentations on test data. Example stimuli from two consonant sounds (afa, atha), and two vowel sounds (bait, bat) are shown. Reference manual segmentation from user-1 (second column) are overlaid on the segmentations from manual segmentation from user-2 (third column), proposed U-net segmentation (fourth column), and the seed-growing segmentation (fifth column). Each inset from the three segmentations also show the corresponding DICE similarity with the reference segmentation.

The seed-growing algorithm was implemented in MATLAB. For every test image, we manually specified the location of the initial seed in a square region of interest (ROI) whose boundaries were defined to be the beginning of the lips, top of the hard palate, the pharyngeal wall, and the bottom of the glottis. The average processing times of all the U-net, seed-growing, and the manual segmentations were recorded and compared on the test set. For seed-growing, the processing time also included the time spent for specifying the seed, and drawing the square ROI.

3. RESULTS AND DISCUSSION

Figure 3 shows representative segmentations of consonant sounds (atha, afa), and vowel sounds (bait, bat) from the test set. The reference segmentations from user-1 are overlaid with the segmentations from the three schemes of the manual segmentation from user-2, proposed U-net segmentation, and the seed-growing segmentation. The DICE coefficients are shown on the images. While seed-growing segments the vocal airspace, it is sensitive to leaking into airspaces beyond the vocal tract. These were observed in regions with low contrast between the air-tissue boundary (eg. near the boundaries of the air-hard palate; air-velum). In contrast, U-net segmentation consistently provided a higher DICE similarity, and depicted good segmentation accuracy. The U-net segmentation has subtle differences with manual segmentations from the second user. In some of the images, U-net segmentation was observed to be sensitive to the spatial piece-wise blurring artifacts typical with total variation regularization. This is observed prominently in the bait stimuli shown in the third row, where there are subtle mismatches of the U-net segmentation with both the manual segmentations.

Figure 4 shows the average DICE similarity of all the methods with reference segmentations on all the 20 test images. The U-net segmentation (mean DICE = 0.9) shows a considerable improvement over seed-growing (mean DICE = 0.86), and a minor difference compared to a manual segmentation from user-2 (mean DICE = 0.91).

U-net provided considerably faster processing times to generate the segmentation. The average mean processing times across 20 test images were 0.21sec/image for U-net segmentation, 11.6sec/image for seed-growing segmentation and 45 sec/image for the manual segmentation from user-2.

4. CONCLUSION

We successfully demonstrated airway segmentation of the vocal tract in speech MRI using the U-net architecture. The architecture is trained to learn the end to end mapping between the mid-sagittal image and the segmented airway. Once trained, the model performs fast airway segmentations at the order of 0.21sec/image on a CPU with 12 cores. We demonstrated U-net to provide considerably improved DICE similarity compared to existing seed-growing segmentation, and minor differences in DICE similarity compared to manual segmentation. Future work includes adaptations to training to be robust to typical reconstruction artifacts, learning disjoint airway masks along the vocal tract as occurred in dynamic imaging, extensions to 3D network architectures (eg. as in [14]), and implementations on a GPU for faster training times.

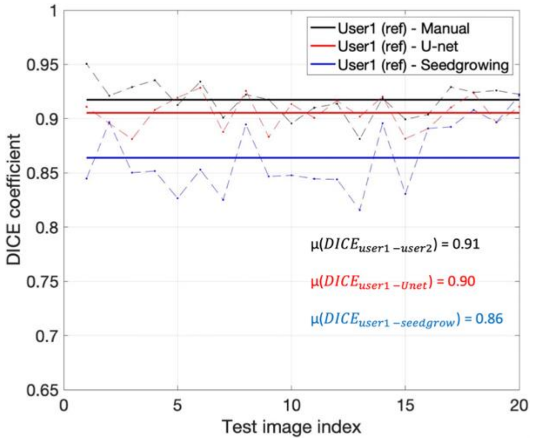

Figure 4: DICE similarity of the reference (user-1) segmentation with segmentations from user-2, U-net, seed-growing across all 20 test images. U-net segmentation has higher mean DICE similarity over seed-growing, and has marginally lower DICE similarity when compared to manual segmentation from user-2.

5. REFERENCES

[1] A. D. Scott, M. Wylezinska, M. J. Birch, and M. E. Miquel, "Speech MRI: Morphology and function," *Phys. Medica*, 2014.

[2] S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, "Recommendations for real-time speech MRI," *Journal of Magnetic Resonance Imaging*, vol. 43, no. 1. pp. 28–44, 2016.

[3] J. L. Perry, B. P. Sutton, D. P. Kuehn, and J. K. Gamage, "Using MRI for assessing velopharyngeal structures and function," *Cleft Palate-Craniofacial J.*, 2014.

[4] P. W. Iltis, J. Frahm, D. Voit, A. A. Joseph, E. Schoonderwaldt, and E. Altenmüller, "High-speed real-time magnetic resonance imaging of fast tongue movements in elite horn players.," *Quant. Imaging Med. Surg.*, vol. 5, no. 3, pp. 374–81, 2015.

[5] S. G. Lingala, A. Toutios, J. Toger, Y. Lim, Y. Zhu, Y. C. Kim, C. Vaz, S. Narayanan, and K. Nayak, "State-of-the-art MRI protocol for comprehensive assessment of vocal tract structure and function," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016.

[6] Y. Lim, Y. Zhu, S. G. Lingala, D. Byrd, S. Narayanan, and K. S. Nayak, "3D dynamic MRI of the vocal tract during natural speech," *Magn. Reson. Med.*, 2019.

[7] Z. I. Skordilis, A. Toutios, J. Toger, and S. Narayanan, "Estimation of vocal tract area function from volumetric Magnetic Resonance Imaging," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017.

[8] Z. I. Skordilis, V. Ramanarayanan, L. Goldstein, and S. S. Narayanan, "Experimental assessment of the tongue incompressibility hypothesis during speech production," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.

[9] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Trans. Med. Imaging*, 2009.

[10] 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 *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.

[11] C. A. Valliappan, R. Mannem, and P. Kumar Ghosh, "Air-tissue boundary segmentation in real-time magnetic resonance imaging video using semantic segmentation with fully convolutional networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015.

[13] T. Sorensen, Z. Skordilis, A. Toutios, Y. C. Kim, Y. Zhu, J. Kim, A. Lammert, V. Ramanarayanan, L. Goldstein, D. Byrd, K. Nayak, and S. Narayanan, "Database of volumetric and real-time vocal tract MRI for speech science," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.

[14] A. G. U. Juarez, H. A. W. M. Tiddens, and M. de Bruijne, "Automatic airway segmentation in chest CT using convolutional neural networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018.