VOCAL TRACT ARTICULATORY CONTOUR DETECTION IN REAL-TIME MAGNETIC RESONANCE IMAGES USING SPATIO-TEMPORAL CONTEXT

S Ashwin Hebbar[2] Rahul Sharma[1] Krishna Somandepalli[1] Asterios Toutios[1] Shrikanth Narayanan[1]

[1]University of Southern California, USA
[2]National Institute of Technology Karnataka, India

ABSTRACT

Due to its ability to visualize and measure the dynamics of vocal tract shaping during speech production, real-time magnetic resonance imaging (rtMRI) has emerged as one of the prominent research tools. The ability to track different articulators such as the tongue, lips, velum, and the pharynx is a crucial step toward automating further scientific and clinical analysis. Recently, various researchers have addressed the problem of detecting articulatory boundaries, but those are primarily limited to static-image based methods. In this work, we propose to use information from temporal dynamics together with the spatial structure to detect the articulatory boundaries in rtMRI videos. We train a convolutional LSTM network to detect and label the articulatory contours. We compare the produced contours against reference labels generated by iteratively fitting a manually created subject-specific template. We observe that the proposed method outperforms solely image-based methods, especially for the difficult to track articulators involved in airway constriction formation during speech.

Index Terms— rtMRI, CNN, convLSTM, segmentation

1. INTRODUCTION

One of the fundamental challenges in understanding the mechanisms of human speech motor control is obtaining accurate information about the movement and shaping of the vocal tract during speech production. Dynamic vocal tract imaging technologies are also crucial for understanding the relationship between speech articulation and acoustics. However, acquiring high-quality data to resolve fine-grained movements of speech articulators and the vocal tract without interfering with speech production remains challenging. In this regard, real-time MRI (rtMRI) has emerged as one of the most appropriate tools to image the entirety of a speaker's vocal airway during speech production [1, 2, 3].

This non-invasive method can capture dynamic information about the coordinated movement of speech articulators such as jaw, lips, tongue, velum, epiglottis, pharyngeal and laryngeal regions [1, 3, 4]. It provides rich spatial information of the entire midsagittal plane at a high temporal resolution of up to 83fps. This allows us to capture detailed articulatory motion at a high temporal resolution. As a result, rtMRI is extensively used in the speech science and linguistic studies to understand the dynamics of speech production across languages, and across health conditions [5]. A crucial step to automate these analyses is the ability to track different articulators such as the tongue, lips, etc. since the shaping of the airway by the articulators, by forming and releasing constrictions (e.g., bringing the tongue tip close to the alveolar ridge to make the sound /s/ or the lips together to make a /b/), is the crucial aspect of speech production. Because rtMRI is typically reconstructed for the midsaggital plane to cover the upper airway maximally, the task of tracking articulators can be formulated as a contour detection problem; the same approach however can be applied to other imaging scan planes as well.

In this work, we address the problem of articulatory contour detection in rtMRI videos of the midsagittal plane of the human vocal system. We primarily focus on detecting 12 key

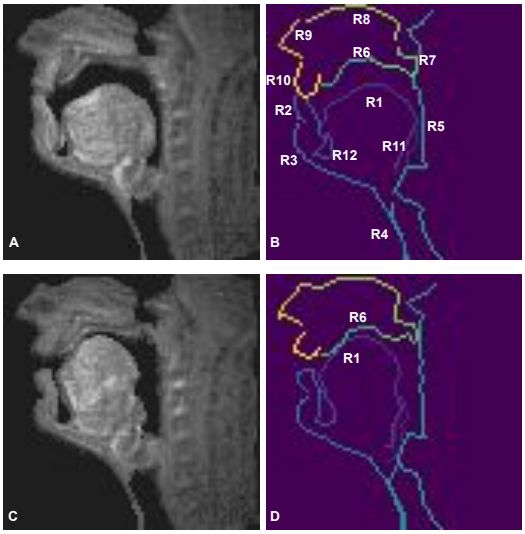

Fig. 1. A,C) Input real-time MR image. B) Vocal tract contour labels: R1 - Tongue, R2 - Lower lip, R3 - Jaw, R4 - Trachea, R5 - Pharynx, R6 - Palate, R7 - Velum, R8 - Nasal cavity, R9 - Nose, R10 - Upper lip, R11 - Epiglottis, R12 - Incisor. D) R1 and R6 involved in constriction

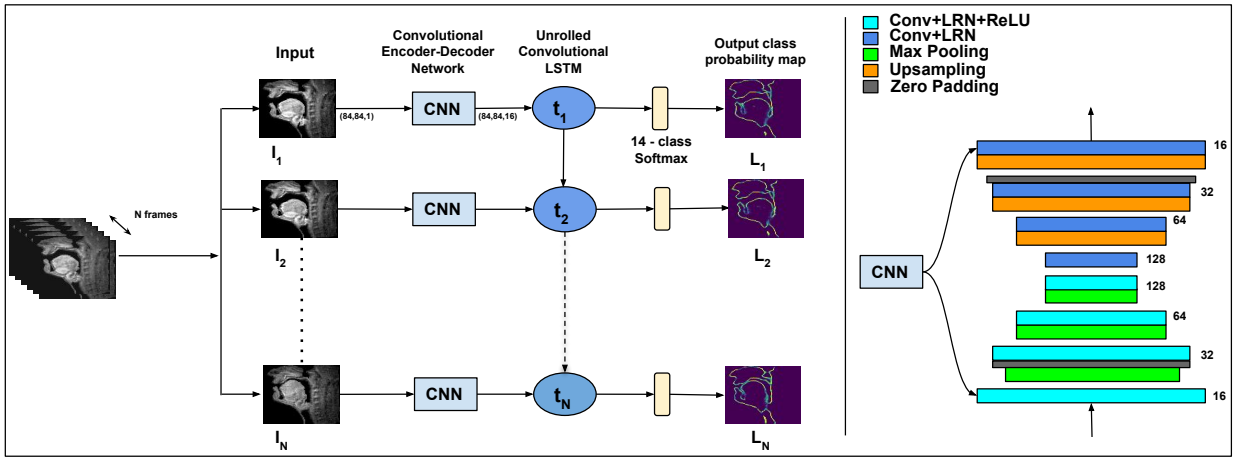

Fig. 2. Complete architecture of the proposed network.

articulator contours, as shown in Figure 1B. The rtMRI data we examine in this work are acquired at a high temporal resolution to resolve finer details of articulatory motion; as a result, the spatial resolution of the images is limited [6] to optimize real-time acquisition. Additionally, these images have a low signal to noise ratio and poor contrast between different tissue or muscle types in the upper airway.

Over the years, several methods have been explored to identify the contours corresponding to articulators of interest [7]. Most of these are methods are semi-automatic and require speaker-dependent information such as speaker-specific templates. Recently, with the advent of Convolutional Neural Networks (CNNs), methods for fast and automatic detection of contours have been explored. Somandepalli et al. [8] used a SegNet CNN architecture [9] with a custom loss function to simultaneously detect and label the articulator contours. Following this, there have been several CNN-based architectures [10, 11, 12] exploring the articulatory contour detection and labeling problem.

The primary limitation of these methods is their inability to detect the air-tissue boundaries when two articulators are involved in constriction since there are no differences in contrast between different articulators. Fig 1D shows the articulators (R1 and R6) involved in constriction. The fundamental component missing in all these approaches is the information from the rich temporal structure that is inherent to articulatory motion. Due to the presence of deformable articulators and that too involved in constriction, it is challenging to detect them without using the temporal context along with the static image frames. In this work, we use Convolutional Long Short Term Memory [13] (ConvLSTM) networks to model the temporal context to track articulatory motion.

There have been several works in the medical imaging domain that employ combinations of CNNs and LSTMs for the application of segmentation. A PyraMiD-LSTM architecture was used in [14] for 3D medical image segmentation. A recurrent fully-convolutional network (RFCN) was used in [15] for multi-slice MRI cardiac segmentation. Bai et al. [16] used a method that uses FCN and C-LSTM layers to learn the spatial and temporal context, respectively, for aortic MR image sequence segmentation. Along a similar direction, we propose to use a fully-convolutional encoder-decoder network for generating frame-level encodings and a convolutional LSTM across time that takes care of the temporal structure.

2. METHODS

We formulate the problem of articulatory boundary detection as a supervised multi-class pixel labeling task. We train a deep neural network in a supervised fashion to classify each pixel into 12 articulator contours (as shown in Figure 1B) or the airway or a tissue, thus formulating it as a 14 class classification problem. The ConvLSTM architecture we employ takes the spatial as well as temporal context into account for detecting the articulatory boundaries.

2.1. Data

We use the dynamic real time magnetic resonance imaging videos that consist of the scans of the human midsagittal plane, recorded while producing running speech. The data we analyze are from 8 individuals (4 female, 4 male) and is made publicly available[1]. The details of the experimental protocol used can be found in [17]. As described in [6], the images were acquired at 83.33 frames per second. We obtain 2D grayscale images of size 84x84 pixels.

The articulatory landmarks for all the MRI scans were produced using the method described in [7]. We followed the method described in [8] to obtain continuous contours as reference (”true”) labels. It is important to note that these ’ground truth labels’ are a noisy approximation of the contours as described in [7].

[1] sail.usc.edu/span/test-retest

2.2. Proposed Architecture

The proposed network consists of two components. The first is the CNN module to account for the spatial structure and the other is an RNN module to incorporate the temporal structure of the speech MRI scans. The CNN module is adapted from the encoder-decoder network architecture proposed in [8] to extract spatial features from the input image frames. We used local response normalization (LRN), and the standard ReLU activation function. We employed a convolutional LSTM (ConvLSTM) layer to connect the obtained CNN representations across time. The motivation behind using the ConvLSTMs is their ability to preserve the spatial structure. Thus the contour probability maps are generated using the spatial as well as temporal context. The architecture is presented in Figure 2. The predicted output is generated by picking the top class at every pixel.

.

3. EVALUATION

3.1. Implementation Details

All our models are implemented using the Keras framework. The input to the network is a video clip of one second duration with each frame being 84×84. The network is trained to minimize the categorical cross-entropy loss. Adadelta optimiser is used with a batch size of 8 video segments. The rtMRI videos have a frame rate of 83.33 fps. While training, we downsampled it by a factor of 3, uniformly picking one of every 3 consecutive frames. Thus each training sample has 28 frames. This serves two purposes: i) Reduces the computational complexity of the network by reducing the number of states of the RNN. ii) Helps in mitigating the effect of the errors in the ground truth labels.

3.2. Results

All experiments were conducted in a leave-one-subject-out (LOSO) fashion, for the eight subjects in our database. We use the SegNet based approach used in [8] as a baseline. To evaluate the performance of our models, we compute the average of the Cityblock distance between every point of the output contour to the true labels, and vice versa. The rectilinear nature of the City-block measure allows us to capture the relative smoothness between two contours. The smaller this measure, the more similar is the output contour to the true labels. Table 1 reports the Cityblock distance averaged over all LOSO models. Figure 3 shows four sample frames with ground truth and the predicted contours.

For each LOSO model, a total of 70 videos (10 per subject) were used for training. As seen in Table 1, the proposed method significantly outperforms the SegNet based implementation. This supports our motivation that the rtMRI frames have a strong temporal context that could be utilized for segmentation. We exclude the incisor contour (R12) from further analysis, as the true labels for this region are not precise. This is due to the fact that the signal acquired in this region through MRI is very low (because teeth and bone structures with low hydrogen content do not image well in MRI). As the variance in the average Cityblock distances were found to be low, we consider only one LOSO model for further analysis.

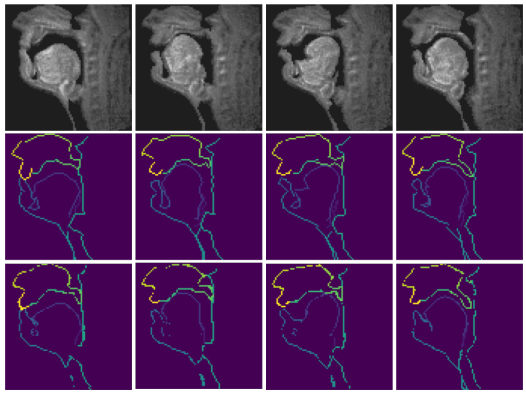

Fig. 3. Row 2: True Contours; Row 3: Predicted Contours Showing the frames with different constrictions

Label	Som. [8]	Ours	Label	Som. [8]	Ours
R1	0.93	1.18	R7	1.92	1.10
R2	1.72	1.19	R8	1.42	1.24
R3	0.96	0.65	R9	1.91	1.00
R4	1.12	0.72	R10	1.52	1.01
R5	1.41	0.79	R11	2.31	1.37
R6	1.70	1.04	Avg	1.54	0.99

Table 1. Average Cityblock distance to the true labels

One of the major shortcomings of all the previous approaches including [8], is the lack of discriminability when two articulators are involved in maximal constriction position, i.e., are in contact (such as in sounds like /p/ and /t/) or very close to one another (such as for /s/ and /sh/). We hypothesize that, provided the temporal information, the network should be able to analyze the motion of the articulators in the temporal neighborhood and thus get a better context to decide the articulator boundaries. To evaluate this hypothesis, we manually annotated, by visual inspection, the rtMRI frames where pairs of articulators are in contact (R2-Lower Lip and R10-Upper Lip, R1-Tongue and R5-Palate, R7 - Velum and R6 - Pharynx) for one of the subjects in the validation set. We then compute the average Cityblock distances for the above articulators across the frames where the respective frames are involved in constriction and the remaining frames, separately. We present the average Cityblock distance for the respective scenarios in tables 2 and 3. We observed that the improvement in the performance of the proposed network over the implementation in [8] is more pronounced in the frames where articulators are involved in constriction.

Label	Somandep [8]	Proposed	% Improvement
R1	1.436	1.344	6.4
R5	1.662	0.708	57.4
R2	2.375	1.759	25.9
R10	0.952	0.645	32.2
R7	1.614	1.003	37.8
Avg	1.609	1.092	32.1

Table 2. Average Cityblock Distances at frames where the respective articulators are involved in constriction

Label	Somandep [8]	Proposed	% Improvement
R1	1.334	1.338	-0.3
R5	1.091	0.751	31.2
R2	1.842	1.333	27.6
R10	0.842	0.664	21.1
R7	1.803	1.588	11.9
Avg	1.382	1.135	17.9

Table 3. Average Cityblock Distances at frames where there is no constriction

4. ABLATION STUDIES

Downsampling: The rtMRI videos, that are generated at 83.33 fps, are downsampled to 28 fps while training (note that speech movement rates on average are about 12 Hz). Since the labels are obtained using the algorithm described in [7], rather than a manual process, we expect the labels are coarse and have some errors. As the frame rate of the rtMRI frames is high, and the videos are natural, thus smooth, it can be assumed that a set of three consecutive frames are visually very close. Therefore, the loss in information while downsampling is nominal. This helps in reducing the substantial adverse effect of the erroneous ground truth. Our experiments supported this hypothesis, showing better performance for the majority of contours when the videos are uniformly downsampled during training when compared with using all the frames for training, as seen in Table 4.

Temporal window size: To provide temporal context, an appropriate window length for input video segments must be chosen. In our experiments, a window size of 1 second was used in the ConvLSTM layer. To see the effect of changing the window size on the performance, we trained a LOSO model with ConvLSTM window sizes of 0.5, 1, and 2 seconds. It can be observed from Figure 4 that the dependency of the performance on the window size various with articulators. One of the reasons may be the different moving speeds of different articulators. For instance, having a longer window may be useful for slow-moving articulators such as the nose (R9), but not for fast-moving contours such as the tongue (R1). On averaging, we observed that 1-second window offers the best trade-off among all articulators.

LSTM vs. BLSTM: It is expected that if we provide 2-way temporal information i.e., forward and reverse sequence, the model should perform better. We replaced the ConvLSTM in our model with Bidirectional ConvLSTM to test this hypothesis. We observed that, due to the increased number of parameters, the training time increased significantly, but improvement in performance was minimal, as seen in Table4.

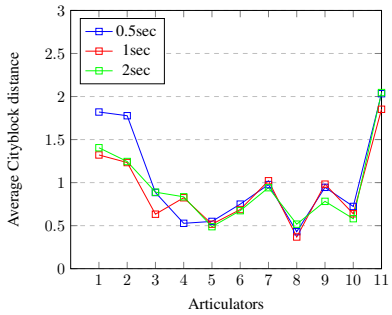

Fig. 4. Evaluation with different temporal windows

Label	ConvLSTM (Without DS)	ConvLSTM (With DS)	Bidirectional ConvLSTM
R1	1.357	1.321	0.981
R2	1.681	1.233	1.390
R3	0.945	0.633	0.741
R4	1.117	0.824	0.626
R5	0.711	0.698	0.451
R6	0.698	0.687	0.839
R7	0.903	1.021	0.844
R8	0.375	0.368	0.616
R9	0.664	0.982	0.796
R10	0.896	0.638	0.545
R11	2.074	1.852	1.574
Avg	1.038	0.916	0.855

Table 4. Comparison of average Cityblock distances when ConvLSTM with and without uniform downsampling, and bidirectional ConvLSTM are used

5. CONCLUSION

In this work, we present a state-of-the-art method for articulatory boundary detection in real-time MRI videos of speech production. We propose to exploit the temporal structure, inherently present in these MRI videos, to improve the detection of the air-tissue boundaries crucial for characterizing the speech generation process. We use a CNN encoder-decoder network to obtain frame-wise representations. These representations are further connected across time using a ConvLSTM layer. We evaluated the proposed system against current image-based methods and established that the addition of temporal information significantly improves the detection of deformable articulators, especially when two articulators are touching each other or in close proximity, as is the case of time points when maximal airway constrictions are formed for several classes of speech sounds.

6. REFERENCES

[1] Shrikanth Narayanan, Krishna Nayak, Sungbok Lee, Abhinav Sethy, and Dani Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, 2004.

[2] Erik Bresch, Yoon-Chul Kim, Krishna Nayak, Dani Byrd, and Shrikanth 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.

[3] Asterios Toutios and Shrikanth S. Narayanan, "Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research," *APSIPA Transactions on Signal and Information Processing*, vol. 5, pp. e6, 2016.

[4] Sajan Goud Lingala, Brad P Sutton, Marc E Miquel, and Krishna S Nayak, "Recommendations for real-time speech mri," *Journal of Magnetic Resonance Imaging*, vol. 43, no. 1, pp. 28–44, 2016.

[5] Christina Hagedorn, Tanner Sorensen, Adam Lammert, Asterios Toutios, Louis Goldstein, Dani Byrd, and Shrikanth Narayanan, "Engineering innovation in speech science: Data and technologies," *Perspectives of the ASHA Special Interest Groups*, vol. 4, no. 2, pp. 411–420, 2019.

[6] Sajan Goud Lingala, Yinghua Zhu, Yoon-Chul Kim, Asterios Toutios, Shrikanth Narayanan, and Krishna S. Nayak, "A fast and flexible mri system for the study of dynamic vocal tract shaping," *Magnetic Resonance in Medicine*, vol. 77, no. 1, pp. 112–125, 2017.

[7] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 28, no. 3, pp. 323–338, March 2009.

[8] 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 *Proc. Interspeech 2017*, 2017.

[9] Vijay Badrinarayanan, Alex Kendall, and Roberto 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.

[10] CA Valliappan, 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 *Interspeech*, 2018, pp. 3132–3136.

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

[12] CA Valliappan, Avinash Kumar, Renuka Mannem, GR Karthik, and Prasanta Kumar Ghosh, "An improved air tissue boundary segmentation technique for real time magnetic resonance imaging video using segnet," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5921–5925.

[13] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[14] Marijn F Stollenga, Wonmin Byeon, Marcus Liwicki, and Juergen Schmidhuber, "Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation," in *Advances in neural information processing systems*, 2015, pp. 2998–3006.

[15] Rudra P. K. Poudel, Pablo Lamata, and Giovanni Montana, "Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation," in *Reconstruction, Segmentation, and Analysis of Medical Images*, Maria A. Zuluaga, Kanwal Bhatia, Bernhard Kainz, Mehdi H. Moghari, and Danielle F. Pace, Eds., Cham, 2017, pp. 83–94, Springer International Publishing.

[16] Wenjia Bai, Hideaki Suzuki, Chen Qin, Giacomo Tarroni, Ozan Oktay, Paul M Matthews, and Daniel Rueckert, "Recurrent neural networks for aortic image sequence segmentation with sparse annotations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 586–594.

[17] Johannes Töger, Tanner Sorensen, Krishna Somandepalli, Asterios Toutios, Sajan Goud Lingala, Shrikanth Narayanan, and Krishna Nayak, "Test–retest repeatability of human speech biomarkers from static and real-time dynamic magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3323–3336, 2017.