Air-Tissue Boundary Segmentation In Real Time Magnetic Resonance Imaging Video Using A Convolutional Encoder-Decoder Network

Renuka Mannem and Prasanta Kumar Ghosh

SPIRE LAB
Electrical Engineering,
Indian Institute of Science (IISc), Bangalore, Karnataka, India

May 17, 2019
Section 1

1 Introduction
2 Methodology
3 Experiments
4 Results
5 Discussion
6 Summary
7 Acknowledgement
Introduction

- **Goal**: Segmentation of the Air-Tissue Boundaries (ATBs) in real time Magnetic Resonance Imaging (rtMRI) video.

![rt-MRI Image](image1.png)  ![Air Tissue Boundaries](image2.png)
Goal: Segmentation of the Air-Tissue Boundaries (ATBs) in real time Magnetic Resonance Imaging (rtMRI) video.

Approach: ATB segmentation using a convolutional encoder-decoder network (CEDN) \(^1\)

---

\(^1\)Yang et. al, ”Object contour detection with a fully convolutional encoder-decoder network,” CVPR, 2016.
Sentence: "They own a big house in the remote countryside"
Introduction

Motivation

Why ATBs?

- Speech production modeling\(^1\)

---

\(^1\) E. Bresch et. al, “Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging,” 2008.
Introduction

Motivation

Why ATBs?

- Speech production modeling\(^1\)
- Text-to-speech synthesis\(^2\)

---

\(^1\) E. Bresch et. al, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," 2008.

Introduction

Motivation

Why ATBs?

- Speech production modeling\(^1\)
- Text-to-speech synthesis\(^2\)
- Analysis of vocal tract morphology\(^3\)

\(^1\) E. Bresch et. al, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," 2008.
Introduction

Motivation

Why ATBs?

- Speech production modeling\(^1\)
- Text-to-speech synthesis\(^2\)
- Analysis of vocal tract morphology\(^3\)
- Automatic visual augmentation\(^4\)

---

Introduction

Dataset

- **USC-TIMIT**\(^1\) corpus
- **MOCHA-TIMIT** sentences
- **2-Female** (F1, F2) and **2-Male** (M1, M2).
- Subset: 16 Videos from each subject.
- Total No of frames: 5779.
- Video: 23.18 fps.
- Spatial resolution of \(68 \times 68\).

\(^1\)S.Narayanan et al, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)", JASA, 2014.
Dataset

- **Manual annotation:**
  1. Complete ATBs
  2. Upper lip (UL)
  3. Lower lip (LL)
  4. Tongue base (AVR)
  5. Velum tip (VEL)
  6. Glottis begin (GLTB)

Number of frames: 1462, 1270, 1642, 1399 for subjects F1, F2, M1, M2 respectively.

SPIRE LAB, IISc, Bangalore
Manual annotation:

1. Complete ATBs
2. Upper lip (UL)
3. Lower lip (LL)
4. Tongue base (AVR)
5. Velum tip (VEL)
6. Glottis begin (GLTB)

Number of frames: 1462, 1270, 1642, 1399 for subjects F1, F2, M1, M2 respectively.
Introduction

Dataset

- Ground truth binary image \((upper, lower)\) generation from manually annotated ATBs.
- Pixel value \(= 1\) if the manually annotated contour traverses through that pixel, otherwise pixel value \(= 0\).
Section 2

1. Introduction
2. Methodology
3. Experiments
4. Results
5. Discussion
6. Summary
7. Acknowledgement
Methodology

Proposed CEDN based ATB segmentation

Illustration of the steps in the proposed CEDN based approach

Test rtMRI Frame → Pre-processing → CEDN based upper contour prediction → CEDN based lower contour prediction → Binary Image Generation and Perimeter Filtering → ATB Prediction

Within Vocal Tract ATB Prediction

SPIRE LAB, IISc, Bangalore
Proposed CEDN based ATB segmentation

Illustration of the steps in the proposed CEDN based approach
Proposed CEDN based ATB segmentation

Preprocessing

- Enhancement using the image processing technique.\(^1\)
- To reduce the image artifacts for better performance of the ATB segmentation.

---

\(^1\)Kim et.al, ”Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data,” ISSP, 2014.
Proposed CEDN based ATB segmentation

Illustration of the steps in the proposed CEDN based approach

Methodology
Proposed CEDN based ATB segmentation

Illustration of the steps in the proposed CEDN based approach
Proposed CEDN based ATB segmentation

CEDN architecture

Encoder: 13 convolutional layered VGG-16 architecture.

**Proposed CEDN based ATB segmentation**

**CEDN architecture\(^1\)**

1. **Encoder:** 13 convolutional layered VGG-16 architecture.
2. **Decoder with less number of layers.**

---

Proposed CEDN based ATB segmentation

CEDN architecture

1. Encoder: 13 convolutional layered VGG-16 architecture.
2. Decoder with less number of layers.
3. Two separate CEDNs for upper and lower contour prediction.

Proposed CEDN based ATB segmentation

CEDN based contour prediction

1. Training: preprocessed input images and ground truth binary images (upper, lower)
Proposed CEDN based ATB segmentation

CEDN based contour prediction

1. Training: preprocessed input images and ground truth binary images (upper, lower)
2. Both encoder and decoder weights are learnt during training.
Methodology

Proposed CEDN based ATB segmentation

1. Training: preprocessed input images and ground truth binary images (upper, lower)
2. Both encoder and decoder weights are learnt during training.
3. Outputs a probability image with pixel values range from 0 to 1 (upper*, lower*).
Proposed CEDN based ATB segmentation

1. Training: preprocessed input images and ground truth binary images (upper, lower)
2. Both encoder and decoder weights are learnt during training.
3. Outputs a probability image with pixel values range from 0 to 1 (upper*, lower*).
4. 1 and 0 indicate the most and least probable ATB pixels respectively.
Proposed CEDN based ATB segmentation

Illustration of the steps in the proposed CEDN based approach

1. Pre-processing
   - Test rtMRI Frame
2. CEDN based upper contour prediction
3. CEDN based lower contour prediction
4. Binary Image Generation and Perimeter Filtering
5. Within Vocal Tract ATB Prediction

Output: $\hat{C}_1^p$, $\hat{C}_2^p$
Proposed CEDN based ATB segmentation

Illustration of the steps in the proposed CEDN based approach
Proposed CEDN based ATB segmentation

1. **Thresholding:** To obtain the binary images \((upper_b^*, lower_b^*)\).
Proposed CEDN based ATB segmentation

Binary Image Generation and Perimeter Filtering

1. Thresholding: To obtain the binary images ($upper_b^*$, $lower_b^*$).
2. Best threshold: Decided based on the performance on the validation data.
Proposed CEDN based ATB segmentation

Binary Image Generation and Perimeter Filtering

1. Thresholding: To obtain the binary images ($upper^*_b$, $lower^*_b$).
2. Best threshold: Decided based on the performance on the validation data.
3. $upper^*_p$, $lower^*_p$: Contain only the perimeter pixels of the detected closed ATB in binary images.
Proposed CEDN based ATB segmentation

Binary Image Generation and Perimeter Filtering

1. Thresholding: To obtain the binary images ($upper_b^*$, $lower_b^*$).

2. Best threshold: Decided based on the performance on the validation data.

3. $upper_p^*$, $lower_p^*$: Contain only the perimeter pixels of the detected closed ATB in binary images.

4. Perimeter pixel: Non-zero and connected to at least one zero-valued pixel with 4-connectivity.
Proposed CEDN based ATB segmentation

Illustration of the steps in the proposed CEDN based approach
Proposed CEDN based ATB segmentation

Within Vocal Tract ATB Prediction

1. Predicts ATBs within the vocal tract from $upper_p^*$, $lower_p^*$ and fixed contour $(C_3)$

---

Proposed CEDN based ATB segmentation

Within Vocal Tract ATB Prediction

1. Predicts ATBs within the vocal tract from $upper_p^*$, $lower_p^*$ and fixed contour ($C_3$)

2. Contour coordinates: pixel indices with value one are sorted in clockwise direction.

---

Proposed CEDN based ATB segmentation

Within Vocal Tract ATB Prediction

1 Predicts ATBs within the vocal tract from $upper_p^*$, $lower_p^*$ and fixed contour ($C_3$)

2 Contour coordinates: pixel indices with value one are sorted in clockwise direction.

3 Contour pruning$^1$ to obtain ATBs within the vocal tract.

---

Methodology

Proposed CEDN based ATB segmentation

Within Vocal Tract ATB Prediction

1. Predicts ATBs within the vocal tract from $upper_p^*$, $lower_p^*$ and fixed contour ($C_3$)
2. Contour coordinates: pixel indices with value one are sorted in clockwise direction.
3. Contour pruning$^1$ to obtain ATBs within the vocal tract.
4. Smoothing using a moving average filter with size $q \times q$.
5. $q$ is decided based on the performance on the validation data.

---

Section 3

Experiments
Experimental Setup

Baselines:

- Maeda grid-line\(^1\) (MG).

\(^1\)Kim et.al, ”Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data,” ISSP, 2014.
Experimental Setup

Baselines:

- Maeda grid-line\(^1\) (MG).
- Fisher-discrimination measure based segmentation\(^2\) (SFDM)

---

\(^1\) Kim et.al, "Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data," ISSP, 2014.

Experiments

Experimental Setup

Baselines:

- Maeda grid-line\(^1\) (MG).
- Fisher-discrimination measure based segmentation\(^2\) (SFDM)
- fully convolutional networks based segmentation\(^3\) (SFCN)

---

\(^1\) Kim et.al, "Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data," ISSP, 2014.
\(^3\) Valliappan CA et. al, Air-tissue boundary segmentation in real-time magnetic resonance imaging video using semantic segmentation with fully convolutional networks," Interspeech, 2018
Experimental Setup

3 types of experiments:

- Seen subject condition
- Unseen subject condition
- Adaptation using unseen subject’s data
Experimental Setup

**Seen subject condition:**

- 16 = 4 videos from per subject
- 4-fold cross validation.

- Training set: \(\sim 2900\)
- Development and Test sets: \(\sim 1443\)
- 30 epochs, early stopping condition.
**Experiments**

**Experimental Setup**

**Seen subject condition:**

- 4-fold cross validation.
- Training set: $\sim 2900$.
- Development and Test sets: $\sim 1443$.
- 30 epochs, early stopping condition.
Experimental Setup

Unseen subject condition:

16 videos from a subject

- Fold 1: F1, F2, M1, M2
- Fold 2: F1, F2, M1, M2
- Fold 3: F1, F2, M1, M2
- Fold 4: F1, F2, M1, M2

Train set: {48 videos}

Test set: {16 videos}

- 4-fold cross validation.
- Training set: \( \sim 4334 \).
- Development and Test sets: \( \sim 1443 \).
- 50 epochs.
Experimental Setup

Adaptation using unseen subject’s data:

16 videos from a subject

- Fold 1
  - F1
  - F2
  - M1
  - M2
- Fold 2
  - F1
  - F2
  - M1
  - M2
- Fold 3
  - F1
  - F2
  - M1
  - M2
- Fold 4
  - F1
  - F2
  - M1
  - M2

- Train set
  - 48 videos

- Adaptation set
  - 100 images

- Validation set

- Test set

- Minimum number of unseen subject’s images required to be better than MG.
Experimental Setup

Adaptation using unseen subject’s data:

16 videos from a subject

Fold 1: F1 F2 M1 M2
Fold 2: F1 F2 M1 M2
Fold 3: F1 F2 M1 M2
Fold 4: F1 F2 M1 M2

- Minimum number of unseen subject’s images required to be better than MG.
- Trained model is adapted from P many frames from adaptation set (P = 0, 10, 20, 30).
Experiments

Experimental Setup

Adaptation using unseen subject’s data:

- Minimum number of unseen subject’s images required to be better than MG.
- Trained model is adapted from $P$ many frames from adaptation set ($P = 0, 10, 20, 30$).
- Last 5 deconvolutional layers are only learned.
Evaluation metric

- **DTW distance**: Measures the closeness of the estimated contour to the ground truth contour (unit: pixel).

---

Section 4

1 Introduction
2 Methodology
3 Experiments
4 Results
5 Discussion
6 Summary
7 Acknowledgement
## Seen subject condition

<table>
<thead>
<tr>
<th>Approach</th>
<th>Upper</th>
<th>Lower</th>
</tr>
</thead>
<tbody>
<tr>
<td>MG</td>
<td>1.13 ± 0.23</td>
<td>1.27 ± 0.36</td>
</tr>
<tr>
<td>SFDM</td>
<td>1.08 ± 0.20</td>
<td>1.14 ± 0.29</td>
</tr>
<tr>
<td>SFCN</td>
<td><strong>1.03 ± 0.20</strong></td>
<td>1.13 ± 0.26</td>
</tr>
<tr>
<td>CEDN</td>
<td>1.10 ± 0.20</td>
<td><strong>1.09 ± 0.24</strong></td>
</tr>
</tbody>
</table>

Average (± standard deviation) DTW distance across all the subjects (blue indicates the least DTW distance)

CEDN based approach gives better performance for lower contours compared to baselines.
### Results

#### Seen subject condition

<table>
<thead>
<tr>
<th>Approach</th>
<th>Upper</th>
<th>Lower</th>
</tr>
</thead>
<tbody>
<tr>
<td>MG</td>
<td>1.13 ± 0.23</td>
<td>1.27 ± 0.36</td>
</tr>
<tr>
<td>SFDM</td>
<td>1.08 ± 0.20</td>
<td>1.14 ± 0.29</td>
</tr>
<tr>
<td>SFCN</td>
<td><strong>1.03 ± 0.20</strong></td>
<td>1.13 ± 0.26</td>
</tr>
<tr>
<td>CEDN</td>
<td>1.10 ± 0.20</td>
<td><strong>1.09 ± 0.24</strong></td>
</tr>
</tbody>
</table>

Average (± standard deviation) DTW distance across all the subjects (blue indicates the least DTW distance)

- CEDN based approach gives better performance for lower contours compared to baselines.
Unseen subject condition

<table>
<thead>
<tr>
<th>Approach</th>
<th>Upper</th>
<th>Lower</th>
</tr>
</thead>
<tbody>
<tr>
<td>MG</td>
<td>1.13 ± 0.23</td>
<td>1.27 ± 0.36</td>
</tr>
<tr>
<td>SFDM</td>
<td>2.34 ± 0.47</td>
<td>2.06 ± 0.78</td>
</tr>
<tr>
<td>SFCN</td>
<td>2.79 ± 0.35</td>
<td>13.3 ± 0.98</td>
</tr>
<tr>
<td>CEDN</td>
<td>1.65 ± 0.30</td>
<td>1.72 ± 0.32</td>
</tr>
</tbody>
</table>

Average (± standard deviation) DTW distance across all the subjects (blue and green colours indicate first and second least DTW distances respectively)

CEDN based approach gives better performance compared to the supervised approaches (SFDM and SFCN).

Better generalizability for new subjects.
Unseen subject condition

<table>
<thead>
<tr>
<th>Approach</th>
<th>Upper</th>
<th>Lower</th>
</tr>
</thead>
<tbody>
<tr>
<td>MG</td>
<td>$1.13 \pm 0.23$</td>
<td>$1.27 \pm 0.36$</td>
</tr>
<tr>
<td>SFDM</td>
<td>$2.34 \pm 0.47$</td>
<td>$2.06 \pm 0.78$</td>
</tr>
<tr>
<td>SFCN</td>
<td>$2.79 \pm 0.35$</td>
<td>$13.3 \pm 0.98$</td>
</tr>
<tr>
<td>CEDN</td>
<td>$1.65 \pm 0.30$</td>
<td>$1.72 \pm 0.32$</td>
</tr>
</tbody>
</table>

Average ($\pm$ standard deviation) DTW distance across all the subjects (blue and green colours indicate first and second least DTW distances respectively)

- CEDN based approach gives better performance compared to the supervised approaches (SFDM and SFCN).
- Better generalizability for new subjects.
Results

Adaptation using unseen subject’s data

Bar plot - DTW distance on the validation data using CEDN, Errorbar - std, Blue line - DTW distance using MG.

CEDN models yield better validation data performance than the MG scheme with 30 adaptation images.
Adaptation using unseen subject’s data

Bar plot - DTW distance on the validation data using CEDN, Errorbar - std, Blue line - DTW distance using MG.

- CEDN models yield better validation data performance than the MG scheme with 30 adaptation images.
Adaptation using unseen subject’s data

<table>
<thead>
<tr>
<th>Sub</th>
<th>Upper Contour</th>
<th>Lower Contour</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MG</td>
<td>CEDN</td>
</tr>
<tr>
<td>F1</td>
<td>1.03 ± 0.27</td>
<td>1.02 ± 0.20</td>
</tr>
<tr>
<td>F2</td>
<td>1.20 ± 0.24</td>
<td>1.16 ± 0.22</td>
</tr>
<tr>
<td>M1</td>
<td>1.23 ± 0.19</td>
<td>1.21 ± 0.24</td>
</tr>
<tr>
<td>M2</td>
<td>1.20 ± 0.24</td>
<td>1.18 ± 0.20</td>
</tr>
<tr>
<td>Avg</td>
<td>1.17 ± 0.23</td>
<td>1.14 ± 0.20</td>
</tr>
</tbody>
</table>

(Average (± std) DTW distance using MG and CEDN (with 30 adaptation images) for test data (blue colour indicates the least DTW distance))
## Results

### Adaptation using unseen subject’s data

<table>
<thead>
<tr>
<th>Sub</th>
<th>Upper Contour</th>
<th>Lower Contour</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MG</td>
<td>CEDN</td>
</tr>
<tr>
<td>F1</td>
<td>1.03 ± 0.27</td>
<td>1.02 ± 0.20</td>
</tr>
<tr>
<td>F2</td>
<td>1.20 ± 0.24</td>
<td>1.16 ± 0.22</td>
</tr>
<tr>
<td>M1</td>
<td>1.23 ± 0.19</td>
<td>1.21 ± 0.24</td>
</tr>
<tr>
<td>M2</td>
<td>1.20 ± 0.24</td>
<td>1.18 ± 0.20</td>
</tr>
<tr>
<td>Avg</td>
<td>1.17 ± 0.23</td>
<td>1.14 ± 0.20</td>
</tr>
</tbody>
</table>

(Average (± std) DTW distance using MG and CEDN (with 30 adaptation images) for test data (blue colour indicates the least DTW distance)

- SFCN and SFDM approaches with 30 adaptation images failed to perform better than the MG approach
Section 5

1. Introduction
2. Methodology
3. Experiments
4. Results
5. Discussion
6. Summary
7. Acknowledgement
Discussion

Reasons for better performance:

1. Supervised nature - overcomes imaging artifacts and grainy noise.

CEDN does not perform better in upper contour predictions in some cases due to having a cluster of points near the velum region.
Reasons for better performance:

1. Supervised nature - overcomes imaging artifacts and grainy noise.
2. Light decoder, learning both encoder and decoder weights, direct prediction of ATBs from network - requires limited number of training images.
Reasons for better performance:

1. Supervised nature - overcomes imaging artifacts and grainy noise.
2. Light decoder, learning both encoder and decoder weights, direct prediction of ATBs from network - requires limited number of training images.
3. Perimeter filtering - precise boundary pixels.
Reasons for better performance:

1. Supervised nature - overcomes imaging artifacts and grainy noise.
2. Light decoder, learning both encoder and decoder weights, direct prediction of ATBs from network - requires limited number of training images
3. Perimeter filtering - precise boundary pixels

CEDN does not perform better in upper contour predictions in some cases due to having cluster of points near velum region.
Section 6

Introduction
Methodology
Experiments
Results
Discussion
Summary
Acknowledgement
Summary

Conclusions

- Proposed method yields better performance than the baselines.
- Better generalizability compared to the supervised baselines.
Conclusions

- Proposed method yields better performance than the baselines.
- Better generalizability compared to the supervised baselines.

Future Work

- Adaptive thresholding to generate binary images from the CEDN output probability images.
The authors thank Pratiksha Trust for their support.
Questions?