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

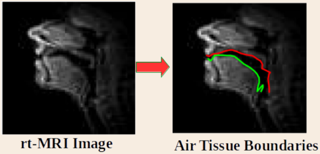
SPIRE LAB

May 17, 2019

Section 1

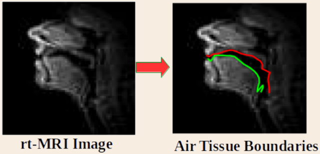
Introduction

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



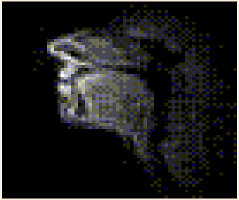
Introduction

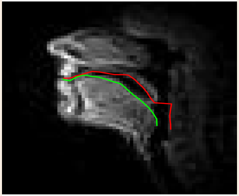
- **Goal**: Segmentation of the Air-Tissue Boundaries (ATBs) in real time Magnetic Resonance Imaging (rtMRI) video.



- **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"

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.

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.
2 Toutios et.al, "Articulatory Synthesis Based on Real-Time Magnetic Resonance Imaging Data," Interpseech, 2016.

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.
[2] Toutios et.al, "Articulatory Synthesis Based on Real-Time Magnetic Resonance Imaging Data, " Interpseech, 2016.
[3] Ramanarayanan et. al, "An investigation of articulatory setting using real-time magnetic resonance imaging," the Journal of the Acoustical Society of America, 2013.

Motivation

Why ATBs?

- Speech production modeling[1]
- Text-to-speech synthesis[2]
- Analysis of vocal tract morphology[3]
- Automatic visual augmentation[4]

[1] E. Bresch et. al, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," 2008.

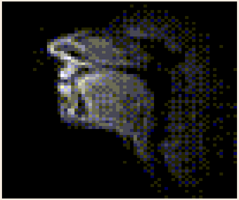
[2] Toutios et.al, "Articulatory Synthesis Based on Real-Time Magnetic Resonance Imaging Data, " Interpseech, 2016.

[3] Ramanarayanan et. al, "An investigation of articulatory setting using real-time magnetic resonance imaging," the Journal of the Acoustical Society of America, 2013.

[4] Chandana et. al, "Automatic visual augmentation for concatenation based synthesized articulatory videos from real-time MRI data for spoken language training," Interspeech, 2018.

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 68×68.



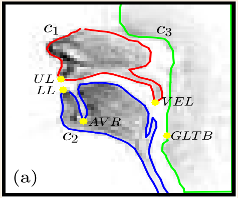
[1] S.Narayanan et. al, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)", JASA, 2014.

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)

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.



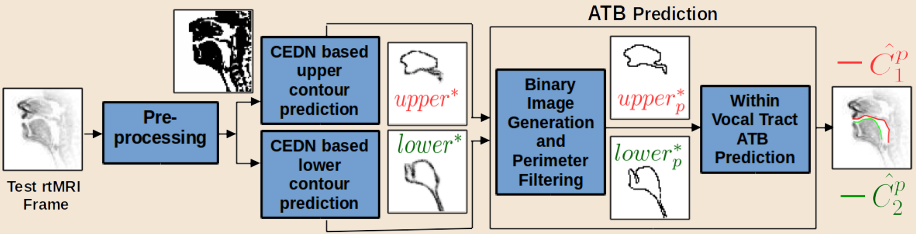
Dataset

- 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

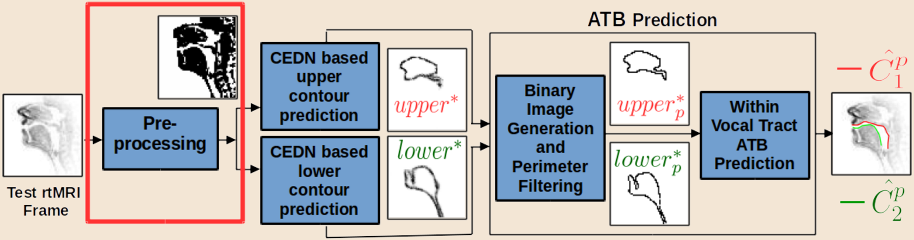
Proposed CEDN based ATB segmentation

Illustration of the steps in the proposed CEDN based approach

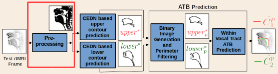


Proposed CEDN based ATB segmentation

Illustration of the steps in the proposed CEDN based approach



Proposed CEDN based ATB segmentation



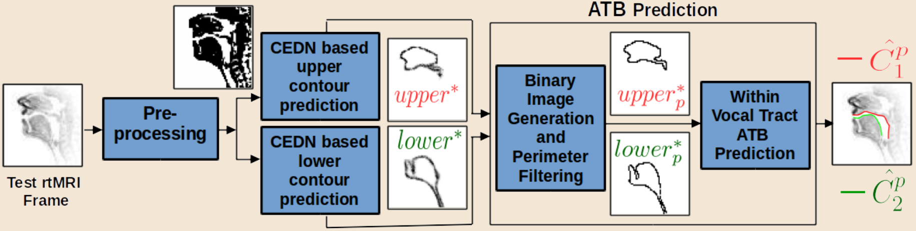
Preprocessing

- 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.

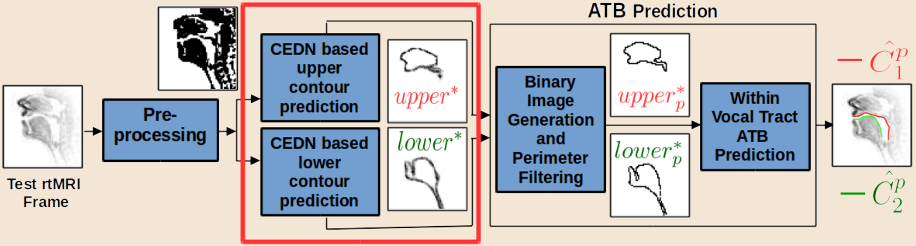
Proposed CEDN based ATB segmentation

Illustration of the steps in the proposed CEDN based approach

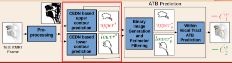


Proposed CEDN based ATB segmentation

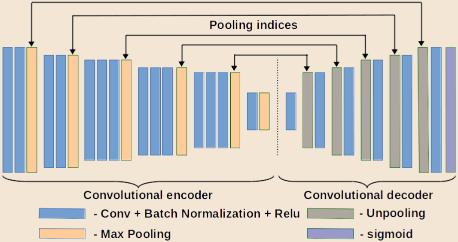
Illustration of the steps in the proposed CEDN based approach



Proposed CEDN based ATB segmentation



CEDN architecture[1]

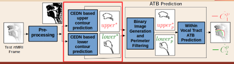


1. Encoder: 13 convolutional layered VGG-16 architecture.

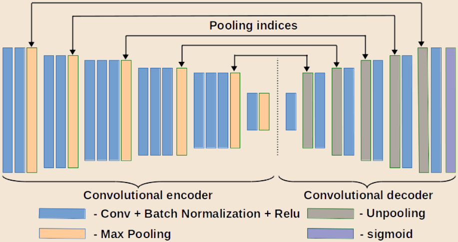
[1] Yang et. al, "Object contour detection with a fully convolutional encoder-decoder network," CVPR, 2016.

Proposed CEDN based ATB segmentation

SPIRE LAB



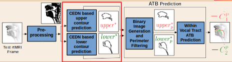
CEDN architecture[1]



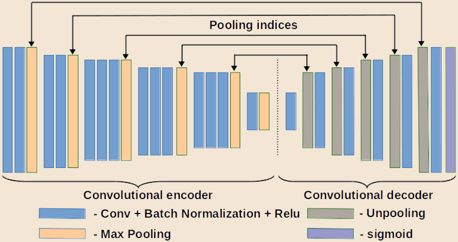
1. Encoder: 13 convolutional layered VGG-16 architecture.
2. Decoder with less number of layers.

[1] Yang et. al, "Object contour detection with a fully convolutional encoder-decoder network," CVPR, 2016.

Proposed CEDN based ATB segmentation



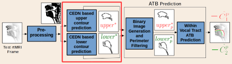
CEDN architecture[1]

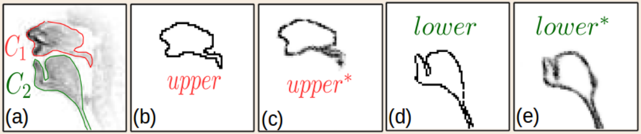


1. Encoder: 13 convolutional layered VGG-16 architecture.
2. Decoder with less number of layers.
3. Two separate CEDNs for upper and lower contour prediction.

[1] Yang et. al, "Object contour detection with a fully convolutional encoder-decoder network," CVPR, 2016.

Proposed CEDN based ATB segmentation

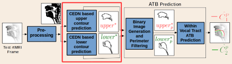


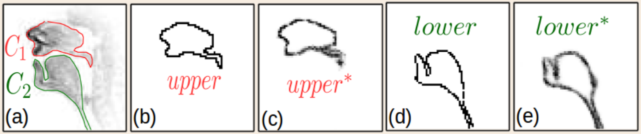


CEDN based contour prediction

1. Training: preprocessed input images and ground truth binary images ($upper$, $lower$)

Proposed CEDN based ATB segmentation

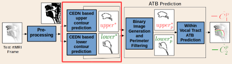


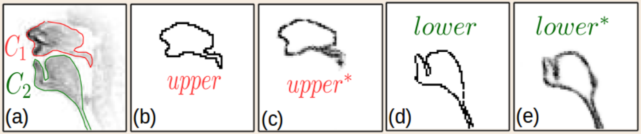


CEDN based contour prediction

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

Proposed CEDN based ATB segmentation

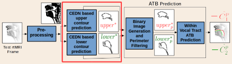


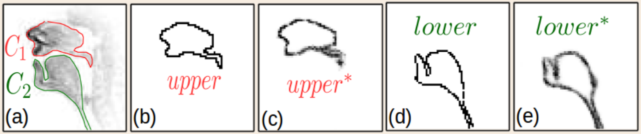


CEDN based contour prediction

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

Proposed CEDN based ATB segmentation



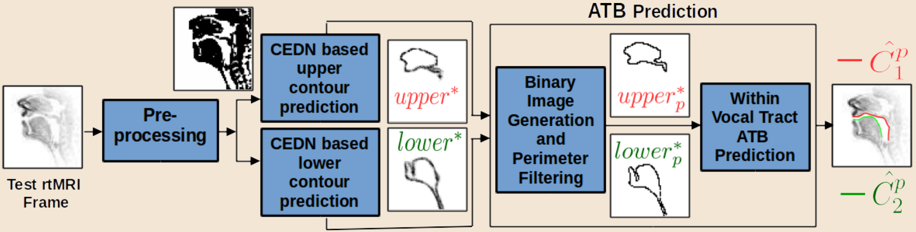


CEDN based contour prediction

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

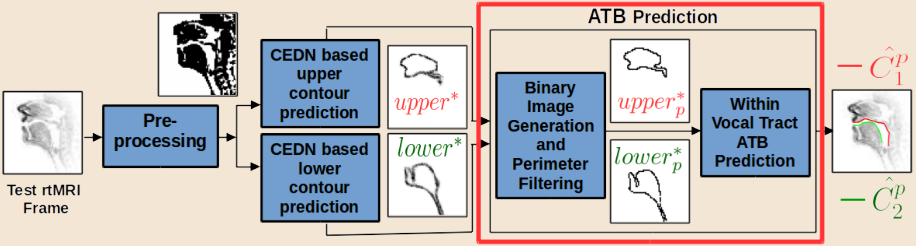
Proposed CEDN based ATB segmentation

Illustration of the steps in the proposed CEDN based approach

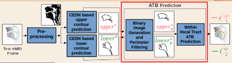


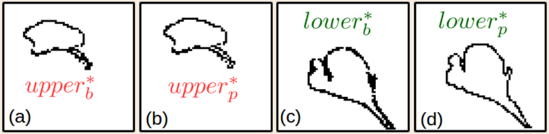
Proposed CEDN based ATB segmentation

Illustration of the steps in the proposed CEDN based approach



Proposed CEDN based ATB segmentation

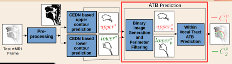


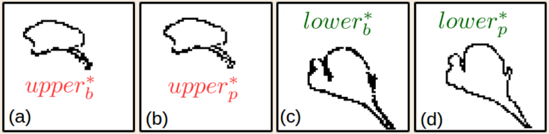


Binary Image Generation and Perimeter Filtering

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

Proposed CEDN based ATB segmentation

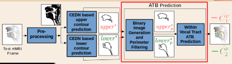


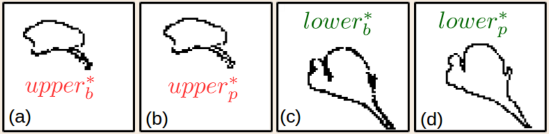


Binary Image Generation and Perimeter Filtering

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

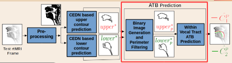


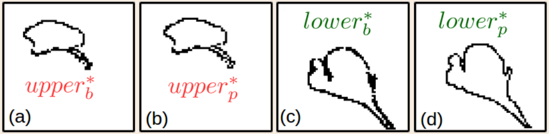


Binary Image Generation and Perimeter Filtering

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



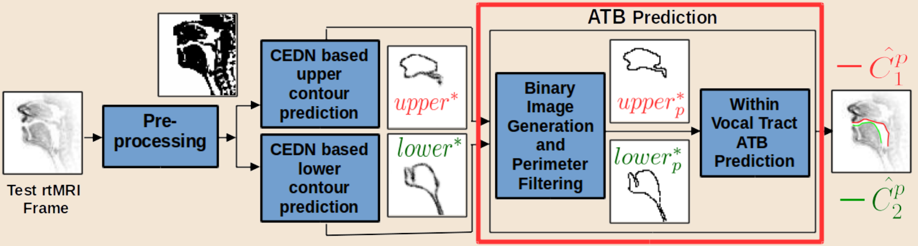


Binary Image Generation and Perimeter Filtering

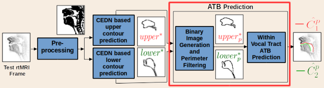
1. Thresholding: To obtain the binary images ($upper^*_b$, $lower^*_b$).
2. Best threshold: Decided based on the performance on the validation data.
3. $upper^*_p$, $lower^*_p$: Contain only the perimeter pixels of the detected closed ATB in binary images.
4. perimeter pixel: Non-zero and connected to 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

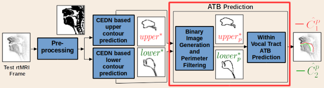


Within Vocal Tract ATB Prediction

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

[1] A. Koparkar et. al, "A supervised air-tissue boundary segmentation technique in real-time magnetic resonance imaging video using a novel measure of contrast and dynamic programming," ICASSP, 2018.

Proposed CEDN based ATB segmentation

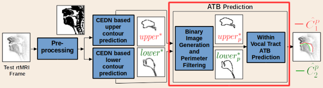


Within Vocal Tract ATB Prediction

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.

[1] A. Koparkar et. al, "A supervised air-tissue boundary segmentation technique in real-time magnetic resonance imaging video using a novel measure of contrast and dynamic programming," ICASSP, 2018.

Proposed CEDN based ATB segmentation

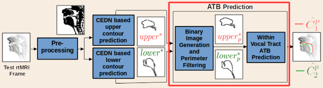


Within Vocal Tract ATB Prediction

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.

[1] A. Koparkar et. al, "A supervised air-tissue boundary segmentation technique in real-time magnetic resonance imaging video using a novel measure of contrast and dynamic programming," ICASSP, 2018.

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.

[1] A. Koparkar et. al, "A supervised air-tissue boundary segmentation technique in real-time magnetic resonance imaging video using a novel measure of contrast and dynamic programming," ICASSP, 2018.

Section 3

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.
[2] A. Koparkar et. al, "A supervised air-tissue boundary segmentation technique in real-time magnetic resonance imaging video using a novel measure of contrast and dynamic programming," ICASSP, 2018.

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.

[2] A. Koparkar et. al, "A supervised air-tissue boundary segmentation technique in real-time magnetic resonance imaging video using a novel measure of contrast and dynamic programming," ICASSP, 2018.

[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

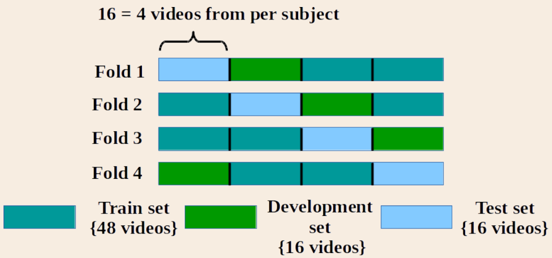
Experimental Setup

3 types of experiments:

- Seen subject condition
- Unseen subject condition
- Adaptation using unseen subject's data

Experimental Setup

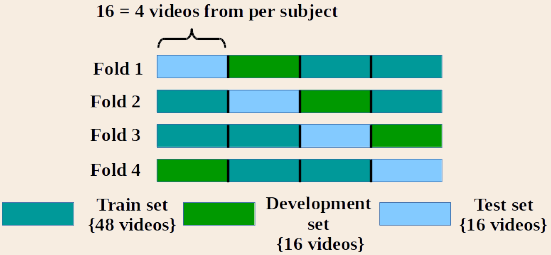
Seen subject condition:



- 4-fold cross validation.

Experimental Setup

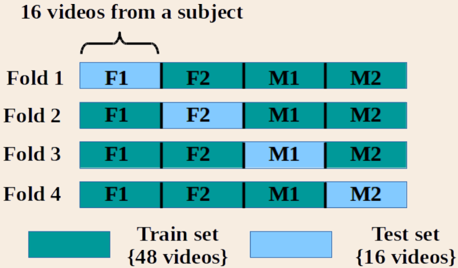
Seen subject condition:



- 4-fold cross validation.
- Training set: ~ 2900.
- Development and Test sets: ~ 1443.
- 30 epochs, early stopping condition.

Experimental Setup

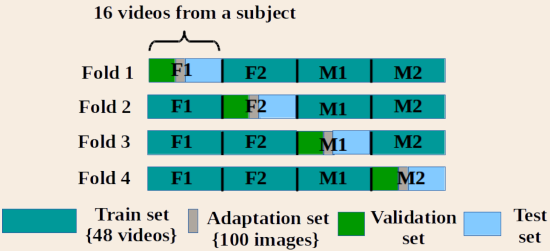
Unseen subject condition:



- 4-fold cross validation.
- Training set: ~ 4334.
- Development and Test sets: ~ 1443.
- 50 epochs.

Experimental Setup

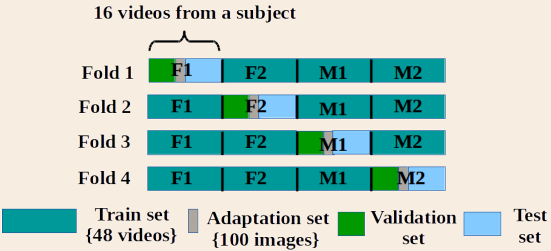
Adaptation using unseen subject's data:



- Minimum number of unseen subject's images required to be better than MG.

Experimental Setup

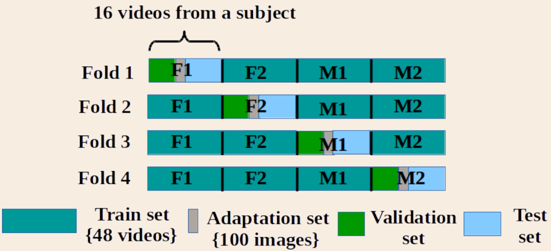
Adaptation using unseen subject's data:



- 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).

Experimental Setup

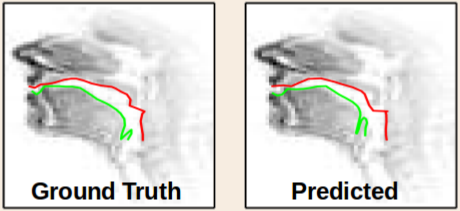
Adaptation using unseen subject's data:



- 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**[1]: Measures the closeness of the estimated contour to the ground truth contour (unit:pixel).



[1] Berndt et. al, "Using dynamic time warping to find patterns in time series," KDD, 1994.

Section 4

Seen subject condition

Approach	Upper	Lower
MG	1.13 ± 0.23	1.27 ± 0.36
SFDM	1.08 ± 0.20	$\mathbf{1.14 \pm 0.29}$
SFCN	$\mathbf{1.03 \pm 0.20}$	1.13 ± 0.26
CEDN	1.10 ± 0.20	$\mathbf{1.09 \pm 0.24}$

Average ($\pm$ standard deviation) DTW distance across all the subjects (blue indicates the least DTW distance)

Seen subject condition

Approach	Upper	Lower
MG	1.13 ± 0.23	1.27 ± 0.36
SFDM	1.08 ± 0.20	$\mathbf{1.14 \pm 0.29}$
SFCN	$\mathbf{1.03 \pm 0.20}$	1.13 ± 0.26
CEDN	1.10 ± 0.20	$\mathbf{1.09 \pm 0.24}$

Average ($\pm$ 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

Approach	Upper	Lower
MG	$\mathbf{1.13 \pm 0.23}$	$\mathbf{1.27 \pm 0.36}$
SFDM	2.34 ± 0.47	2.06 ± 0.78
SFCN	2.79 ± 0.35	13.3 ± 0.98
CEDN	$\mathbf{1.65 \pm 0.30}$	$\mathbf{1.72 \pm 0.32}$

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

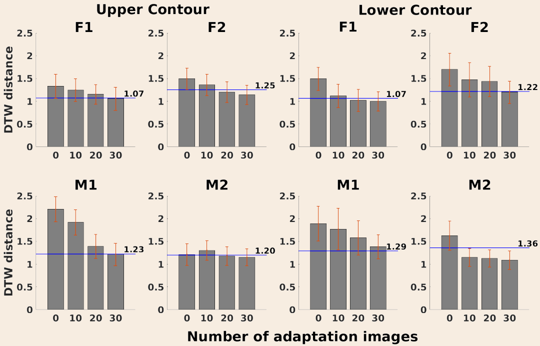
Unseen subject condition

Approach	Upper	Lower
MG	$\mathbf{1.13 \pm 0.23}$	$\mathbf{1.27 \pm 0.36}$
SFDM	2.34 ± 0.47	2.06 ± 0.78
SFCN	2.79 ± 0.35	13.3 ± 0.98
CEDN	$\mathbf{1.65 \pm 0.30}$	$\mathbf{1.72 \pm 0.32}$

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

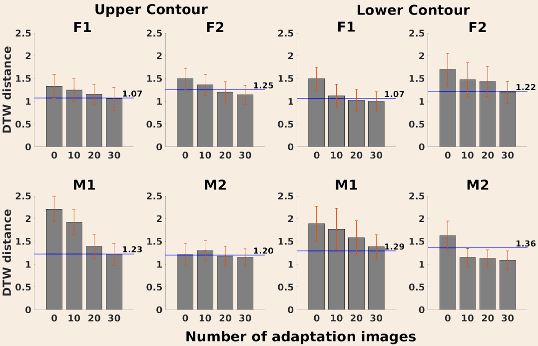
- CEDN based approach gives better performance compared to the supervised approaches (SFDM and SFCN).
- Better generalizability for new subjects.

Adaptation using unseen subject's data



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

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

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	$\mathbf{1.17 \pm 0.21}$

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

Adaptation using unseen subject's data

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	$\mathbf{1.17 \pm 0.21}$

(Average ($\pm$ 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

Reasons for better performance:

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

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

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.

Section 7

The authors thank Pratiksha Trust for their support.

Questions?