Supervised Air-Tissue Boundary Segmentation of real-time Magnetic Resonance Imaging Video

Advait Koparkar [1], Prasanta Kumar Ghosh [2]

[1] Department of Electrical and Electronics Engineering
Birla Institute of Technology and Science, Pilani
K K Birla Goa Campus, India

[2] SPIRE LAB
Electrical Engineering,
Indian Institute of Science (IISc), Bangalore, India

Section 1

Motivation for Using rtMRI

- real-time Magnetic Resonance Imaging (rtMRI) - tool for analyzing articulatory mechanisms in the vocal tract
- Non-invasive and safe method to capture shapes of speech articulators
- More effective than other methods such as X-Ray, Ultrasound, and Electromagnetic Articulograph

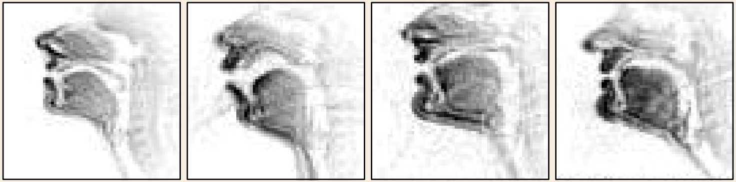


Figure: rtMRI frames from various subjects

Air-Tissue Boundaries (ATB)

- rtMRI data contains spatio-temporal information of the varying shape of the vocal tract and speech articulators
- ATBs - contours marking boundary between air cavity and tissue of the vocal tract
- ATBs are defined as:

$$C_k \triangleq \{(x_{ki}, y_{ki}); 1 \leq i \leq M_k\} \ \forall \ 1 \leq k \leq 3$$

C_1, C_2, C_3 - Upper, Lower, Pharyngeal ATB

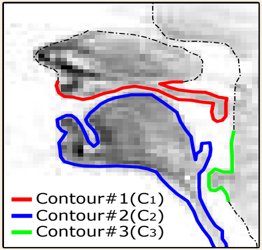


Figure: Air-Tissue Boundaries

Section 2

Database

- rtMRI videos taken from USC-TIMIT rtMRI database [1]
- 2 Male (M1, M2) and 2 Female (F1, F2) subjects - 10 sentences each
- 68×68 pixel videos recorded at 23.18 frames/s
- MATLAB based GUI for manually tracing ATBs of all rtMRI frames

[1] Shrikanth Narayanan 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

Database

- Upper lip (UL), lower lip (LL), tongue base (TB), velum tip (VEL) and glottis begin (GLTB) were also marked for each frame

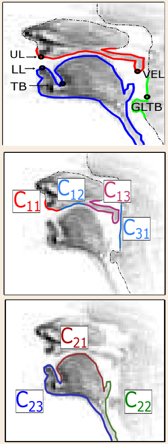


Figure: Manually marked ATBs

Database

- Upper lip (UL), lower lip (LL), tongue base (TB), velum tip (VEL) and glottis begin (GLTB) were also marked for each frame
- C_1 - C_{11} (Upper Lip), C_{12} (Hard Palate) and C_{13} (Velum)

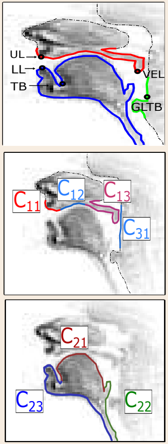


Figure: Manually marked ATBs

Database

- Upper lip (UL), lower lip (LL), tongue base (TB), velum tip (VEL) and glottis begin (GLTB) were also marked for each frame
- C_1 - C_{11} (Upper Lip), C_{12} (Hard Palate) and C_{13} (Velum)
- C_2 - C_{21} (Tongue Tip), C_{22} (Tongue Root) and C_{23} (Lower Lip)

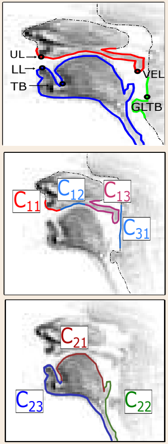


Figure: Manually marked ATBs

Database

- Upper lip (UL), lower lip (LL), tongue base (TB), velum tip (VEL) and glottis begin (GLTB) were also marked for each frame
- C_1 - C_{11} (Upper Lip), C_{12} (Hard Palate) and C_{13} (Velum)
- C_2 - C_{21} (Tongue Tip), C_{22} (Tongue Root) and C_{23} (Lower Lip)
- C_{31} - pharyngeal wall till GLTB

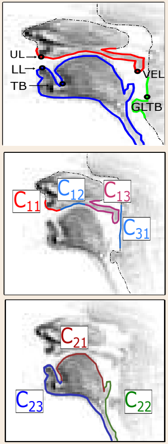


Figure: Manually marked ATBs

Section 3

Problem Statement: ATB Segmentation

To estimate the Upper and Lower ATBs ($\hat{\mathcal{C}}_1$ & $\hat{\mathcal{C}}_2$) for a given rtMRI video sequence $\mathcal{I}_{Test}$ containing N_{Test} frames such that the predicted ATBs correspond to contours of **maximal contrast**, while maintaining **temporal smoothness** across consecutive frames of $\mathcal{I}_{Test}$.

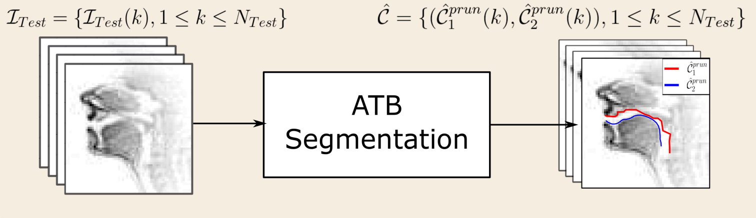


Figure: ATB Segmentation to obtain estimated upper and lower ATBs

Problem Statement: ATB Segmentation

Features of proposed supervised learning approach:

- Robust to imaging artifact thus increasing accuracy
- Exploits slowly varying nature of vocal tract morphology

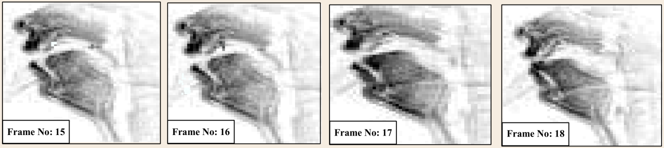
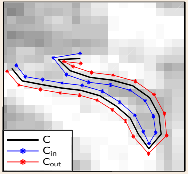


Figure: Smoothly varying vocal tract morphology

Section 4

Fisher Discriminant Measure of Contrast

- Consider $C = \{(x_i, y_i), 1 \leq i \leq M\}$ on an image frame I
- Inner contour C_{in} and Outer contour C_{out} are constructed from C
- Bicubic Interpolation [2] used for finding pixel values along C_{in} and C_{out}



$$I_{in} = \{I(x_i, y_i) \mid (x_i, y_i) \in C_{in}\}$$
$$I_{out} = \{I(x_i, y_i) \mid (x_i, y_i) \in C_{out}\}$$

Figure: Pixel Intensities along C_{in} & C_{out}

[2] Robert Keys, "Cubic Convolution Interpolation for Digital Image Processing" in IEEE transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-29, no. 6, pp. 1153–1160, 1981

Fisher Discriminant Measure of Contrast

The Fisher Discriminant Measure (FDM) $\mathcal{D}_F(C, I)$ is defined as:

$$\mathcal{D}_F(C, I) \triangleq \frac{(\overline{I_{in}} - \overline{I_{out}})^2}{{\sigma^2}_{I_{in}} + {\sigma^2}_{I_{out}}} \quad (1)$$

where,

$$\overline{I_{in}}, \overline{I_{out}} = \text{ mean pixel intensities along } C_{in} \text{ \& } C_{out}$$

$${\sigma^2}_{I_{in}}, {\sigma^2}_{I_{out}} = \text{ variance of intensities along } C_{in} \text{ \& } C_{out}$$

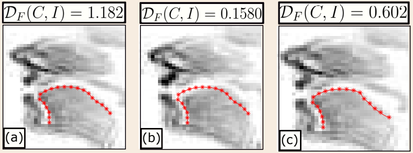


Figure: FDM example

Measure of Proximity between Two Contours

- An optimal alignment map between the points of C_a and C_b is found by the following optimization:

$$\{(m_a(l), m_b(l)), 1 \leq l \leq L\} = \underset{\substack{1 \leq m'_a(l) \leq M_a, \\ 1 \leq m'_b(l) \leq M_b}}{\operatorname{argmin}} \sum_{l=1}^{L} ||C_a(m'_a(l)) - C_b(m'_b(l))||_2 \tag{2}$$

where $C_a(i)$, $C_b(j) \in \mathbb{R}^2$ correspond to the i-th and j-th point of C_a and C_b

[3] Donald J Berndt and James Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series" in KDD workshop, vol. 10, no. 16, pp.359–370, 1994

Measure of Proximity between Two Contours

- An optimal alignment map between the points of C_a and C_b is found by the following optimization:

$$\{(m_a(l), m_b(l)), 1 \leq l \leq L\} = \underset{\substack{1 \leq m'_a(l) \leq M_a, \\ 1 \leq m'_b(l) \leq M_b}}{\operatorname{argmin}} \sum_{l=1}^{L} ||C_a(m'_a(l)) - C_b(m'_b(l))||_2 \quad (2)$$

where $C_a(i)$, $C_b(j) \in \mathbb{R}^2$ correspond to the i-th and j-th point of C_a and C_b

- $\mathcal{D}_D(C_a, C_b)$ - DTW distance [3] measures the alignment of any 2 contours (C_a, C_b)

$$\mathcal{D}_D(C_a, C_b) \triangleq \frac{1}{L} \sum_{l=1}^{L} ||C_a(m_a(l)) - C_b(m_b(l))||_2 \quad (3)$$

[3] Donald J Berndt and James Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series" in KDD workshop, vol. 10, no. 16, pp.359–370, 1994

ATB Prediction using Dynamic Programming

ATB Prediction: task of mapping manually traced training contours ($\mathcal{C}^{Tr}$) to test rtMRI video ($\mathcal{I}$)

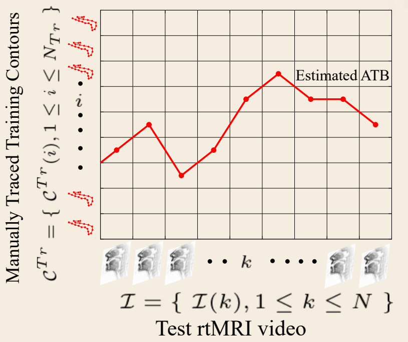


Figure: ATB Prediction: Mapping $\mathcal{C}^{Tr}$ to $\mathcal{I}$

ATB Prediction using Dynamic Programming

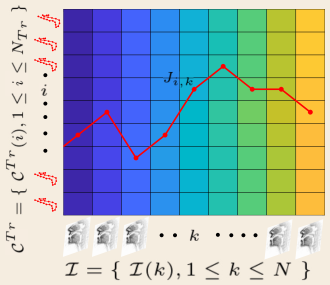
To obtain accurate, smoothly varying predicted contours the following objective function is defined:

$$J(\mathcal{C}^{Tr}(i), \mathcal{I}(k)) = \mathcal{D}_F(\mathcal{C}^{Tr}(i), \mathcal{I}(k)) + \max_{1 \leq i' \leq N_{Tr}} \{J(\mathcal{C}^{Tr}(i'), \mathcal{I}(k-1)) - \lambda \mathcal{D}_D(\mathcal{C}_{Tr}(i'), \mathcal{C}_{Tr}(i))\} \quad (4)$$

where,

$$\mathcal{I} = \{\mathcal{I}(k),\ 1 \leq k \leq N\}$$
$$\mathcal{C}^{Tr} = \{\mathcal{C}^{Tr}(i),\ 1 \leq i \leq N_{Tr}\}$$
$$\lambda = \text{ Temporal Stiffness Factor}$$

ATB Prediction using Dynamic Programming



Algorithm 1 Estimating $\mathcal{C}^*$ using Dynamic Programming

procedure ESTIMATEATB($\mathcal{C}^{Tr}, \mathcal{I}$)
    for $i \leftarrow 1$ to N_{Tr} **do** ▷ Initializing
        $J_{i,1} \leftarrow \mathcal{D}_F\ (\mathcal{C}^{Tr}(i), \mathcal{I}(1))$
    end for

    for $k \leftarrow 2$ to N **do** ▷ Iterative Step
        for $i \leftarrow 1$ to N_{Tr} **do**
            $J_{i,k} \leftarrow \mathcal{D}_F(\mathcal{C}^{Tr}(i), \mathcal{I}(k)) + \max_{1 \le i' \le N_{Tr}} \{J_{i',k-1} - \lambda \mathcal{D}_D(\mathcal{C}^{Tr}(i'), \mathcal{C}^{Tr}(i))\}$
            $\mathrm{BackTrack}_{i,k} \leftarrow \arg\max_{1 \le i' \le N_{Tr}} \{J_{i',k-1} - \lambda \mathcal{D}_D(\mathcal{C}^{Tr}(i'), \mathcal{C}^{Tr}(i))\}$
        end for
    end for

    $i \leftarrow \arg\max_{1 \le i' \le N_{Tr}} \{J_{i',N}\}$ ▷ Back-Tracking to Find $\mathcal{C}^*$
    $\mathcal{C}^*(N) \leftarrow \mathcal{C}^{Tr}(i)$
    for $k \leftarrow N$ down to 2 **do**
        $i \leftarrow \mathrm{BackTrack}_{i,k}$
        $\mathcal{C}^*(k-1) \leftarrow \mathcal{C}^{Tr}(i)$
    end for
return $\mathcal{C}^*$
end procedure

Figure: Estimating $\mathcal{C}^*$ using Dynamic Programming

ATB Prediction using Dynamic Programming

Smoothly varying ATB, $\mathcal{C}^*$, are estimated for a test rtMRI video sequence $\mathcal{I}$ by selecting the best contour from the training set $\mathcal{C}^{Tr}$ by maximizing the following objective function:

$$J(\mathcal{C},\mathcal{I}) = \sum_{k=2}^{N} \mathcal{D}_F(\mathcal{C}(k),\mathcal{I}(k)) - \lambda \mathcal{D}_D(\mathcal{C}(k),\mathcal{C}(k-1)) \quad (5)$$

$$\mathcal{C}^* = \{\mathcal{C}^*(k), 1 \leq k \leq N\} = \underset{\mathcal{C}\in\mathcal{C}^{Tr}}{\operatorname{argmax}}\{J(\mathcal{C},\mathcal{I})\} \quad (6)$$

where,

$$\mathcal{I} = \{\mathcal{I}(k),\ 1 \leq k \leq N\}$$
$$\mathcal{C}^{Tr} = \{\mathcal{C}^{Tr}(i),\ 1 \leq i \leq N_{Tr}\}$$
$$\lambda = \text{ Temporal Stiffness Factor}$$

FDM-based rtMRI Segmentation: Overview

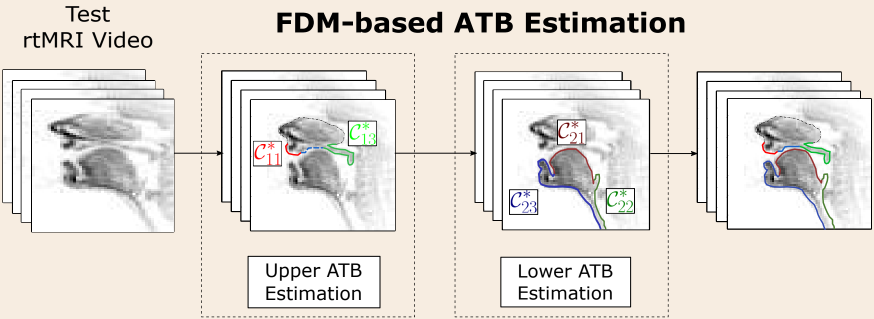


Figure: Order followed while performing ATB Estimation

Contour Stitching and Pruning

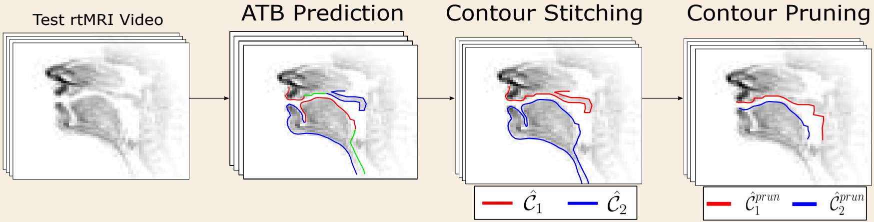


Figure: Stitching and Pruning of Predicted ATB

Section 5

Experimental Setup

- ATBs are estimated using five-fold cross-validation setup separately for each subject
- 8 training, 2 test rtMRI videos - round-robin fashion
- 5 training and 3 development videos in each fold
- $\mathcal{C}_1^{Tr}$ and $\mathcal{C}_2^{Tr}$ are obtained from the manually traced boundaries

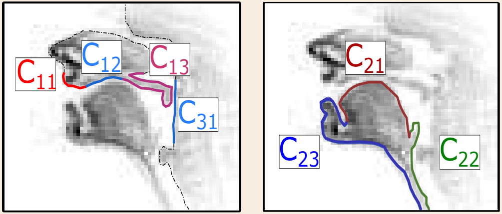


Figure: Manually traced ATBs

Evaluation of Predicted Contours

- Evaluation Measure: DTW Distance ($\mathcal{D}_D$) between predicted and manually traced ATBs
- Two kinds of evaluation performed:
  - Evaluation of Complete ATBs ($\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$)
  - Evaluation of Pruned ATBs ($\hat{\mathcal{C}}_1^{prun}$ and $\hat{\mathcal{C}}_2^{prun}$)

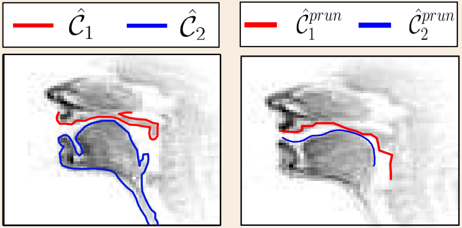


Figure: Evaluation schemes used for experiments

Results

Maeda Grid (MG) [4] based approach used as baseline for comparing with proposed FDM approach

	Lower ATB		Upper ATB	
Sub	**MG**	**FDM**	**MG**	**FDM**
F1	1.09 ± 0.22	1.02 ± 0.24	1.00 ± 0.17	0.95 ± 0.17
F2	1.28 ± 0.29	1.27 ± 0.26	1.42 ± 0.35	1.20 ± 0.22
M1	1.31 ± 0.57	1.25 ± 0.26	1.18 ± 0.19	1.10 ± 0.20
M2	1.38 ± 0.31	1.17 ± 0.28	1.37 ± 0.23	1.17 ± 0.24

Table: $\hat{\mathcal{C}}_1^{prun}$ and $\hat{\mathcal{C}}_2^{prun}$ prediction error in pixels (mean $\pm$ standard deviation)

[4] Jangwon Kim et al. "Enhanced Airway-Tissue Boundary Segmentation for real-time Magnetic Resonance Imaging Data" in International Seminar on Speech Production, ISSP, pp. 222–225, 2014.

Results

- Higher accuracy of proposed approach due to robustness of Fisher Discriminant Measure (FDM) to grainy noise
- Temporal constraint ensures smoothly varying contours across frames

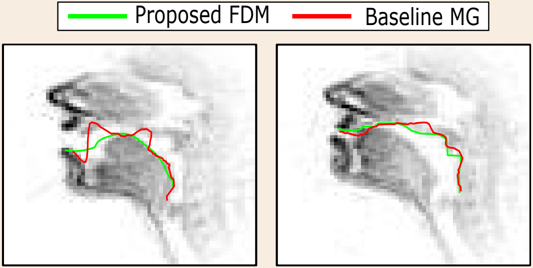


Figure: Improved accuracy of FDM as compared to Baseline MG

Results

- Value of FDM reduces if articulators come in contact with other tissue - may affect accuracy
- FDM can only provide best fitting contour from training set - not all configurations of articulators can be predicted

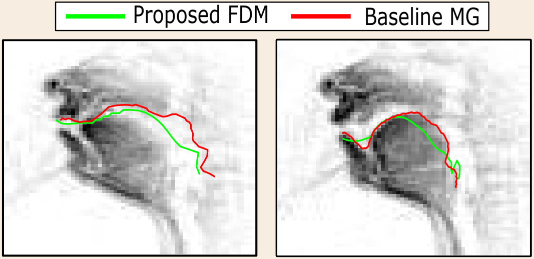


Figure: Shortfalls of FDM

Section 6

Conclusion and Future Works

- ATB shapes learned from training data - "best fit" approach
- Temporal continuity of ATBs ensured across successive frames
- Further improvement in accuracy possible - deform-able model for estimated ATBs

Thank you

Advait Koparkar

advaitkoparkar@gmail.com

The authors thank the Pratiksha Trust for their support.

Appendix: Contour Stitching

Parts of Upper and Lower ATBs ($\mathcal{C}^*$) are stitched to form smooth contours $\hat{\mathcal{C}}_1$ & $\hat{\mathcal{C}}_2$

- $\hat{\mathcal{C}}_1$ obtained by concatenating $\mathcal{C}^*_{11}$ (Upper Lip), $\mathcal{C}_{12}$ (Hard Palate) and $\mathcal{C}^*_{13}$ (Velum)

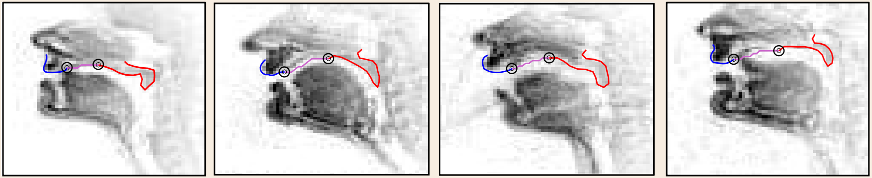


Figure: Stitching $\mathcal{C}_{11}$, $\mathcal{C}_{12}$ and $\mathcal{C}_{13}$

Appendix: Contour Stitching

- $\hat{\mathcal{C}}_2$ obtained by stitching contours $\mathcal{C}^*_{21}$ (Tongue Tip), $\mathcal{C}^*_{22}$ (Tongue Root) and $\mathcal{C}^*_{23}$ (Lower Lip)
- Continuity of $\hat{\mathcal{C}}_2$ maintained at junctions of $\mathcal{C}^*_{21}$, $\mathcal{C}^*_{22}$ and $\mathcal{C}^*_{23}$

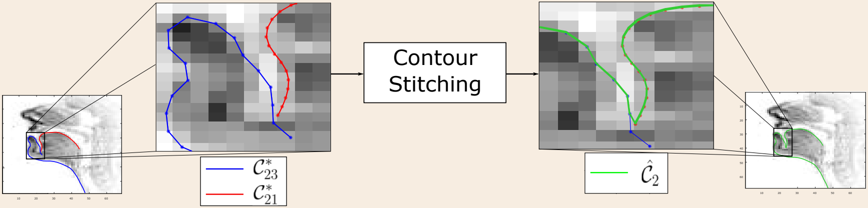


Figure: Stitching of lower contours to obtain $\hat{\mathcal{C}}_2$

Appendix: Contour Pruning

$\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ are pruned to obtain boundaries within the vocal tract

- $\hat{\mathcal{C}}_1$ pruned from UL to VEL tip and concatenated with $\mathcal{C}_{31}$ till GLTB to obtain $\hat{\mathcal{C}}_1^{prun}$

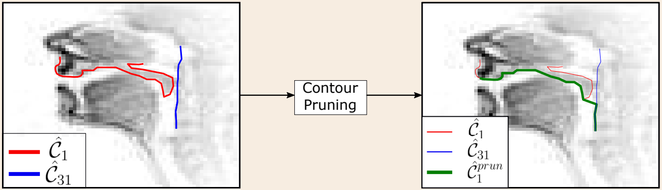


Figure: Contour Pruning for $\hat{\mathcal{C}}_1$

Appendix: Contour Pruning

- $\hat{\mathcal{C}}_2$ pruned from Lower Lip (LL) to Glottis Begin (GLTB)
- Segment of $\hat{\mathcal{C}}_2$ near tongue base ($\mathcal{C}_{tb}$) replaced by a smooth boundary denoted by $\mathcal{C}_{sm}$

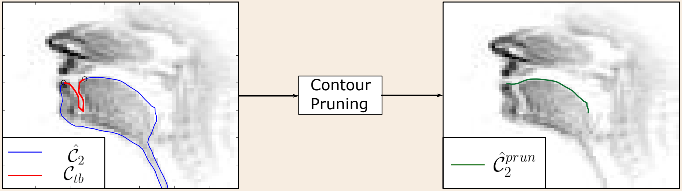
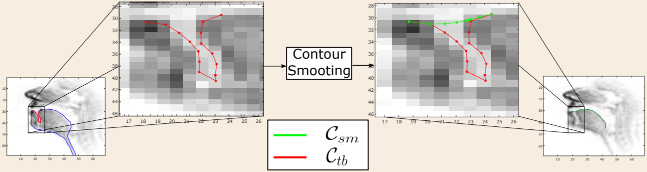


Figure: Contour Pruning for $\hat{\mathcal{C}}_2$

Appendix: Contour Pruning



Algorithm 2 Smoothing $\mathcal{C}_{tb}$ to obtain $\mathcal{C}_{sm}$

procedure SmoothContour($\mathcal{C}_{tb}$)

Given, $\mathcal{C}_{tb} = \{(x_i^{tb}, y_i^{tb}), 1 \leq i \leq N_{tb}\}$

Assume, $\mathcal{C}_{sm} = \{(x_i^{tb}, y_i^{sm}), 1 \leq i \leq N_{tb}\}$

$$y_i^{sm} \triangleq a_0 + a_1 x_i^{tb} + a_2 (x_i^{tb})^2$$

$$\{a_0, a_1, a_2\} = \underset{\alpha,\beta,\gamma}{\operatorname{argmin}} \sum_{i=1}^{N_{tb}} (y_i^{tb} - (\alpha + \beta x_i^{tb} + \gamma (x_i^{tb})^2))^2$$

$$\text{subject to } \alpha + \beta x_i^{tb} + \gamma (x_i^{tb})^2 \leq y_i^{tb}, \forall i$$

return $\mathcal{C}_{sm}$

end procedure

Figure: Smoothing $\mathcal{C}_{tb}$ to obtain $\mathcal{C}_{sm}$

Appendix: Full ATB Evaluation

- Values indicate average euclidean distance (in pixels) between points of predicted contour and ground truth
- Manually traced ATBs ($\mathcal{C}_1^{Tr}$ and $\mathcal{C}_2^{Tr}$) used as ground truth for evaluating $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$

Sub	**Lower ATB**	**Upper ATB**
F1	0.93 ± 0.13	0.92 ± 0.12
F2	0.99 ± 0.17	1.09 ± 0.19
M1	0.98 ± 0.16	1.13 ± 0.18
M2	0.98 ± 0.18	1.17 ± 0.25

Table: $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ prediction error in pixels (mean $\pm$ standard deviation)