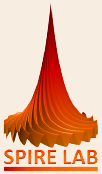
An error correction scheme for improved air-tissue boundary in real-time MRI video for speech production

Anwesha Roy, Varun Belagali, Prasanta Kumar Ghosh

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



ICASSP 2022

Overview

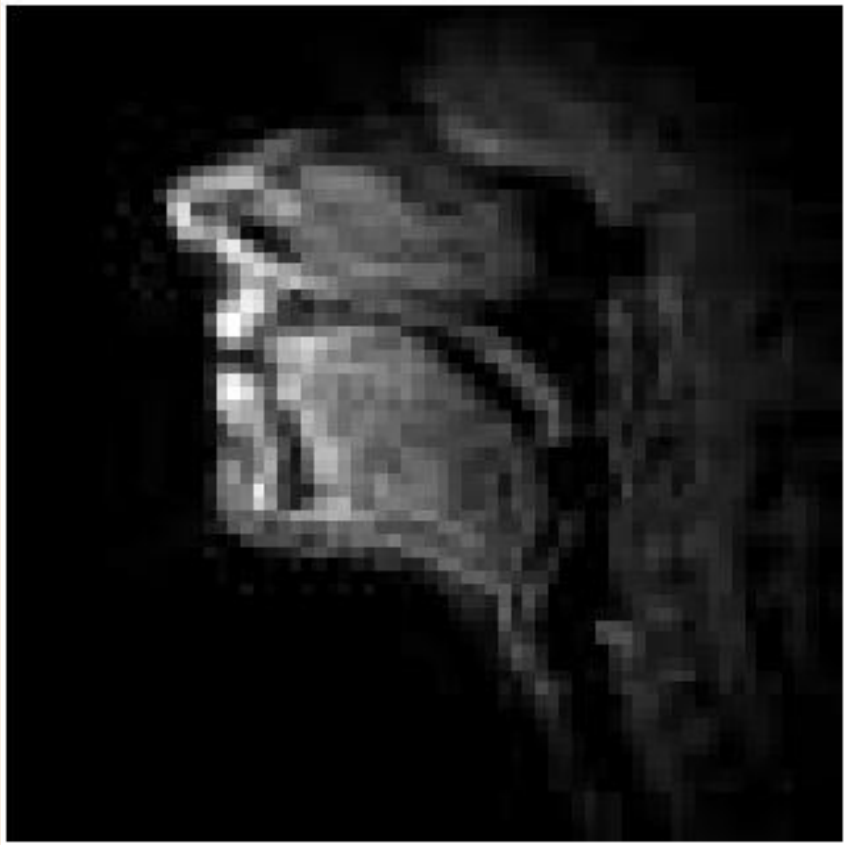
Real-time Magnetic Resonance Imaging (rtMRI)

- Real-time Magnetic Resonance Imaging (rtMRI) is a safe and non-invasive imaging method which captures a complete picture of the vocal tract
- A common step before using these rtMRI videos is obtaining the Air-Tissue Boundary (ATB) segmentation in every frame

1. Hagedorn, Christina, et al. "Engineering innovation in speech science: Data and technologies." Perspectives of the ASHA Special Interest Groups 4.2 (2019): 411-420.

2. Bresch, Erik, et al. "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]." IEEE Signal Processing Magazine 25.3 (2008): 123-132.

Air Tissue Boundary (ATB)



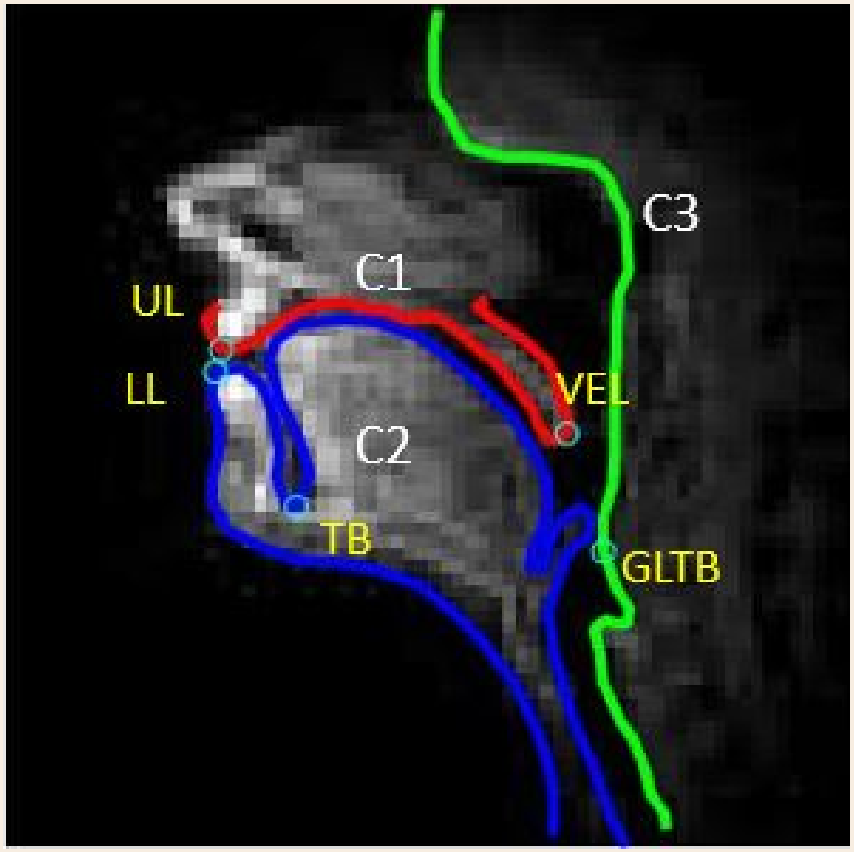


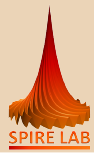
Figure: Illustration of (a) a rtMRI frame, (b) three manually annotated contours: contour1 (C1), contour2 (C2) and contour3 (C3)

Motivation

- Best ATB segmentation scheme in literature: **3D-CNN**[1]
- Evaluation for ATB segmentation done in the past using Dynamic Time Warping (**DTW**) distance between complete annotation & prediction
- DTW may not capture local errors in the entire contour
- We propose:
  - detection & correction schemes for local errors
  - new regional evaluation metrics

1. Mannem, Renuka, Navaneetha Gaddam, and Prasanta Kumar Ghosh. "Air-Tissue Boundary Segmentation in Real Time Magnetic Resonance Imaging Video Using 3-D Convolutional Neural Network." INTERSPEECH. 2020.

Types of errors



- Errors in **C1**
  - **incomplete contour** - velum (**VEL**) portion is missing
  - **frame error** - entire C1 has defects
- Errors in **C2**
  - **TB error** - tongue base (**TB**) dip not predicted properly
    TB error occurs mainly because of the low number of pixels with low intensity present in the dip region
  - **frame error** - entire C2 has defects

Observed errors

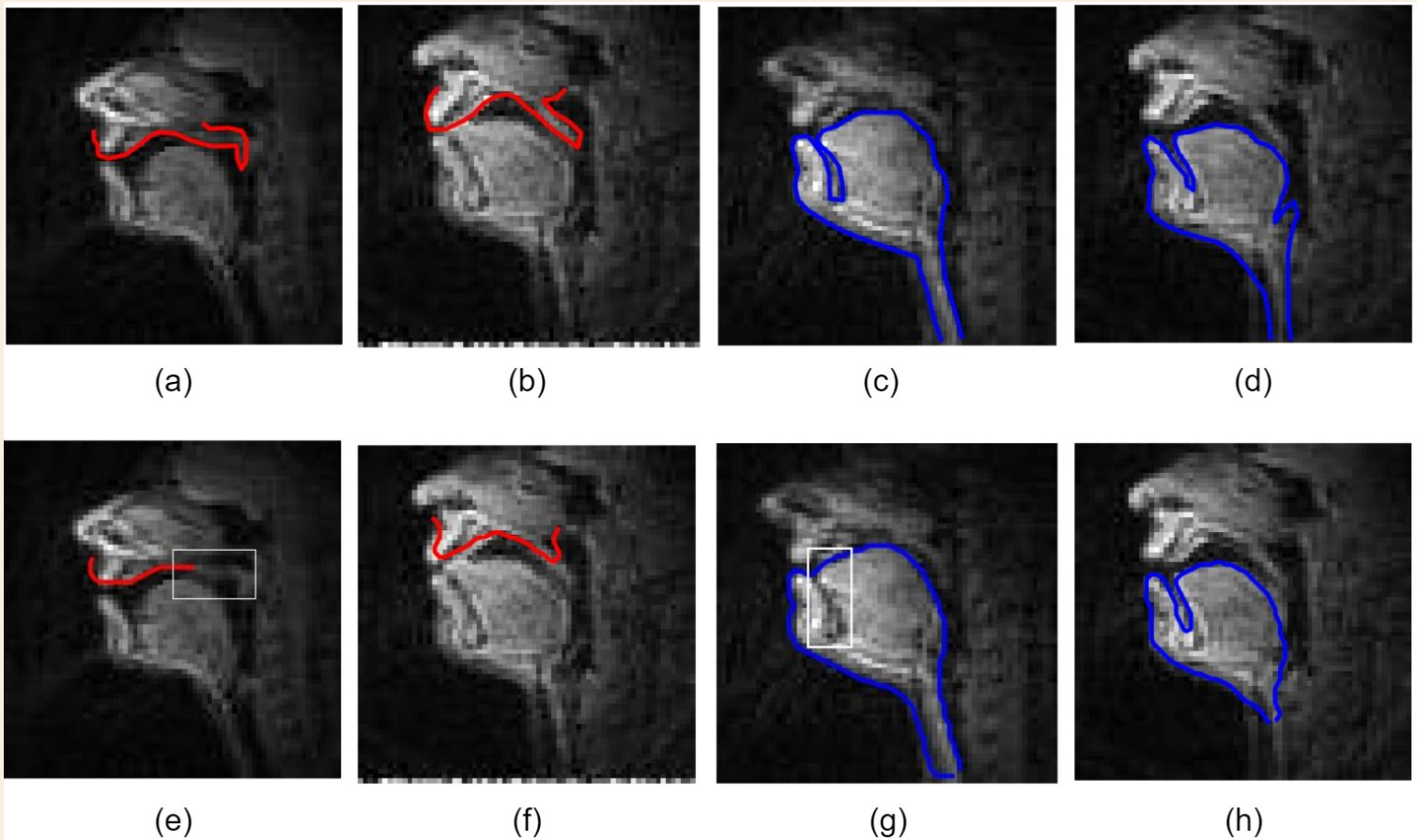
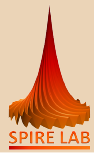


Figure: Manual annotations (a,b,c,d) and corresponding erroneous predictions for C1 incomplete (e), C1 frame (f), C2 TB (g) and C2 frame (h) errors

Overview

Dataset Description



- **USC-TIMIT** corpus[1] is used in this work:
  - rtMRI videos of the upper airway in the mid-sagittal plane
  - 5 female (F1, F2, F3, F4, F5) and 5 male (M1, M2, M3, M4, M5) subjects
  - Each of them speak 460 sentences from MOCHA-TIMIT database[2]
  - Frame rate is 23.18 frames/sec & spatial resolution is 68×68 (pixel dimension of 2.9mm $\times$ 2.9mm).
- 3D-CNN trained on 90 videos (9 videos from each subject)
- ATBs predicted on 100 videos (10 videos from each subject) not seen in training

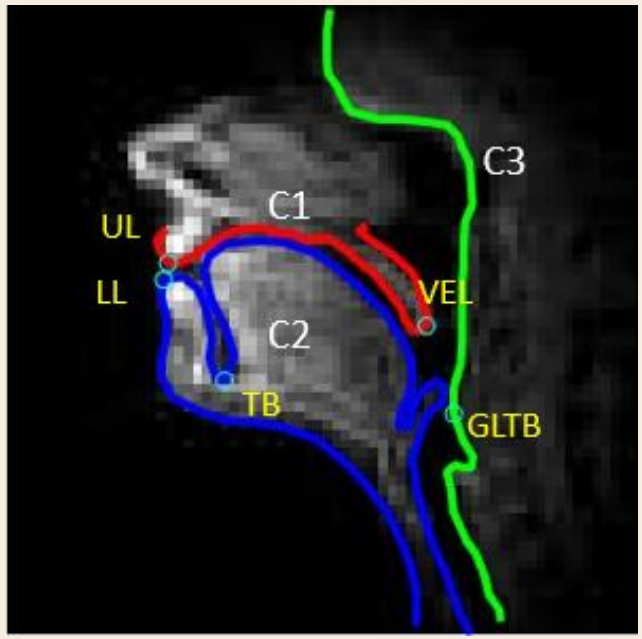
---

1. Narayanan, Shrikanth, et al. "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)." The Journal of the Acoustical Society of America 136.3 (2014): 1307-1311.

2. Wrench, Alan A. "A multichannel articulatory database and its application for automatic speech recognition." In Proceedings 5 th Seminar of Speech Production. 2000.

Manual annotation

- ATBs are manually annotated using a MATLAB GUI[1]. 3 contours and 5 points are marked.



- **Contour1 (C1)** starts from UL, goes through the hard palate till VEL and goes around the fixed nasal tract
- **Contour2 (C2)** covers the jawline, LL, tongue blade and extends below the epiglottis
- **Contour3 (C3)** marks the pharyngeal wall
- The 5 points - upper lip (**UL**), lower lip (**LL**), tongue base (**TB**), velum (**VEL**) and glottis begin (**GLTB**)

1. Pattem, Ashok Kumar, et al. "Optimal sensor placement in electromagnetic articulography recording for speech production study." Computer speech language 47 (2018): 157-174.

Overview

Observations

- **MATLAB GUI** developed to observe both annotation and prediction in each frame
- If prediction deviates a lot in any region from the annotation, frame is declared **erroneous** and the defective contour is noted (C1 or C2)
- Predicted C3 not found to have any observable defects
- Observations cross-checked by an unbiased viewer
- Erroneous frames selected based on subjective criteria considered as **ground truth** for error classification

Lineplot

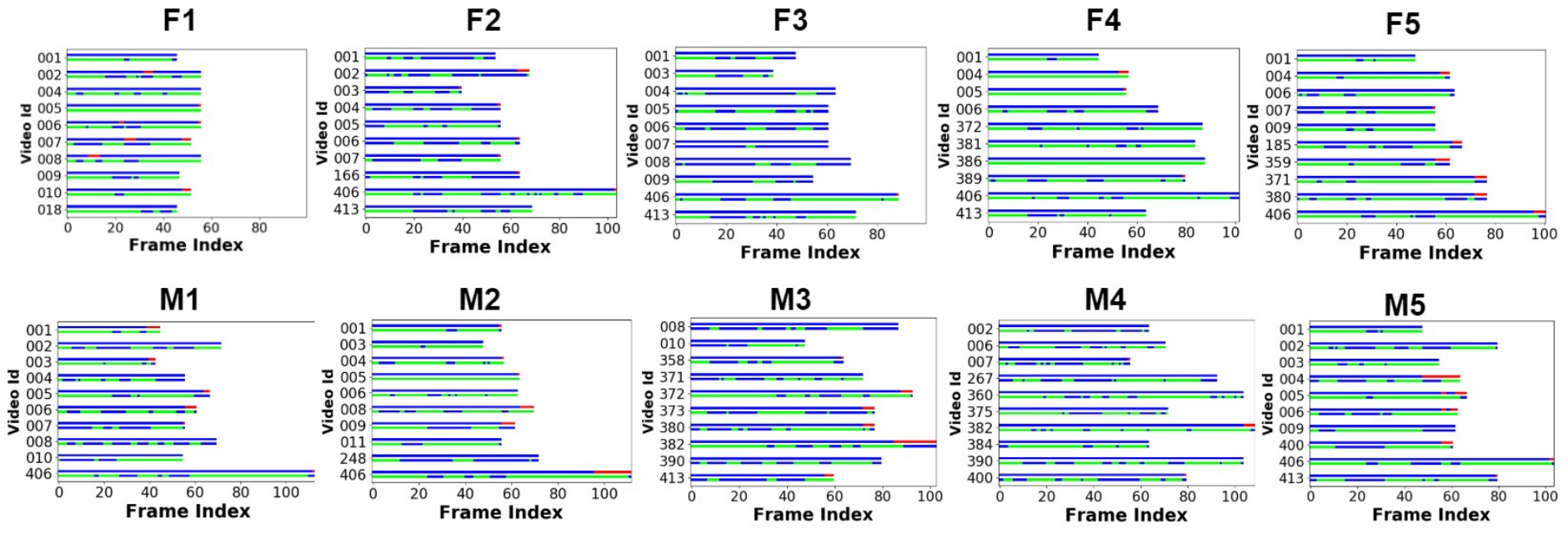


Figure: Illustration of position of error frames in each video for all subjects. There are two line plots for each video index, where each line shows correct frames in blue. C1 errors are shown in red in first line and C2 errors in green in second line

Lineplot

- It is observed that C2 (especially near TB) errors occur repeatedly throughout the video.
- C1 errors are mostly observed at the end of videos. This may be due to the end frame padding done in 3D-CNN model during prediction.

Contour1

- Subjective analysis shows **207 frames (3.07 %)** have C1 error
- Mean $\pm$ standard deviation (std) DTW distance between annotated & predicted C1 for error frames is 2.15 $\pm$ 1.48 pixels compared to 1.11 $\pm$ 0.18 pixels for correct frames
- But range of DTW distance of erroneous and correct frames overlap

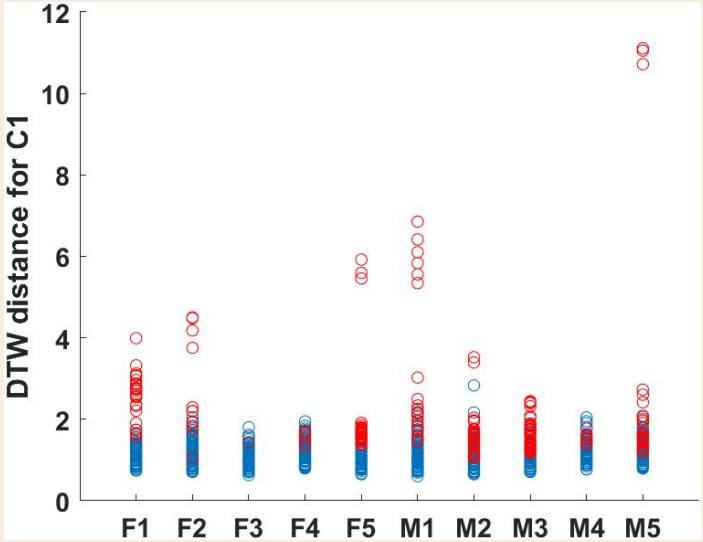


Figure: DTW analysis plot for C1 over all frames, where red bubbles represent erroneous frames and blue are correct

Contour2

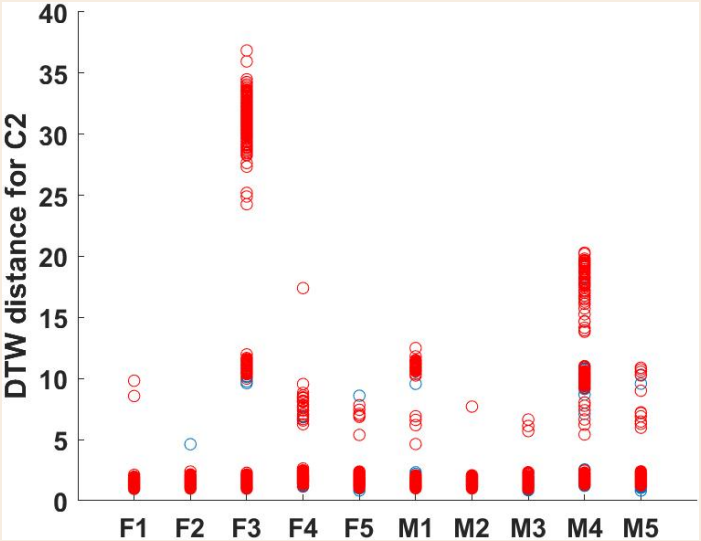


Figure: DTW analysis plot for C2 over all frames, where red bubbles represent erroneous frames and blue are correct

- **325 frames (4.82 %)** have C2 frame error & **65.73 %** of the frames have C2 TB error
- Mean $\pm$ std DTW dist for error frames is 3.61 ± 3.49 pixels and for correct frames is 1.92 ± 1.89 pixels
- High variability in global DTW distance value, which does not reflect C2, especially TB dip, errors

Overview

Summary

	Error type	Evaluation metric	Detection method	Correction method
Contour1	Incomplete	EVEL, VELrDTW	Deviation from mean VEL, VEL to pharyngeal wall dist	Interpolation + Appending
	Frame			Interpolation
Contour2	TB	ETB, TBrDTW	LL to TB slope, LL to TB distance, Combined	Otsu thresholding + Contour warping
	Frame		No. of points	Interpolation

Summary of proposed metric, detection and correction schemes

Experimental setup

- **4 fold cross-validation** set-up where, in each fold, 3 randomly chosen subjects are taken for validation and the rest of the 7 subjects for test
- VEL point predicted by finding dip in the row index of the contour in the region around velum in C1
- TB in C2 found by locating the dip in the region between LL and uppermost point on tongue

Evaluation metrics

- **Detection:** F-score used as metric
- **Correction:**
  - **Contour1**
    - **VELrDTW** distance: DTW dist of C1 around the VEL region, taking 30% of total number of points of C1 on the pharyngeal wall side
    - **EVEL**: Euclidean distance between predicted and annotated VEL point
  - **Contour2**
    - **TBrDTW** distance: DTW dist of C2 in the region between LL and uppermost tongue point
    - **ETB**: Euclidean distance between predicted and annotated TB point
  - **Global DTW** is also reported for both contours

Error Detection

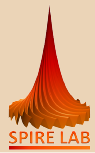
- We propose different methods of error detection for C1 based on relative position of VEL point
- Detection is done for C2 on the basis of position of TB and number of points
- Hyperparameters used by the methods in each fold are selected based on F-score achieved on validation set

C1 detection

- **Deviation from mean VEL**
  - Euclidean distance between VEL in a frame and mean VEL found
  - Thresholds applied based on validation data are 3.5, 4, 4, 4 for 4 folds
  - F-score of **0.86 ($\pm$ 0.01)** on test set

- **Distance of VEL from pharyngeal wall**
  - Nearest point on C3 from VEL found and fixed for each subject from manual annotations
  - Thresholds applied on distance between this point and VEL, based on validation data, are 8, 8, 8, 7.5 for 4 folds
  - F-score of **0.85 ($\pm$ 0.02)** on test set

C1 detection



- **Combined**
  - Frame declared erroneous if it satisfies either one of the two aforementioned error criteria
  - F-score of **0.86 ($\pm$ 0.02)** achieved on test set

C2 detection

Number of points

- Number of points of predicted C2 low
- Average no. of points in C2 over all videos in validation set found
- Threshold set to 65% of it to find erroneous frames
- F-score of **0.941 ($\pm$ 0.02)** achieved

LL to TB slope

- For TB errors, slope of line joining TB and lower lip observed to be low
- Thresholds on this slope, selected based on validation data, are 0.7, 0.7, 0.8, 1 for 4 folds
- F-score of **0.85 ($\pm$ 0.02)** on test set

C2 detection

- **LL to TB distance**
  - For TB errors, distance from the lower lip to TB is short
  - Thresholds on distance, based on validation data, are 8, 7, 10, 10 for 4 folds
  - F-score of **0.88 ($\pm$ 0.02)** on test set
  - Distance thresholding performs better than slope
- **Combined**
  - Frame declared erroneous if it satisfies either one of the three aforementioned error criteria
  - F-score of **0.90 ($\pm$ 0.02)**

C1 correction

- Frame to frame variance low because of temporal continuity of 3D-CNN
- For all error frames detected using the **combined method**, C1 generated by **linear interpolation** using neighbouring frame contours
- **Incomplete C1 errors:**
  - Section-wise DTW dist analysis shows that incomplete part of the predicted C1 is actually correct and rest is missing
  - End point of original C1 found on interpolated one & rest of the interpolated C1 appended to the existing contour
- **C1 Frame errors:**
  The interpolated contour is taken completely

C2 correction

- For all detected **C2 frame errors**, entire C2 is generated by linear interpolation. These frames & error frames detected by combined method considered for C2 TB correction
- TB observed to be within a **15 $\times$ 20 patch** between lip and tongue
- **Otsu thresholding**[1] done on this patch to find darker pixels that lie within C2. Lowest point in class-0 region of binary image within C2 is marked as **corrected TB**
- We adjust C2 in vicinity of 3D-CNN predicted TB location mapping shift of neighbouring points in a gradient based fashion to find corrected C2 in TB dip region

1. Otsu, Nobuyuki. "A threshold selection method from gray-level histograms." IEEE transactions on systems, man, and cybernetics 9.1 (1979): 62-66.

Illustration

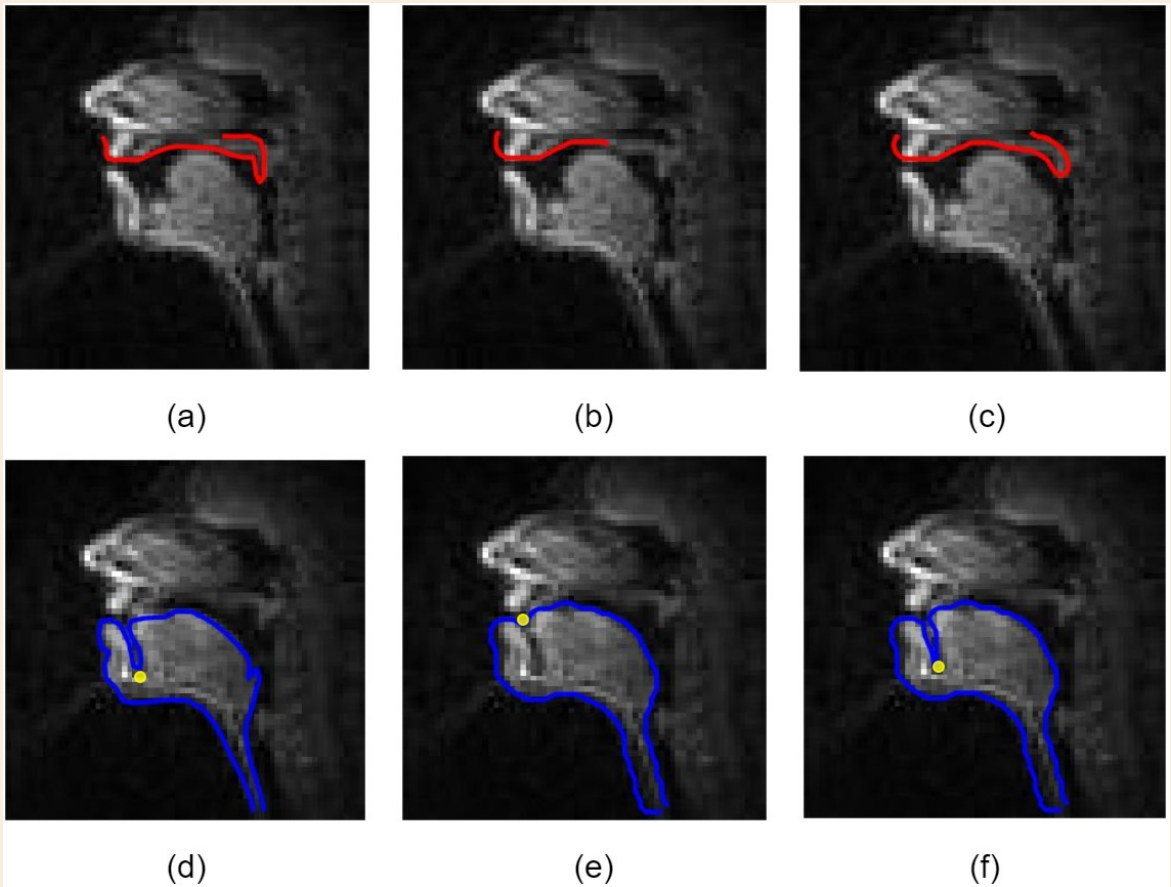


Figure: Annotations (a,d), erroneous predictions (b,e) and corrected contours (c,f) for VEL and TB errors respectively, where yellow point represents TB

Results

	Evaluation Metric	Pre-correction	Post-correction
C1	EVEL	8.10 ± 2.33	3.09 ± 1.34
	VELrDTW	4.12 ± 1.56	1.59 ± 0.43
	DTW	2.04 ± 1.19	1.13 ± 0.19
C2	ETB	11.31 ± 3.40	3.64 ± 2.71
	TBrDTW	4.26 ± 1.26	3.05 ± 1.06
	DTW	2.06 ± 1.22	1.98 ± 1.32

Table: Mean $\pm$ standard deviation of evaluation metrics (in pixels) before and after correction for C1 and C2

Results

- **Contour1**
  - **EVEL** decreases by 61.8% after correction, whereas **VELrDTW** improves by 61.4% for these frames
  - For all frames, **DTW** distance over entire C1 improves by 44.6% after correction, which is not as significant as the change in VELrDTW
- **Contour2**
  - **ETB** and **TBrDTW**, improve by 67.8% and 28.4% respectively, after correction
  - Global **DTW** distance, on the other hand, does not show any significant improvement

Overview

Key Takeaways

- Careful analysis reveals different types of errors are present in the results of 3D-CNN model like VEL or TB region defects
- Automatic methods are proposed to **detect** and **correct** such observed errors
- Further, **region specific metrics** are proposed for evaluation of the quality of the predicted and corrected contours
- The proposed methods show observable refinement of ATBs, which is reflected in the improvement in proposed metrics

Future Work

- Robust neural network approaches using region specific loss functions, which target specific problems in particular contour regions
- Active Appearance Models[1] (**AAM**) for correction of regional errors like TB or VEL
- Performance of these corrected ATBs when used for different downstream applications

1. Cootes, Timothy F., Gareth J. Edwards, and Christopher J. Taylor. "Active appearance models." European conference on computer vision. Springer, Berlin, Heidelberg, 1998.

THANK YOU

Have Questions/Suggestions?
Write to us @ `spirelab.ee@iisc.ac.in`