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

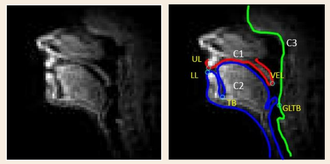
Anwesha Roy, Varun Belagali, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bangalore, India-560 012

Introduction

- **Air-tissue boundary (ATB)**
  - Air-Tissue Boundary (ATB) segmentation is a common pre-processing step before rtMRI videos are applied in different domains like text-to-speech synthesis, speaker verification, and visual augmentation for synthesized articulatory videos.

Figure illustrates an rtMRI frame, and the corresponding ATB, including 3 contours: contour1 (C1), contour2 (C2) and contour3 (C3), & 5 points: upper lip (UL), lower lip (LL), tongue base (TB), velum (VEL), glottis begin (GLTB).

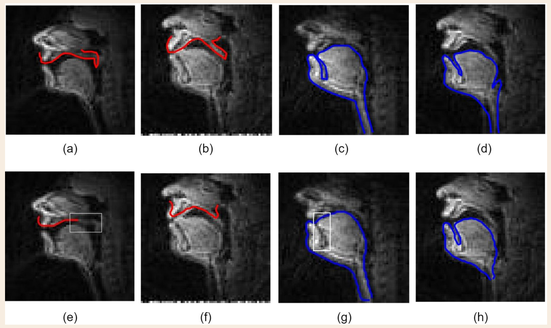


- **Motivation**:
  - 3D CNN [1] gives best performance in literature. But global Dynamic Time Warping (DTW) distance is used as evalulation metric, which might not capture regional errors
  - In this work, we analyze such errors, & propose a novel detection and correction scheme

Data set

- **USC-TIMIT** 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) subject, each speaking 460 sentences from MOCHA-TIMIT database
  - Frame rate is 23.18 frames/sec & spatial resolution is 68 × 68 (pixel dimension of 2.9mm × 2.9mm)
- 3D-CNN trained on 90 videos (9 videos from each subject)
- ATBs predicted on 100 videos (10 from each subject), not seen in training, are used for error detection and correction.

Error Types



- Contour1 errors observed to be of 2 types - incomplete contours, where VEL portion is missing (Fig. e), & frame errors, where the entire C1 has defects (Fig. f)
- Contour2 errors observed to be of 2 types - TB error, where TB dip is not predicted properly (Fig. g), & frame error, where entire frame is wrongly predicted (Fig. h)

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

Analysis and Metrics

- **Analysis**:
  - MATLAB GUI is developed to observe both annotation and prediction in each frame. A frame is labelled as erroneous if the prediction deviates a lot from annotation, in any region .
  - For C1, mean ± std DTW distance between annotated and predicted contour for error frames is 2.15 ± 1.48 pixels, and 1.11 ± 0.18 pixels for correct frames.
  - For C2, mean ± std DTW distance for error frames is 3.61 ± 3.49 pixels and for correct frames it is 1.92 ± 1.89 pixels.
  - Even though mean DTW distance is higher for error frames, range of DTW distance of erroneous and correct frames overlap.
  - Global DTW distance does not reflect regional errors. Hence, new region-specific metrics metrics are proposed.
- **Proposed evaluation metrics**:
  - Contour1:
    - **VELrDTW**: DTW dist. between annotated & predicted 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**: TBrDTW distance: DTW dist. between annotated & predicted C2 in region between LL and uppermost tongue point.
    - **ETB**: Euclidean distance between predicted and annotated TB point

Error Detection

- C1 Detection: Threshold applied on:
  - Euclidean distance between VEL in a frame and mean VEL across all frames in the video
  - Euclidean distance between VEL and nearest point on C3 (fixed for a subject) in a frame
- C2 Detection: Threshold applied on:
  - Number of points in C2 (for frame error)
  - Slope of the line joining LL and TB
  - Euclidean distance between LL and TB

Error Correction

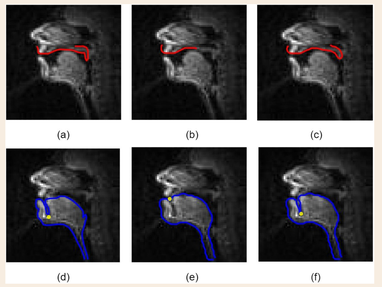
- C1 Correction:
  - For detected error frames, C1 is generated by linear interpolation using neighbouring frame contours.
  - For incomplete C1 errors, VEL part of interpolated C1 is appended to existing contour
  - For C1 frame errors, the interpolated contour is taken completely.
- C2 Correction:
  - For all detected C2 frame errors, entire C2 is generated by linear interpolation.
  - For frames with TB error, otsu thresholding is used in 15 X 20 patch, between LL and tongue, to find the corrected TB. C2 is adjusted in TB dip region by contour warping in a gradient based fashion.

Results

Table: Mean ± std of evaluation metrics (in pixels) before and after correction for C1 and C2

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

- For C1, EVEL decreases by 61.8% after correction, whereas VELrDTW improves by 61.4%.
- For C2, ETB and TBrDTW, improve by 67.8% and 28.4% respectively, after correction.
- Gloabl DTW also shows slight improvement for C1 & C2.



- Fig. (a), (b) and (c) illustrate manual annotation, incomplete C1 error and corresponding corrected contour.
- Fig. (d), (e) and (f) illustrate manual annotation, TB dip error and corresponding corrected contour, where yellow point represents TB.

Conclusion

- Automatic methods are proposed to detect and correct regional errors in 3D-CNN. New regional evaluation metrics are also proposed for evaluation of the quality of the predicted ATBs.
- **Future work** : Robust neural network approaches using region specific loss functions, which target specific problems in particular contour regions.

References

[1] R. Mannem, et. al., “Air-tissue boundary segmentation in real time magnetic resonance imaging video using 3-d convolutional neural network.” INTERSPEECH 2020.

anwesharoy@iisc.ac.in, varunbelagali98@gmail.com, prasantg@iisc.ac.in