USC University of Southern California

USC Viterbi School of Engineering

Ming Hsieh Department of Electrical Engineering

Spatial and temporal alignment of multimodal human speech production data: real time imaging, flesh point tracking and audio

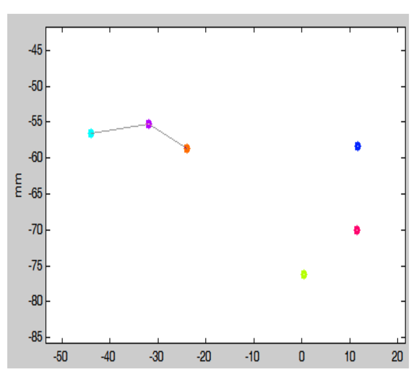
Jangwon Kim, jangwon@usc.edu, Electrical Engineering Dept. Signal Analysis and Interpretation Laboratory (SAIL)

Motivation & Introduction

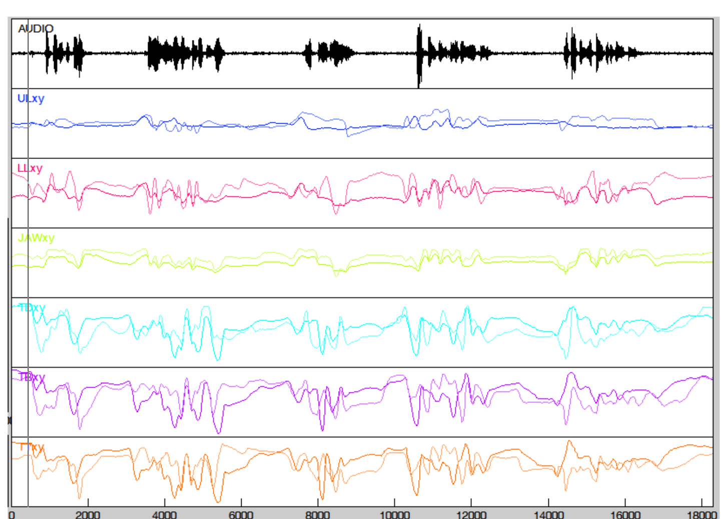
Objective: to obtain detailed vocal tract dynamics from MRI video aligned with EMA sensor trajectories

Data characteristics:
20 sentences by 1 speaker.

MRI video



EMA sensors



EMA sensor trajectories

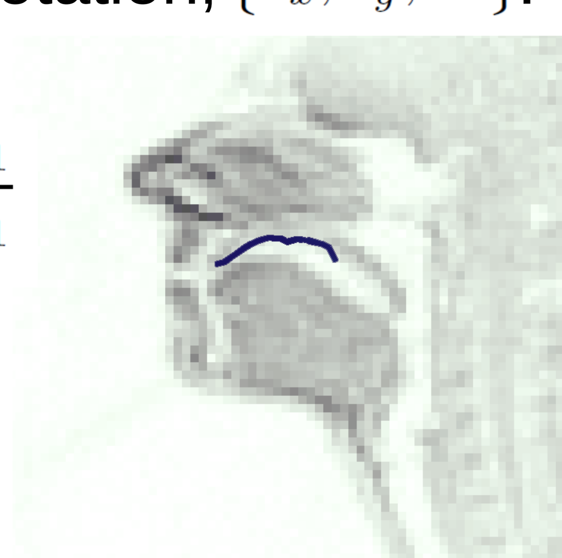
MRI video	EMA sensors
Pixel image sequence	Motion captures
(2D) complete midsaggital view 68×68 pixels	(3D) 6 sensor trajectories => 18 dimension
23.180 samples/sec	100 samples/sec

Spatial alignment

Method: align the hard-palate tracking of EMA and hard palate boundary of MRI by grid search over variety of translation along x and y axes, and rotation, $\{\delta_x, \delta_y, \theta\}$.

$$\{\delta_x^\star, \delta_y^\star, \theta^\star\} = \arg\max_{\delta_x, \delta_y, \theta} \sum_{\forall i,j \in palate\ trace} \frac{p_{i,j-1}}{p_{i,j+1}}$$

where $p_{i,j}$ is pixel intensity at (i,j) of standard deviation MRI matrix.



Aligned palate trace

Conclusion & Future Work

Conclusion:

•JAATA generates the best MRI image regions from which the EMA-like articulatory features.

•Average phone boundary difference is improved from 50.101 msec (MFCC) to 44.198 msec (JAATA): 12%↑

Future work:

•More flexible specifications (size, shape, numbers) of automatic pixel region selection

•New articulatory features of MRI image fitting better to articulatory features of EMA

Temporal alignment by JAATA

Joint Acoustic-Articulatory based Temporal Alignment:

•Temporal alignment with automatic feature extraction

•Temporal alignment: DTW with articulatory features + acoustic features (MFCC)

•Automatic feature extraction: to find the square pixel regions on MRI images whose behaviors are most similar to each EMA sensor trajectory in terms of Euclidean distance of their (smoothed) derivatives.

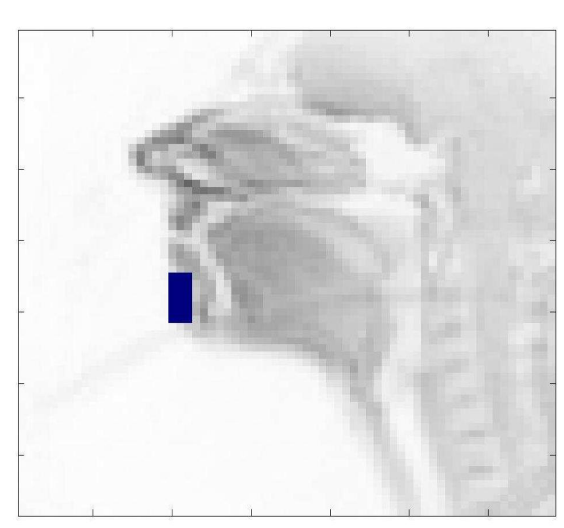
•To alternate DTW and feature extraction until the cost of a optimization cost function converges.

$$J(\lambda, \{\mathbf{W}_{M,f}, \mathbf{W}_{E,f}\}, \{\mathbf{s}_{q,M}, 1 \le q \le 12\})$$

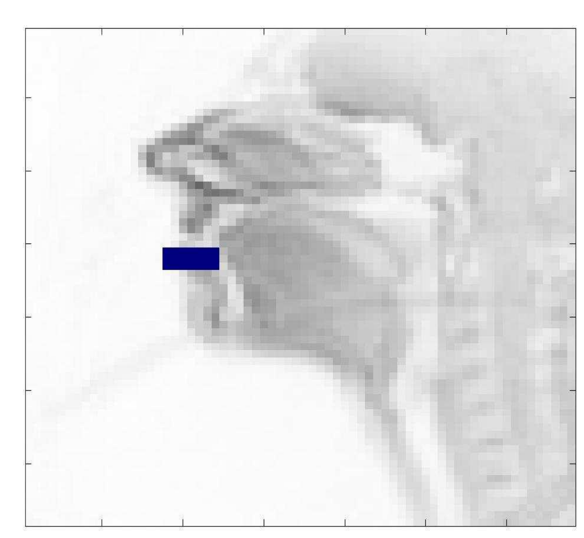
$$= \sum_{f=1}^{F} \left\{ \lambda \left(\left|\left| \mathbf{X}_{M,f}\mathbf{W}_{M,f} - \mathbf{X}_{E,f}\mathbf{W}_{E,f} \right|\right|_F^2 \right) + (1-\lambda) \left(\sum_{q=1}^{12} \left|\left| \frac{1}{A} \mathbf{s}_{q,M}^{\mathrm{T}} \mathbf{Y}_{M,f}\mathbf{W}_{M,f} - \left(\mathbf{z}_{E,f}^{q}\right)^{\mathrm{T}} \mathbf{W}_{E,f} \right|\right|^2 \right) \right\}$$

where **W** is path matrix, X is acoustic feature sequence, **f** denotes f-th sentence, ***M*** denotes MRI, ***E*** denotes EMA, **Y** is MRI video sequence, **s** is making matrix (non-zero elements selects a submatrix), λ is weighting term.

Automatic feature extraction results:

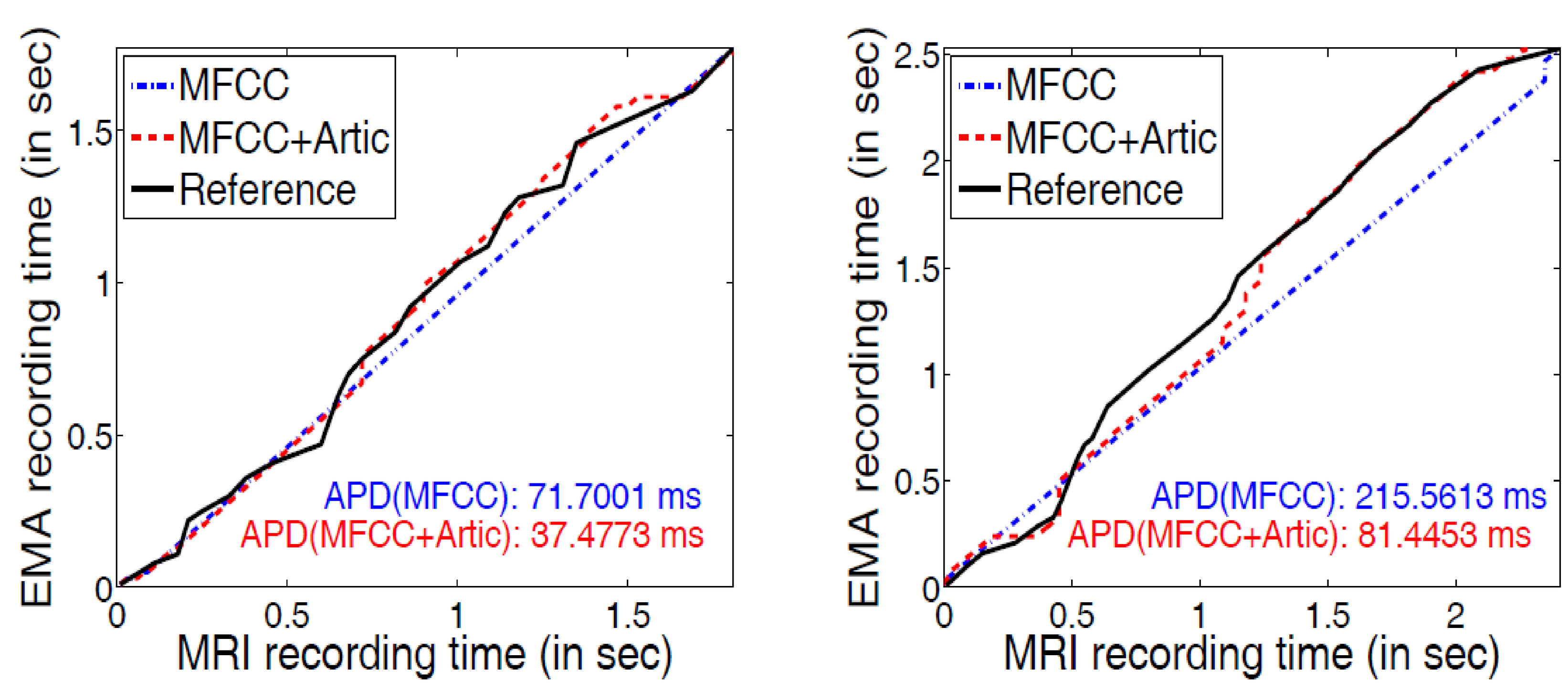


Pixel box for LIx(rho=0.68)



Pixel box for LLy(rho=0.67)

Temporal alignment results:



(a) Sentence 3 (b) Sentence 19

DTW alignment maps of two sentences

Contact information, sponsors etc.

Ming Hsieh Institute
Ming Hsieh Department of Electrical Engineering