Co-registration of articulographic and real-time magnetic resonance imaging data for multimodal analysis of running speech

Jangwon Kim, Adam Lammert, Michael Proctor, Shrikanth Narayanan
jangwon@usc.edu,lammert@usc.edu, mike.i.proctor@gmail.com, shri@sipi.usc.edu

[Supported by NIH and NSF grants]

Introduction

Motivation:

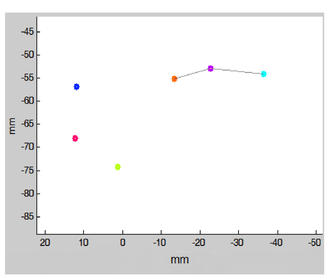
- Electromagnetic articulography (EMA) and real-time MRI (rtMRI) provide complementary imaging advantages
  - EMA: higher temporal resolution & flesh-point tracking
  - rtMRI: higher spatial resolution & complete midsagittal view
- Combine advantages of each modality by fine-grain spatial and temporal alignment in order to maximize preserved information (e.g., token-wise variability)

Potential applications:

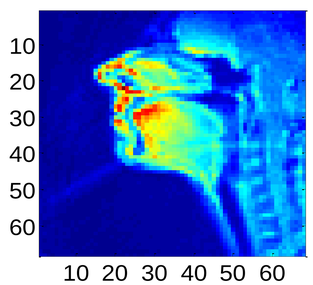
- Assessment of completed tongue contours from EMA
- Improved palate traces for EMA (and other structural characteristics)
- Estimated flesh-point tracking in rtMRI
- Improved interpolation of rtMRI video frames

Experimental setup

Database: EMA TIMIT & rtMRI dataset



(a) EMA sensors



(b) MRI image

- EMA and rtMRI data are recorded with the same stimuli (460 sentences from Mocha TIMIT database) in different time.
- Articulatory data + parallel speech waveform
- Electromagnetic articulography (EMA):
  - 6 sensor trajectories: tongue tip, tongue body (TB), tongue dorsum (TD), upper lip, lower lip(LL), lower incisor (Jaw)
  - frame rate: 100 frame/sec
- Real-time MRI:
  - 68 × 68 pixels in the midsagittal plane of head and neck
  - frame rate: 23.180 frame/sec

Alignment methods

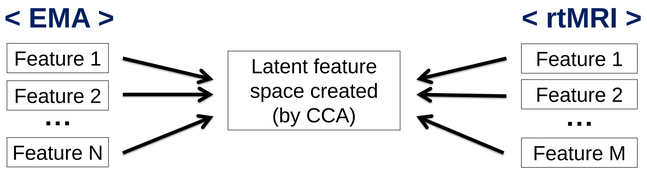
Spatial alignment: Align EMA palate trace to rtMRI image

- Manual initialization of placement
- Brute-force search for optimal rigid transformation
- Objective: maximize contrast across palate trace

Obj. fn.: $\arg\min_i f(i) = p_{i-1,j} / p_{i+1,j}$, $p_{i,j}$ is a pixel at *(i,j)* in the image

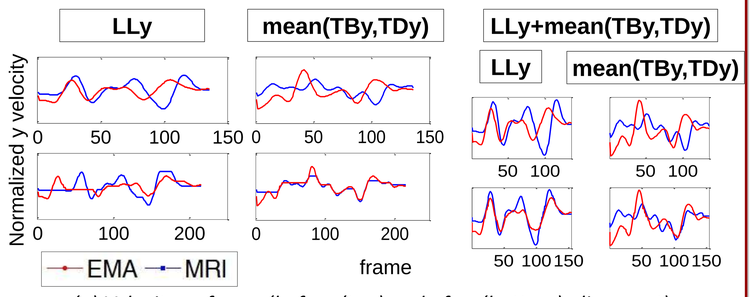
Temporal alignment of articulatory data:

- Canonical Time Warping (CTW): Canonical Correlation Analysis (CCA) + Dynamic Time Warping (DTW)

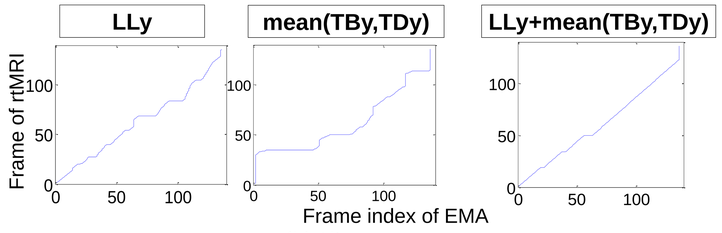


Articulatory alignment results (DTW)

- DTW on the **velocity** of mean pixel intensity for rtMRI and the **velocity** of corresponding sensor trajectories. y: vertical axis



(a) Velocity vs frame (before (top) and after (bottom) alignment)



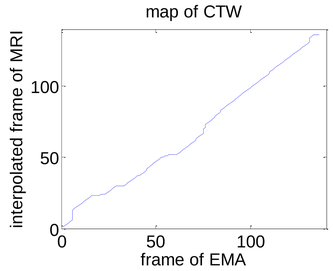
(b) Alignment map

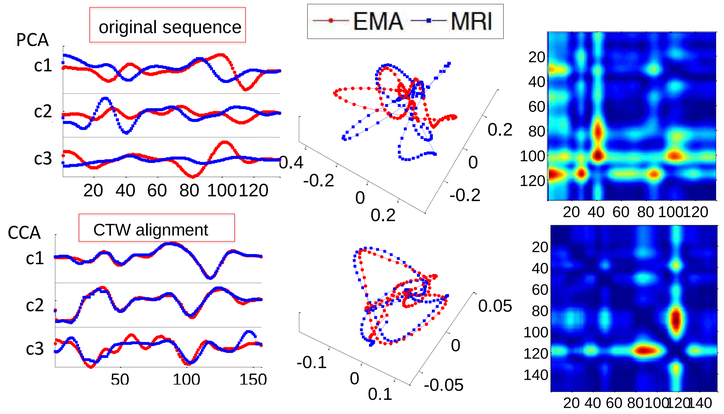
More articulatory information helps better alignments!

Articulatory alignment results (CTW)

- We need temporal alignment with more EMA sensors and more information in rtMRI.
- CTW allows the flexibility of dimensionality of two feature streams: It can align with different number (and nature) of features of two feature streams

Example of CTW results with the velocity (than position) of EMA sensor trajectories (6 features) and (manually chosen) mean pixel intensity of 4 pixel windows of rtMRI





(a) Projected features vs frame *(b) Projected features in space* *(c) cost fn. Map*

Conclusions & future work

Conclusion:

- Alignment performance:
  - DTW: useful for knowledge-based features
  - CTW performs well with mean pixel intensity features
  - CTW fails with all pixels in the vocal tract, probably because of the noise of MRI images and nonlinear relationship)
  - Articulatory features: velocity > position

Future work:

- (Noise-robust) feature developments / Noise reduction
- Automatic selection of pixel regions in rtMRI video
- Evaluation method for comparing the alignment performances of different methods and features.