Dynamic 3D Visualization of Vocal Tract Shaping during Speech

Y. Zhu[1], Y-C. Kim[1], M. I. Proctor[1], S. S. Narayanan[1], and K. S. Nayak[1]
[1]Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, United States

Introduction: Speech researchers have utilized 2D cine [1, 2], 2D real-time [3] and 3D static [4] MRI to provide valuable insights into the shaping of the vocal tract. These methods have informed our knowledge of the goals of speech production, but still require a trade-off between spatial and temporal resolution. In this work, we present a novel approach for dynamic 3D vocal tract MRI that is based on 1) 2D real-time imaging of multiple repetitions of a speech task with synchronized audio recording [5], and 2) alignment of the 2D real-time movies using dynamic time warping based on the recorded audio tracks.

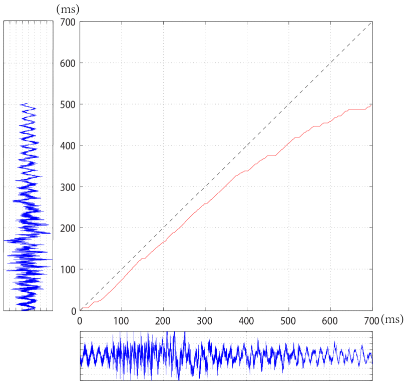


Fig 1. Minimum-distance warping path for two recordings of /*asa*/ during real-time imaging, using DTW.

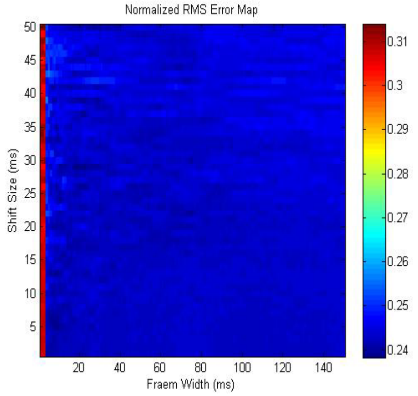


Fig 2. Normalized RMS error map of images on wide ranges of frame width and shift size.

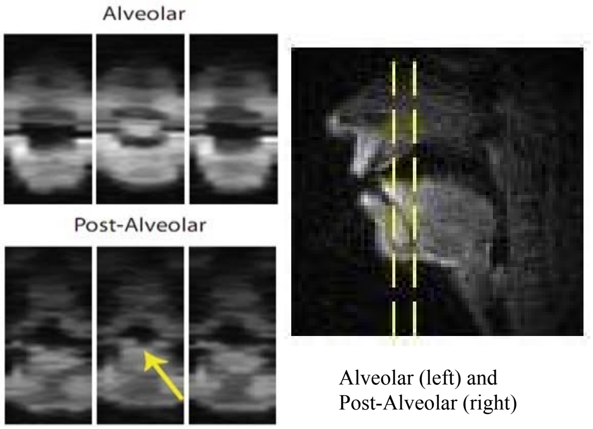
Methods and Results:

Experiments: Experiments were performed on a clinical 1.5T MRI scanner, using a custom 4-channel upper airway receiver coil. Images were acquired using a custom real-time vocal tract imaging setup [6, 7], (spiral GRE, TR = 6ms, 13 interleaves, sliding window reconstruction). Three native speakers of English repeated VCV sequences /*asa*/, /*asha*/, /*ala*/ and /*ara*/ while their vocal tracts were progressively imaged using a series of 15 parallel sagittal slices (5 mm thickness, 2 mm overlap).

Alignment Methods: Synchronized audio tracks were recorded during imaging [5], and aligned by a dynamic time warping (DTW) procedure [8], using acoustic features [9]: 1) mel-frequency cepstral coefficients (MFCC) were computed from overlapping segments of audio. 2) DTW was used to align pairs of MFCC sequences and to generate a "minimum-distance warping path" (Fig. 1). 3) The warping path was then used to align the real-time 2D movies and generate arbitrary new views and 3D movies.

Validation and Parameter Selection: In order to validate the method, a single mid-sagittal slice was imaged 2 times during production of /*asa*/, /*asha*/, /*ala*/ and /*ara*/. DTW was performed using a range of frame widths and shift sizes. The resulting RMS image error between aligned warped videos is shown in Figure 2. This error was flat over a broad range of parameters; we therefore chose settings within that range that best matched the MRI acquisition parameters, frame width = 6 ms, shift size = 1 ms, for the remainder of the studies.

3D Dynamic Visualization: The vocal tract was manually segmented from each frame of aligned videos. Stacks of 2D contours were interpolated and visualized using Matlab *3-D Visualization* package. Frames taken from two synthesized coronal videos during production of /*asa*/ are shown in Figure 3, where tongue grooving is clearly revealed (see the arrow). Example frames taken from dynamic 3D tongue models during production of /*asa*/ and /*ara*/ are shown in Figure 4; characteristic tongue grooving and cupping (see the arrows) are evident in the reconstructed lingual surfaces.



Alveolar (left) and
Post-Alveolar (right)

Fig 3. Synthesized coronal images of /*asa*/ at alveolar and post-alveolar positions.

Discussion: These experiments demonstrated the validity of the use of MFCC-DTW to reconstruct dynamic 3D models of vocal tract shaping, based on 2D real-time MRI with synchronized noise-cancelled audios. The resulting 3D movies reveal many vocal tract features that cannot be resolved by planar imaging alone, and therefore present unique value to speech research. Limitations and extensions of this approach are currently being explored.

Acknowledgments: This work is supported by NIH Grant R01 DC007124-01.

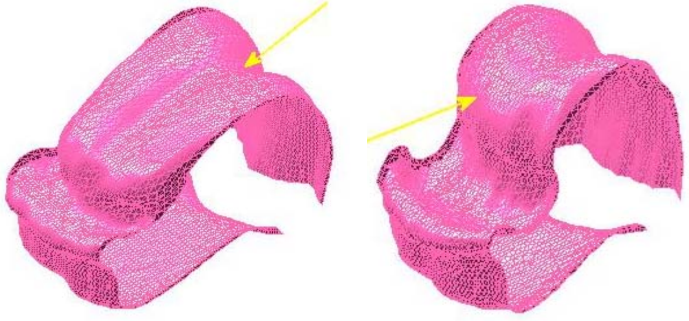


Fig 4. Individual frames from 3D dynamic movies of /*asa*/ and /*ara*/.

References: [1] Stone, et al., JSLHR:44:1026-1040, 2001; [2] Takemoto, et al., J.Acoust.Soc.Am.:119:1037-1049, 2006; [3] Byrd, et al., J.Phon.:37:97-110, 2009; [4] Story, et al., J.Acoust.Soc.Am.:100:537-554, 1996; [5] Bresch, et al., J.Acoust.Soc.Am.:120:1791-1794, 2006; [6] Santos, et al., ISMRM, p468, 2002; [7] Bresch, et al., IEEE Sig.Proc.Mag.:25:123-132, 2008; [8] Salvador, et al., KDD workshop, p70-80, 2004; [9] Zheng, et al., J.Comput.Sci.Technol.:16:582-589, 2001.