Semi-automated Tracking of Tongue Movements in Dynamic MRI of Speech

Bradley P Sutton[1], Andrew Naber[1], Jason Wang[1], Jamie L. Perry[2], and David P Kuehn[3]

[1]Bioengineering Department, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [2]Department of Communication Sciences and Disorders, East Carolina University, Greenville, NC, United States, [3]Department of Speech and Hearing Sciences, University of Illinois at Urbana-Champaign, Champaign, IL, United States

Introduction: Dynamic MRI of speech can provide useful information on the physiology of normal movements during speech. It can provide quantitative and qualitative information on the effect of pathology, disorders, or cultural differences on normal speech function. Studies correlating structural movement to speech sounds traditionally employ manual measurement techniques [1]. These manual techniques can suffer from subjective placements of tracking points and can be very time consuming given current imaging frame rates of 20 frames per second or higher [2,3]. In this work, we demonstrate a simple edge detection and processing algorithm to track points between the tongue tip, dorsum, and blade through a structured speech sample.

Methods: *Tracking Algorithm:* Robust measures on dynamic speech data are difficult due to the low contrast and high noise of fast, dynamic MRI images. We developed a semi-automated procedure that uses a user-identified region to help the algorithm converge to edges of interest. The algorithm starts by displaying the average image over the whole time series, allowing the user to identify a box containing all the places that the tongue travels to during the time series. This is the only user input required. A Canny edge detection is performed with a weak/strong threshold of 0.22/0.55 (relative to the maximum gradient intensity), $\sigma = \sqrt{2}$. The edges are fed into an edge connecting and labeling algorithm [4]. The longest edge in our region of interest is the tongue surface. The tongue tip was identified as the most anterior portion of the tongue contour and the tongue dorsum was the most superior point of the posterior region of the tongue.

Scanning: A custom spiral FLASH sequence was used to acquire MRI images at 20 frames per second [2]. The images were reconstructed using a sliding window technique to 30 frames per second. Subjects were recruited in accordance with the Institutional Review Board. Subjects were instructed to repeat "an-sa" while paced by a 2 Hz tone played through headphones.

Comparison Measures: The output of this program was compared with manual tracing performed by three trained speech scientist, referred to as H1, H2, H3 in the plots. Since we are primarily interested in the shaping of the oral track at the tongue tip and dorsum during speech, we compared the output of our semi-automated tracking algorithm (SA) to that of the three human tracers (H1, H2, H3) only for the vertical position of the tongue tip and dorsum. There was not much anterior/posterior movement at the resolution of the movies and the tracking algorithm was not as accurate in determining the position in this direction. Vertical measurements were made relative to the hard palate (roof of mouth) line.

Results: Figure 1 shows the bounding box identified by the user for the semi-automated tracking routine along with output of the algorithm. Correlation coefficients were examined between the automated tracking routine and the trained manual tracers for the measures of the vertical position of the tongue tip and tongue dorsum. The correlation coefficients were averaged across data from 4 subjects by Fisher z-transformation. The resulting average correlation coefficients are shown in Figure 2A. Figure 2B shows the normalized root mean squared error of the disagreement between all pairs of tracers (manual and algorithm), given in percent.

Conclusion: The semi-automatic tracing algorithm performed well, providing comparable results to the trained manual tracers. The amount of time to track an entire time series of images (1515 images or 50 s worth of scanning) was less than 15 minutes on a dual-core Pentium workstation. This is compared to an average of 3.5 hours for tracings by hand. The algorithm will enable correlations of speech acoustics with movements of oral structures across a range of subjects in larger studies.

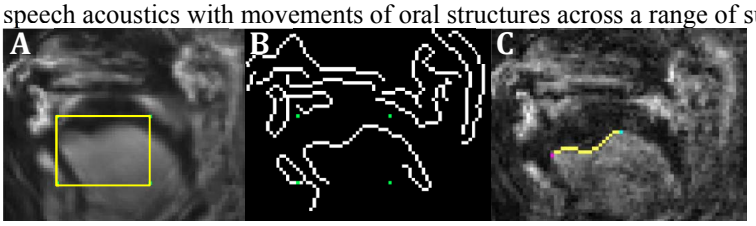

Figure 1 (above): A) User-identified box shown by 4 points. **B)** Edges from Canny filter. **C)** Outline of tongue from tongue tip to tongue dorsum.

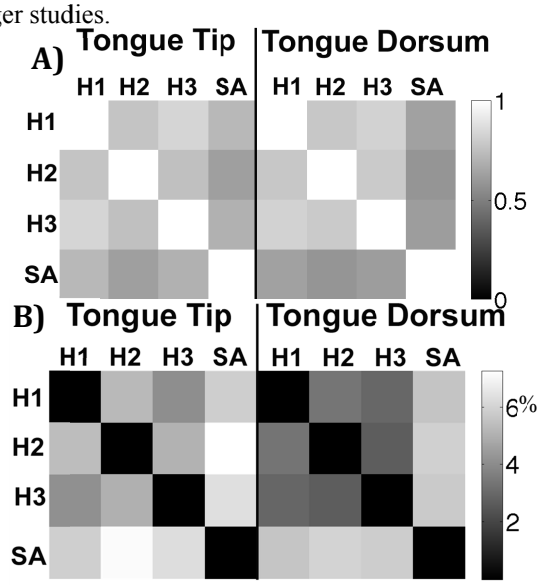

Figure 2 (right): A) Correlation coefficients (R) between tracings by three trained tracers (H1, H2, H3) and the semi-automatic program (SA). **B)** NRMSE in % between the three trained tracers (H1, H2, H3) and the semi-automatic tracing program (SA).

Acknowledgements: This work was supported by a grant from NIH 1R03DC009676-01A1.

References: [1] Bae, et al. 2011. Cleft Pal Craniofac J 48:695. [2] Sutton, et al. 2010 J Magn Reson Imag. 32:1228. [3] Uecker, et al. 2010 NMR Biomed 23:986. [4] Kovesi P. www.csse.uwa.edu.au/~pk/research/matlabfns/