Co-registration of speech production datasets from electromagnetic articulography and real-time magnetic resonance imaging

Jangwon Kim[a)]
Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089
jangwon@usc.edu

Adam C. Lammert
Department of Computer Science, University of Southern California, Los Angeles, California 90089
lammert@usc.edu

Prasanta Kumar Ghosh
Department of Electrical Engineering, Indian Institute of Science, Bangalore, Karnataka, India
prasantg@ee.iisc.ernet.in

Shrikanth S. Narayanan
Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089
shri@sipi.usc.edu

Abstract: This paper describes a spatio-temporal registration approach for speech articulation data obtained from electromagnetic articulography (EMA) and real-time Magnetic Resonance Imaging (rtMRI). This is motivated by the potential for combining the complementary advantages of both types of data. The registration method is validated on EMA and rtMRI datasets obtained at different times, but using the same stimuli. The aligned corpus offers the advantages of high temporal resolution (from EMA) and a complete mid-sagittal view (from rtMRI). The co-registration also yields optimum placement of EMA sensors as articulatory landmarks on the magnetic resonance images, thus providing richer spatio-temporal information about articulatory dynamics.

PACS numbers: 43.72.Ar [DOS]
Date Received: October 22, 2013 **Date Accepted:** December 11, 2013

1. Introduction

Speech production studies often require articulatory data. No single modality can provide direct articulatory data covering all aspects of the production process concurrently. Therefore, researchers have been using a variety of techniques, such as x-ray microbeam,[1] ultrasound,[2] electromagnetic articulography (EMA),[3,4] and magnetic resonance imaging (MRI),[5,6] to get information about specific relevant aspects of articulation. Each modality offers a different type of articulatory information each with its relative advantages and disadvantages, e.g., EMA provides flesh-sensor trajectories, while real-time magnetic resonance imaging (rtMRI) provides image sequences of the full airway cross section represented by pixel intensity. In particular, the advantages of

[a)]Author to whom correspondence should be addressed.

EMA over rtMRI are (1) higher temporal resolution: 100, 200, or 400 Hz for EMA and, typically in speech production study, 23.18–33.18 frames/s for rtMRI, and (2) potential for better speech audio recording quality: EMA allows a cleaner speech audio recording environment, while rtMRI recording generates acoustic noise in the scanner, thus requiring noise cancellation.[7] On the other hand, rtMRI can offer richer spatial information in terms of imaging in any plane including the full upper airway along the mid-sagittal plane of the head and neck, compared to tracking just a handful of sensor locations along the airway afforded by EMA.

While it may be desirable to simultaneously acquire data with multiple modalities, it is not currently feasible due to technological limitations or incompatibility such as in the case of EMA and rtMRI. One possible way to obtain some of the combined benefits of EMA and rtMRI is by spatial and temporal alignment of datasets recorded with the same stimuli, by the same speaker, but at different times. However, differences in the dimensionality and quality of the measured articulatory and acoustic data across these two modalities make the alignment problem challenging. We recently proposed a spatial and temporal alignment method[8] where the spatial alignment of EMA data and rtMRI data was computed by using the estimated palatal trace of EMA and the standard deviation image of rtMRI, and the temporal alignment was computed using the Joint Acoustic-Articulatory based Temporal Alignment (JAATA) algorithm,[8] which performs both temporal alignment and selection of EMA-like features from rtMRI data, thereby overcoming the dimensionality mismatch between the two modalities.

The goal of this paper is two-fold. The first is to experimentally validate the effectiveness of the temporal alignment method proposed by Kim *et al.* (2013) on larger datasets. The second is to illustrate the utility of the aligned corpus for speech production studies, offered by the temporally and spatially richer articulatory information in the aligned corpus than by either modality individually. We illustrate the advantage of the aligned corpora over only EMA data (Sec. 4.1) through measurement of velic and pharyngeal constriction degree, and the advantage over rtMRI data (Sec. 4.2) through anatomical landmark tracking in the up-sampled magnetic resonance (MR) images.

2. Co-registration techniques

In the co-registration of EMA and rtMRI, both spatial and temporal alignments are important to obtain a correspondence between different degrees of spatio-temporal dynamics captured by these data modalities. Spatial alignment is achieved by requiring the best alignment between the estimated palate trace of EMA data and the estimated upper vocal airway surface in MR images by means of a linear transformation composed of translations and rotation. The optimum values of translations and rotation are obtained by means of a grid search. The scaling factor is known, since the spatial resolution of MR images is specified, e.g., as $2.9\,\text{mm} \times 2.9\,\text{mm}$ per pixel, during rtMRI data collection. The temporal alignment uses acoustic features computed from the speech signal acquired in MRI and EMA recordings. To compensate for the relatively poorer audio quality in MRI recording a specialized algorithm (JAATA)[8] is used for temporal alignment which utilizes articulatory features in addition to acoustic features during alignment. While the derivatives of raw EMA sensor values are used as articulatory features for EMA data, similar flesh-point features are found from the raw MRI video and used as articulatory features in order to achieve optimum alignment. Let $\mathbf{X}_{M,f}$ and $\mathbf{X}_{E,f}$ be the acoustic feature sequence matrices of MRI audio and EMA audio, respectively, of the fth sentence among F sentences in total. Let $\mathbf{Y}_{M,f}$ be the articulatory feature sequence matrices of MR images of the fth sentence. Also, let $\mathbf{W}_{M,f}$ and $\mathbf{W}_{E,f}$ be the time alignment paths of MRI data and EMA data, respectively, for fth sentence. Then, the optimum alignment is obtained by minimizing the following objective function:

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

$\mathbf{s}_{q,M}$ is the masking matrix, whose non-zero elements select a sub-matrix (articulatory features) for qth EMA sensor trajectory among Q trajectories in total, from the MR images. A is the number of pixels in the selected articulatory features. $\mathbf{z}_{E,f}^{q}$ is the qth EMA trajectory of the fth sentence. λ is the weighting factor on the acoustic features for the temporal alignment. JAATA uses an iterative approach involving automatic extraction of EMA-like features and DTW (Dynamic Time Warping) to obtain best temporal alignment between MRI and EMA recordings. The co-registration software package which contains the MATLAB codes for the spatio-temporal alignment techniques and the subsets of data for demonstration is freely available in Ref. 9.

3. Validation of co-registration techniques

3.1 Datasets and experimental setup

Experiments were performed on the MRI TIMIT data[10] and the EMA TIMIT data, which were collected with the same stimuli and subjects, but at different times. One subject, a female native speaker of American English, was selected from the database. Data from this speaker included 460 read English sentences identical to those in MOCHA TIMIT.[11] The MRI TIMIT data comprise rtMRI image sequences of the mid-sagittal plane of the upper airway and simultaneously recorded speech waveforms that were subsequently denoised using model-based acoustic noise cancellation.[7] The EMA TIMIT data comprise three-dimensional coordinates of six flesh sensors, monitoring the movements of tongue tip, tongue blade, tongue dorsum, upper lip, lower lip, and lower incisor. Post-processing was performed on the data: Smoothing and interpolation of sensor trajectories and occlusal plane correction, on the EMA TIMIT data and principal component analysis based noise reduction on the images of the MRI TIMIT. The EMA sensor trajectories in the x and y axes are used in the experiments of this paper, because they lie in the mid-sagittal plane. Detailed specifications and post-processing procedures applied to the MRI TIMIT data and the EMA TIMIT data are described in our previous papers.[7,8,10]

Previous work[8] demonstrated the performance of JAATA with a small set of data, i.e., 20 utterance pairs (~40 s for each modality), where an utterance was spoken in each modality. In the present paper, we examine the performance of JAATA with substantially more data that includes all English phones with 114 utterance pairs, excluding initial and final silence of the utterances (~260 s for EMA TIMIT data and ~270 s for MRI TIMIT data). We evaluated temporal alignment performance by Average Phonetic-boundary Distance (APD). Since APD requires identical parallel phonetic transcriptions for each utterance pair, some utterances were excluded from the calculation of APD in the following way. First, we excluded utterance pairs if either utterance was found to contain a speaking error (i.e., deletion, addition, and substitution) through listening. Second, we excluded utterance pairs whose phonetic transcriptions were different from each other. Finally, 114 utterance pairs were chosen from the whole MRI TIMIT data and EMA TIMIT data. The phonetic transcriptions were obtained by forced-alignment with the SONIC aligner.[12]

For acoustic features, we used 13-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) with 25 msec window and 10 msec shift. For articulatory features, we used the derivatives of EMA sensor trajectories and MRI pixel intensities. APD of DTW with only MFCCs was used as a baseline.

3.2 Experimental results

We performed temporal alignment by JAATA with area parameter A from 15 to 50, except prime numbers, and weighting factor λ from 0.0001 to 0.5 with an interval of 0.0001. The minimum APD, 4.31 msec, was obtained with $A = 27$ and $\lambda = 0.0008$, which indicates that both acoustic and articulatory features were used for the best registration result. Compared to the baseline performance of 4.84 msec, JAATA reduces APD by 11%, indicating that JAATA is effective for reducing temporal alignment error.

We also examined the improvement of temporal alignment by JAATA for each phone in terms of APD. Figure 1 shows the change of APD for each phone by JAATA compared to MFCC-only alignment. Note that the phoneme notation in Fig. 1 follows the American English phoneme set used by the SONIC,[12] since APD was computed based on the automatically generated boundaries by using the tool. The phonemes are sorted in descending order of the amount of reduction in APD value. It can be seen that JAATA improves temporal alignment in terms of APD for most of the phonemes.

4. Using co-registered data

The co-registered data can offer spatially or temporally richer articulatory information than either EMA or rtMRI data by themselves. This section illustrates some ways in which co-registered data can be used for taking advantage of both EMA and rtMRI data for speech production research.

4.1 Information from more speech articulators

Articulatory information that is not directly available from EMA sensors, e.g., constrictions in the velar and pharyngeal regions, can be measured from MR images in the co-registered dataset. An example of this can be seen in Fig. 2, which shows three articulatory time series extracted during articulation of the word "harms." The velic and pharyngeal opening parameters were extracted from rtMRI data using Region-Of-Interest (ROI) analysis.[13] Labial opening was extracted from EMA data as the Euclidean distance between the upper and lower lip sensors in the mid-sagittal plane. The action of the lips is accurately captured, and the closure of the lips during production of /m/ can be clearly seen. Moreover, labial closure is coordinated in time with the velic opening to produce the nasal sound, with both time series showing a similar time course. The pronounced pharyngeal constriction is also well captured, during the production of /ɑ/ and /ɹ/ and preceding the nasal.

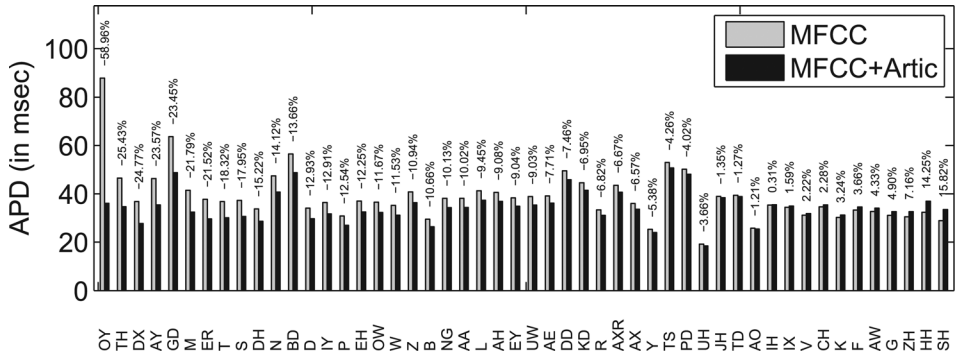

Fig. 1. APD of each phoneme for MFCC-only alignment and MFCC + articulatory (Artic) alignment. The number on top of each bar is the percentage of change from MFCC-based alignment to MFCC + Artic-based alignment. Phone list is sorted by the percentage from low to high. "−" indicates that APD decreases by adding articulatory information.

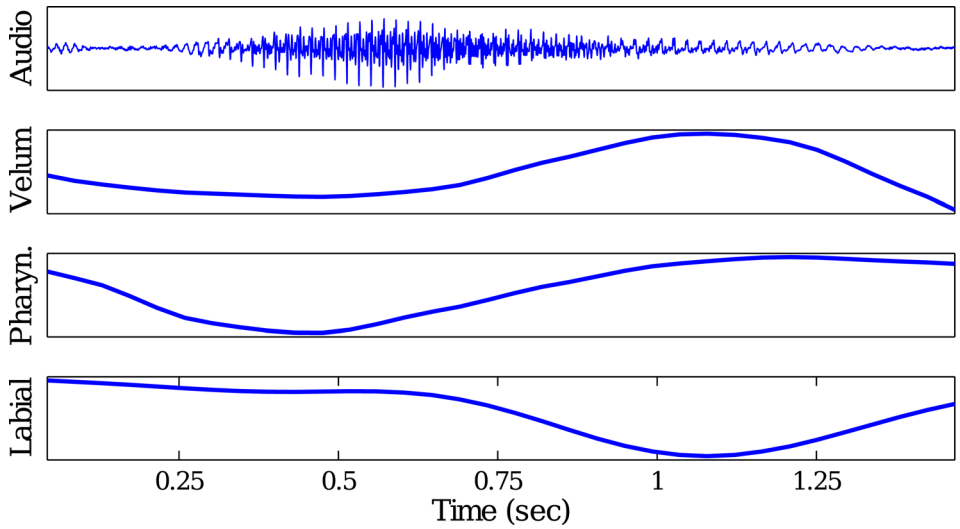

Fig. 2. (Color online) Clean speech waveform (top plot) for the word "harms" and corresponding time series of velic (the second plot), pharyngeal (the third plot), and labial (bottom plot) opening. The velic and pharyngeal opening parameters extracted from rtMRI are synchronized with the labial opening parameter extracted from the EMA by JAATA.

4.2 Higher temporal information and tongue landmarks for rtMRI data

Spatio-temporal alignment of rtMRI and EMA can be used for articulatory landmark tracking in the MR images with improved temporal resolution as a result of co-registration. Anatomical landmarks are not always conspicuous in MR images (e.g., tongue tip) because certain speech articulators, particularly the tongue, change drastically in shape over time. These shape changes can obscure or make indistinguishable anatomical landmarks, and can present challenges for landmark tracking in rtMRI. The spatio-temporal alignment can provide information about which point in each MR image corresponds to each EMA sensor that was placed at an anatomical landmark in the vocal tract. In addition, the alignment map between rtMRI and EMA can assist in up-sampling rtMRI data by utilizing the higher temporal resolution of EMA to interpolate between rtMRI frames. The left-most plot in Fig. 3 shows an example plot of an MR image overlaid with EMA sensors (circles in the plot). Sample MRI

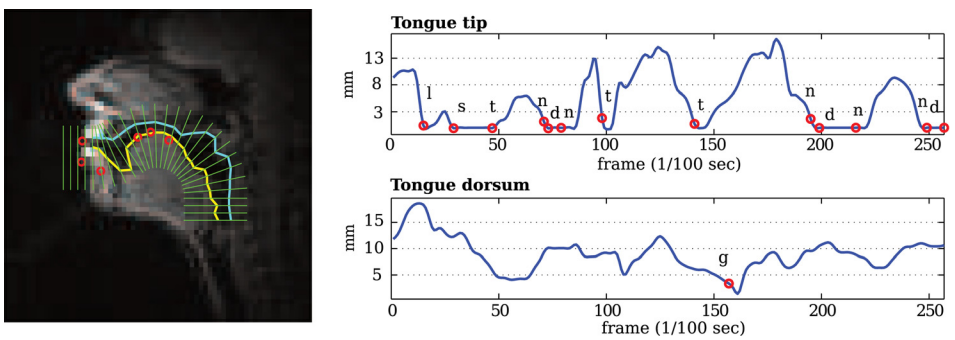

Fig. 3. (Color online) Left: Six EMA sensors (circles) overlaid on the MRI image with estimated vocal tract boundaries (outer and inner lines in the vocal tract) and grid lines after co-registration. Right: Constriction degrees of the tongue tip (top plot) and tongue dorsum (bottom plot) extracted from up-sampled rtMRI data for the sentence "Publicity and notoriety go hand in hand." The circle for each phone is placed on the trajectory of the critical articulator of the phone, indicating the frame index for the phone in the registered data.

videos with up-sampled rtMRI data, vocal tract tissue boundaries, overlaid EMA sensors before and after spatio-temporal alignment can be found at the "Demo video"' section in Ref. 9. The right-most plot in Fig. 3 illustrates the estimated constriction degrees of two landmark points (tongue tip and tongue dorsum) extracted from up-sampled MR images for the sentence "Publicity and notoriety go hand in hand." The mean of start and end times of each phone is indicated by a circle, where the phonetic boundaries were estimated by an adaptive speech-text alignment tool, SailAlign,[14] followed by manual correction. The constriction degrees from rtMRI were computed by measuring the Euclidean distance between the upper (outer line in the vocal tract in the left-most plot in Fig. 3) and the lower (inner line) air-tissue boundary points on the closest vocal tract grid line to each EMA sensor of the tongue. Note that the tongue tip sensor is usually placed about 5 mm behind the anatomical tongue tip for minimizing its interference on its natural movement. For a rough comparison, the tongue tip constriction degree was measured on the next grid line anterior to the closest grid line to the superimposed tongue tip sensor position. Air-tissue boundaries were determined using a MATLAB-based software[15] for analyzing rtMRI data. The right-most plot in Fig. 3 suggests that the estimated landmarks in the registered data capture the closure gestures of the tongue tip and the tongue dorsum well.

5. Discussion and future work

Spatially and temporally aligned EMA and rtMRI data can assist speech production research by combining the advantages of both modalities. On top of the illustrated benefits of each articulatory measurement modality, another possible advantageous combination would be to substitute the clean speech audio collected in conjunction with EMA data for the degraded rtMRI audio after temporal alignment. In addition, it may also be possible to reconstruct the tongue contour (as shown by Qin *et al.*[16]) from EMA sensors by learning the statistical relationships between the EMA sensor positions and the mid-sagittal contours visible in rtMRI. The aligned data can also be used to extract articulatory features for subsequent modeling, including for automatic speech recognition[17] and speaker verification[18] which use speech production knowledge.

Although the present paper tested our approach with more data, compared to our previous paper,[8] its robustness against additional sources of intra- and inter-speaker variability needs to be examined. For example, the effects of the variation in speaking rate and type (e.g., casual vs formal) need to be examined. Another future direction is to continue improving the proposed alignment techniques. For example, more flexible specifications (size, shape, numbers) of ROI selection might generate articulatory features leading to better alignment. Although the mean pixel intensity of some rectangular windows in rtMRI images behaves similarly to certain EMA sensors, pixel-wise tracking in rtMRI could be even more similar. Finally, our co-registration approach is potentially applicable for datasets collected by other modalities, e.g., ultrasound. Selecting a subset of EMA sensors (for alignment) depending on the corresponding available articulatory information in ultrasound data or proper feature engineering are needed so that the articulatory features from the two modalities behave similarly.

Acknowledgments

This work was supported by NSF IIS-1116076 and NIH DC007124.

References and links

[1]O. Fujimura, S. Kiritani, and H. Ishida, "Computer controlled radiography for observation of movements of articulatory and other human organs," Comp. Biol. Med. **3**(4), 371–384 (1973).
[2]M. Stone, "A guide to analyzing tongue motion from ultrasound images," Clin. Ling. Phon. **19**(6–7), 455–501 (2005).
[3]K. Iskarous, M. Pouplier, S. Marin, and J. Harrington, "The interaction between prosodic boundaries and accent in the production of sibilants," in *ISCA Proceedings of the 5th International Conference on Speech Prosody*, Chicago (2010), pp. 1–4.

[4]J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. Jackson, "Electromagnetic mid-sagittal articulometer systems for transducing speech articulatory movements," J. Acoust. Soc. Am. **92**(6), 3078–3096 (1992).
[5]S. S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," J. Acoust. Soc. Am. **115**(4), 1771–1776 (2004).
[6]T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. **106**(3), 1511–1522 (1999).
[7]E. Bresch, J. Nielsen, K. Nayak, and S. S. Narayanan, "Synchronized and noise-robust audio recordings during real-time magnetic resonance imaging scans," J. Acoust. Soc. Am. **120**, 1791–1794 (2006).
[8]J. Kim, A. Lammert, P. Kumar Ghosh, and S. S. Narayanan, "Spatial and temporal alignment of multimodal human speech production data: Real-time imaging, flesh point tracking and audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver (May 2013), pp. 3637–3641.
[9]http://sail.usc.edu/old/software/Registration_EMA_rtMRI (Last viewed September 23, 2013).
[10]S. S. Narayanan, E. Bresch, P. Kumar Ghosh, L. Goldstein, A. Katsamanis, Y.-C. Kim, A. Lammert, M. I. Proctor, V. Ramanarayanan, and Y. Zhu, "A multimodal real-time MRI articulatory corpus for speech research," in *ISCA Proceedings of Interspeech*, Florence, Italy (August 2011).
[11]A. A. Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *ISSP Proceedings of the 5th Seminar of Speech Production*, Kloster Seeon, Bavaria, Germany (2000), pp. 305–308.
[12]B. Pellom and K. Hacioglu, "SONIC: The University of Colorado Continuous Speech Recognizer," Technical Report TR-CSLR-2001-01, May 2005.
[13]A. Lammert, M. Proctor, and S. S. Narayanan, "Data-driven analysis of real-time vocal tract MRI using correlated image regions," in *Proceedings of Interspeech*, Makuhari, Japan (2010).
[14]A. Katsamanis, M. Black, P. G. Georgiou, L. Goldstein, and S. S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Philadelphia, PA (January 2011).
[15]M. Proctor, D. Bone, A. Katsamanis, and S. S. Narayanan, "Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis," in *ISCA Proceedings of Interspeech*, Makuhari, Japan (2010), pp. 1576–1579.
[16]C. Qin and M. A. Carreira-Perpinan, "Reconstructing the full tongue contour from EMA/X-ray microbeam," in *IEEE International Conference on Acoustics Speech and Signal Processing* (March 2010), pp. 4190–4193.
[17]S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," J. Acoust. Soc. Am. **121**(2), 723–742 (2007).
[18]M. Li, J. Kim, P. Ghosh, V. Ramanarayanan, and S. S. Narayanan, "Speaker verification based on fusion of acoustic and articulatory information," in *ISCA Proceedings of Interspeech* (2013).