SPATIAL AND TEMPORAL ALIGNMENT OF MULTIMODAL HUMAN SPEECH PRODUCTION DATA: REAL TIME IMAGING, FLESH POINT TRACKING AND AUDIO

Jangwon Kim, Adam Lammert, Prasanta Ghosh†*, Shrikanth S. Narayanan*

University of Southern California, Los Angeles, CA, U.S.A
†Indian Institute of Science (IISc), Bangalore, India

jangwon@usc.edu, lammert@usc.edu, †prasantg@ee.iisc.ernet.in, shri@sipi.usc.edu

ABSTRACT

In speech production research, the integration of articulatory data derived from multiple measurement modalities can provide rich description of vocal tract dynamics by overcoming the limited spatio-temporal representations offered by individual modalities. This paper presents a spatial and temporal alignment method between two promising modalities using a corpus of TIMIT sentences obtained from the same speaker: flesh point tracking from Electromagnetic Articulography (EMA) that offers high temporal resolution but sparse spatial information and real time Magnetic Resonance Imaging (MRI) that offers good spatial details but at lower temporal rates. Spatial alignment is done by using palate tracking of EMA, but distortion in MRI audio and articulatory data variability make temporal alignment challenging. This paper proposes a novel alignment technique using joint acoustic-articulatory features which combines dynamic time warping and automatic feature extraction from MRI images. Experimental results show that the temporal alignment obtained using this technique is better (12% relative) than that using acoustic feature only.

Index Terms— Speech production, spatial alignment, temporal alignment, automatic feature extraction, EMA, MRI, TIMIT corpus

1. INTRODUCTION

Speech production research crucially relies on articulatory data acquired by various acquisition methods. Each method has its advantage in terms of the nature of information it offers, while at the same time limited in important ways, notably in terms of the spatio-temporal details offered. Popular techniques include ultrasound, X-ray microbeam, Electropalatography, Electromagnetic articulography (EMA) and recently (real-time) magnetic resonance imaging (MRI). For example, EMA offers motion capture of several flesh-point sensors in two (sagittal) or three dimensional (parasagittal) coordinates with high temporal resolution (100 samples/second in WAVE system), while real-time MRI (rtMRI) provides complete midsagittal (or along any arbitrary 2D scan plane) view of the vocal tract in relatively low temporal resolution (68×68 pixel images at 23.180 samples/second [1]). Combining the information from these multimodal sources can be beneficial, but simultaneous acquisition with these techniques is usually not possible because of the cognizant technology requirements and limitations. Hence algorithmically co-registering and integrating these datasets is the most plausible avenue.

This study aims at obtaining the combined benefits of "multiple" data acquisition methods in modeling speech production dynamics by both spatial alignment and temporal alignment of these multimodal data. Specifically, it aims to obtain detailed vocal tract dynamics from MRI video aligned with EMA sensor trajectories. The alignment of multiple data will not only provide us finer and richer articulatory information, but also offer new opportunities for speech production research and modeling, i.e., temporal reconstruction (i.e., upsampling) of rtMRI based on EMA information, tongue reconstruction and complete tongue movement representation from EMA pellets, palate reconstruction from EMA pellets, and their evaluations.

We use a corpus of TIMIT sentences collected from the same speakers, but at different times, with rtMRI and EMA as the basis for this study. The speech waveform and corresponding articulatory data (recorded simultaneously) within each dataset is provided as synchronized by the acquisition system itself (EMA by WAVE) or by an algorithm in the case of rtMRI [2]. However, EMA TIMIT data and MRI TIMIT data need time warping alignment, because they were recorded separately. The temporal alignment of the two datasets is not straightforward due to several reasons. First, the nature of articulatory information of the two datasets is different: EMA is motion capture of flesh-point sensors and MRI is image stream. Second, rtMRI has grainy image noise and suffers from acoustic distortion in the speech audio signal. Lastly, the complex structure of articulators and their movements in rtMRI images make it hard to directly use spatio-temporal alignment techniques on the articulatory data.

In order to overcome the limitation of co-registering relying on any individual modality, such as using just acoustic feature based temporal alignment, we propose a novel temporal alignment using both acoustic and articulatory features, working with dynamic time warping (DTW) [3]. The goal of this work is to examine how articulatory features can be used to improve temporal alignment. For instance, spatial alignment of articulatory data can be solved by transformation based on relatively stationary "reference" structures such as using palate tracking of both EMA TIMIT and MRI TIMIT. The automatic feature extraction technique in the novel temporal alignment formulation determines the set of pixels whose mean pixel intensity behaves similar to each EMA sensor trajectory.We demonstrate the performance of this alignment method on a subset of the TIMIT corpus [1] elicited from a female speaker of American English.

This paper is organized as following. Section 2 explains the relation of our new algorithm to prior work. Section 3 describes a multimodal speech production database, the USC EMA TIMIT and MRI TIMIT corpora, along with the details of post-processing them after acquisition. Section 4 describes our spatial alignment method and results. Next, section 5 explains our temporal alignment method followed by its results in section 6. Finally, discussions, conclusions and future works follow in sections 7 and 8.

2. RELATION TO PRIOR WORK

There have been spatio-temporal alignment studies in various domains including multimedia, medical imaging [4, 5, 6]. Although these methods have shown successful alignment results on their dataset of interest, they are not directly applicable to our multimodal data. This is mainly due to the different spatio-temporal nature of the multimodal data streams. Recently, canonical time warping (CTW) [7] was introduced for alignment task, which deals with different nature of data by alternating between the linear transformation of two original data spaces to a common latent space and temporal

Thanks to NIH and NSF for funding.

alignment. However, CTW based alignment is likely to fail when the two original feature streams have complex (nonlinear) relationships such as exhibited by the EMA sensor trajectories and MRI image streams. In fact we have found poor performance of CTW based alignment on our corpus (see section 7 for details).

Accurate information about the shape of the palate can be obtained by explicit measurements of the palate (i.e., taken from a dental cast), although in practice this can be labor intensive and uncomfortable for subjects. Previous work has tried to measure palate shape from flesh-point tracking data by asking subjects to sweep the tongue tip sensor across the palate, but this can be unreliable because subjects have trouble keeping the tongue tip sensor directly against the palate and precisely in the midsagittal plane [8]. Palate shape can also be inferred from flesh-point tracking data, using all the sensor positions observed from an entire acquisition, for instance by taking the convex hull of those sensor positions [9]. In the current study, palate shape is inferred from all tongue sensor positions in the data using a windowed technique which allows for more detail about palate shape to be preserved in the inference.

3. DATA

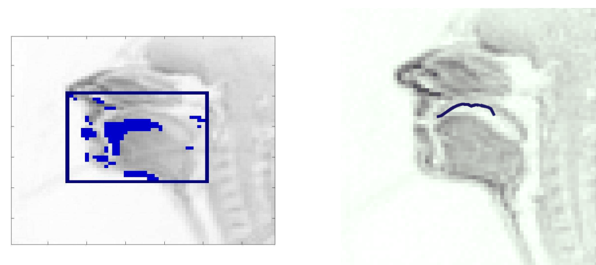
We have developed the technology for rtMRI of the vocal tract during speech with simultaneous recording of speech audio [1]. Using this we have created a speech production corpus using the same MOCHA TIMIT stimuli of 460 English sentences [10], called MRI TIMIT [1], information available in http://sail.usc.edu/span/mri-timit/. The frame rate of MRI images is 23.180 frames/sec, and spatial resolution is 68×68 pixels (2.9 mm $\times$ 2.9 mm). More details of the database, data collection and post-processing, including noise cancellation on speech audio, are explained in [1, 2]. Figure 1(a) shows a sample MRI video frame along with top 3% high variance pixels. With the same stimuli and subjects of MRI TIMIT we also collected, at a different time, flesh-point tracking EMA data using WAVE system (referred to as EMA TIMIT), which includes the trajectories of 6 flesh-point sensors on tongue tip (TT), tongue blade (TB), tongue dorsum (TD), upper lip (UL), lower lip (LL) and lower incisor (LI), at a sampling rate of 100 Hz and simultaneously recorded speech audio. Following the procedure outlined in [11], we performed post-processing which includes smoothing and occlusal plane correction on EMA sensors. The x,y co-ordinate trajectories of six EMA sensors (i.e., 12 EMA trajectories) are used for our experiments. EMA TIMIT also contains palate tracking. In palate tracking, a subject scans the upper surface of the vocal tract from the alveolar ridge to the soft palate, using the TT sensor. This palate tracking along with MRI image is used for spatial alignment. For analyzing the performance of temporal alignment, we use identical set of 20 sentences ($\sim$40 sec) from the MRI TIMIT and EMA TIMIT such that they cover all phonemes. The sentences were spoken by a native female speaker of American English.

4. SPATIAL ALIGNMENT OF EMA SENSORS ON MRI IMAGE

The goal of spatial alignment is to align the reference midsagittal plane (i.e., x-y plane) in EMA recording with MRI scan plane such that EMA sensor coordinates on the midsagittal plane correspond to the respective points on the MRI image. The spatial alignment is achieved by estimating the transformation of EMA sensors on the MRI image. We uses MRI image and palate tracking of EMA sensors for this task. The spatial alignment of articulatory sensors on MRI image can be done by applying the same transformation on the sensor coordinates. We estimate the palate contour from EMA palate tracking data as well as all tongue sensor data by choosing the highest vertical point in each adjacent bins (L/20 mm in length, no overlap), where L is the length of the palate tracking data, along x axis. To find a location for the palate trace in the MRI image plane, we firstly scaled down EMA sensors by 2.9 (Note that unit of EMA sensors is mm, and the pixel size of MRI image is 2.9 mm). Then, after manual initialization, we perform a grid search over a variety of translations, δ_x and δ_y (along x and y axis), from -5 to +5 pixels at increments of 0.5 and rotations θ from -π/4 to π/4 radians at increments of π/32 radians. The manual initialization is done at (horizontal pixel = 25th, vertical pixel = 23th, rotation = 0). The optimum translation and rotation is found to be ($\delta_x^\star$ = 25.5, $\delta_y^\star$ = 24, $\theta^\star$ = -π/32). $\delta_x^\star$, $\delta_y^\star$, and $\theta^\star$ are found by maximizing the contrast across palate trace as follows:

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

where $p_{i,j}$ is a pixel at (i, j) of standard deviation (SD) MRI matrix. The SD MRI matrix contains the standard deviations of MRI image pixels. In SD MRI matrix the palate is clearly visible as a region of high contrast just above the oral cavity and it also guards against the false palate problem unlike the raw MRI image matrix. Due to the unavailability of ground truth we visually examine the spatial alignment result. Figure 1(b) shows the optimum palate trace location of EMA on MRI image. Visually it appears that the transformation of EMA results in a good match between EMA palate trace and the palate visible in MRI image.



(a) High variance MRI pixels (b) Aligned palate trace

Fig. 1. (a) Top 3% highest variance pixels are highlighted (along with their bounding box), which includes articulatory movements in vocal tract region. (b) Spatial alignment result - dark blue line is the estimated palate trace on MRI image.

5. TEMPORAL ALIGNMENT USING ACOUSTIC AND ARTICULATORY FEATURES

Below we describe our proposed automatic algorithm for temporal alignment of MRI and EMA recordings using both acoustic and articulatory features. We refer to this automatic algorithm as Joint Acoustic-Articulatory based Temporal Alignment (JAATA). A key feature of JAATA is that it computes EMA-like features from raw MRI video in order to achieve optimum alignment.

5.1. Objective function

Suppose we need to perform temporal alignment of MRI and EMA recording of F sentences. Suppose the f-th ($1 \leq f \leq F$) sentence has N_M and N_E frames in MRI and EMA recordings, respectively. Let $\mathbf{X}_{M,f} = [\mathbf{x}_{1,M} \ \cdots \ \mathbf{x}_{N_M,M}]$ denote the acoustic feature sequence matrix of MRI audio of the f-th sentence where $\mathbf{x}_{l,M}$ is the acoustic feature vector at the l-th frame. Similarly, let $\mathbf{X}_{E,f} = [\mathbf{x}_{1,E} \ \cdots \ \mathbf{x}_{N_E,E}]$ denote the acoustic feature sequence matrix of EMA audio. We vectorize MRI video in each frame, i.e., at l-th frame MRI video matrix $V_{l,M}$ (68$\times$68) is converted to MRI video vector $\mathbf{y}_{l,M}$ ($68^2\times$1) such that $\mathbf{y}_{l,M}(68j + i) = V_{l,M}(i, j)$, $0 \leq i, j \leq 67$. Thus, for the f-th sentence, we obtain the MRI video sequence matrix $\mathbf{Y}_{M,f} = \left[\mathbf{y}_{1,M} \ \cdots \ \mathbf{y}_{N_M,M}\right]$. The 12 EMA sensor trajectory matrix is denoted by $\mathbf{Y}_{E,f} = \left[\mathbf{y}_{1,E} \ \cdots \ \mathbf{y}_{N_E,E}\right] = \left[\mathbf{z}^1_{E,f} \ \cdots \ \mathbf{z}^{12}_{E,f}\right]^{\mathrm{T}}$, where $\mathbf{y}_{l,E}$ (12$\times$1) represents the 12 EMA sensor values at the l-th frame and $\mathbf{z}^q_{E,f}$ ($N_E\times$1) is the trajectory of the q-th EMA sensor for f-th sentence. T is the matrix transpose operator. We obtain the best temporal alignment between MRI and EMA recordings of all F sentences by minimizing the following objective function:

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

The objective function J is obtained by summing objective functions J_f corresponding to each sentence. J_f has two terms which are convexly combined using weight λ - the first term measures the Euclidean distance between acoustic features of MRI and EMA audio after alignment and the second term measures the same for articulatory features. $||\mathbf{U}||_F^2 = \mathrm{Tr}(\mathbf{U}^{\mathrm{T}}\mathbf{U})$ designates the Frobenious norm. $\mathbf{W}_{M,f},\ \mathbf{W}_{E,f}$ encode the time alignment path for f-th sentence (for details see [7]). $\mathbf{s}_{q,M}$ ($68^2 \times 1$) is a masking matrix, whose non-zero elements selects a submatrix (of size $K \times L, K, L \in \mathcal{Z}$) from the MRI image matrix. Thus, $\frac{1}{A}\mathbf{s}_{q,M}^{\mathrm{T}}\mathbf{Y}_{M,f}$ is the articulatory trajectory derived from MRI video corresponding to q-th EMA trajectory. The number of pixels or the area of the submatrix is denoted by $A (= KL)$, which is user-specified before optimizing J. The elements of $\mathbf{s}_{q,M}$ can take value of 0 or 1. Thus, $\mathbf{s}_{q,M}^{\mathrm{T}}\mathbf{1} = A$, where $\mathbf{1}$ is a column vector of all '1's.

5.2. Optimization of the objective function

Minimization of J is a non-convex optimization problem with respect to the optimization variables $\mathbf{W}_{M,f},\ \mathbf{W}_{E,f}$ (time alignment matrices), $\{\mathbf{s}_{q,M}, 1 \leq q \leq 12\}$ and λ. Hence we use an iterative approach comprising two main steps - 1) Optimize $\mathbf{W}_{M,f},\ \mathbf{W}_{E,f}$ using DTW given $\{\mathbf{s}_{q,M}, 1 \leq q \leq 12\}$ and λ, 2) Given $\mathbf{W}_{M,f},\ \mathbf{W}_{E,f}\ \forall f$ and λ, optimize $\{\mathbf{s}_{q,M}\}$ sequentially $\forall q$ by searching over K, L such that $KL = A$. λ is optimized by performing a grid search. It is easy to show (from (2)) that in each of these steps J decreases monotonically. Thus the iterative process of optimization stops when the value of J reaches a local minima. The iterative process is initialized with the temporal alignment obtained by acoustic-only features using DTW and Euclidean distance between acoustic features as the distance measure.

6. TEMPORAL ALIGNMENT EXPERIMENTS

6.1. Experimental setup

We use 13 dimensional mel-frequency cepstrum coefficient (MFCC) vector as the acoustic feature $\mathbf{X}_M$ and $\mathbf{X}_E$ for both MRI TIMIT and EMA TIMIT audio. MFCCs are computed at a frame rate of 100 Hz. Note that 12 EMA trajectories are also at a frame rate of 100 Hz. We applied smoothing on the EMA trajectories by butterworth filter with a cut-off frequency at 8 Hz. 8 Hz is chosen by the frequency analysis in a previous work in [12]. We have computed the derivative of EMA trajectories and denote them as $\mathbf{Y}_E$. Similar to the EMA trajectories, we also low-pass filtered MRI video pixel trajectories using a butterworth filter with a cut-off frequency at 8 Hz. Since MRI videos have a lower frame rate, we have upsampled the MRI video at a sampling rate of 100Hz such that both acoustic and articulatory data streams are at identical frame rate. This frame rate was chosen to match the frame resolution of the phone boundary, which is used for evaluation of temporal alignment. Derivatives of the upsampled MRI pixel trajectories are computed and used as $\mathbf{Y}_M$. We normalized both EMA and MRI articulatory feature trajectories between 0 and 1 for each sentence. We have found that derivative computation and normalization contribute to better temporal alignment performance.

As discussed in Section 5, for each EMA trajectory, the optimum rectangular region on the MRI image is estimated as a by-product of the temporal alignment formulation. Trajectory of the derivative of the mean pixel intensity of MRI in the optimized area is used for temporal alignment. To reduce the search space for finding the location of the optimum rectangular area, we restrict the search to a bounding box of the top 3% high variance pixels (see Figure 1(a)) which contains the surface movement of articulators. The λ values used for optimization are $\{(k-1) \times 0.05, 1 \leq k \leq 20\}$.

For evaluation of the temporal alignment, we have used an objective measure of how the phonetic boundaries of MRI audio correspond to those of the EMA audio when mapped using the optimized alignment path. We call this measure as Average Phonetic-boundary Distance (APD). Phonetic boundaries obtained from forced alignment [13] are manually corrected to be used in this evaluation. APD is computed as the root mean square (RMS) value of the difference between the manually corrected phonetic boundaries and the estimated phonetic boundaries in EMA audio obtained by mapping phonetic boundaries of MRI audio using the temporal alignment.

6.2. Results

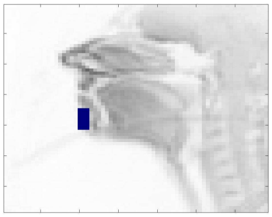
We experimented with different values of rectangular area A - 9, 12, 15, 18, 21, 24, 30, 32, 36. For all these different choices of A, the optimum value of λ turns out to be 0.1. For different choices of A APD averaged over all sentences reduces by $\sim$6 msec when articulatory features are used in addition to MFCC by JAATA. The minimum APD, 44.198 msec occurs with A=21 compared to an APD of 50.101 msec using only MFCCs. To have deeper insights, we, therefore, investigate the quality of alignment for each sentence with A=21.

We firstly examine the optimum rectangular region on MRI image for each EMA trajectory. Figure 2 shows the estimated regions of MRI image with A = 21 for four different EMA trajectories, namely LIx, LIy, TTy, TBy. From Figure 2 it is clear that the regions correspond to the respective articulators on the MRI image. The mean pixel intensity indicates the constriction degree in the region of selected pixels. Constriction degree measurement of a specific vocal tract region of rtMRI data has been used in earlier speech production studies i.e., [14, 15]. However, finding the "best" region corresponding to each EMA trajectory by hand is not straightforward. Varying morphological structure of subjects sometimes makes it hard to decide the best region. Thus our proposed optimization for temporal alignment offers a solution in this regard. To examine how correlated the mean pixel trajectory is with the corresponding EMA trajectory, we also report correlation coefficient (ρ) between the two. ρ, when averaged over all articulators, is 0.59 with a SD of 0.10. ρ values for different articulators ranges from 0.36 (ULy) to 0.68 (LIx). ρ values suggest that, on an average, the features from the mean intensity over optimum MRI regions are linearly correlated to the respective EMA trajectories.

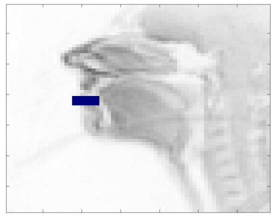
Figure 3 shows example alignment maps for four different sentences obtained using only MFCC and with both MFCC and articulatory features (MFCC+Artic) using JAATA. As a reference alignment, we have also shown an alignment based on phonetic boundaries (Reference). These four cases are chosen to illustrate the sentences where use of articulatory features led to better as well as worse alignment compared to only MFCC based alignment. For example, APD decreases by 134 msec for sentence 19 (Figure 3(b)) and by 34 msec for sentence 3 (Figure 3(b)) by using automatically extracted articulatory features in addition to MFCC. However for sentence 12, we observed that APD increases by 52 msec (Figure 3(d)).

7. DISCUSSIONS

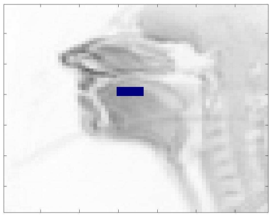
This study includes two alignment tasks, spatial alignment and temporal alignment. The performance of our temporal alignment tech-



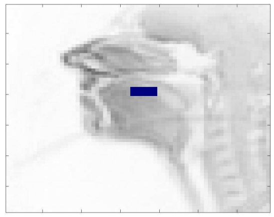
(a) for LIx (ρ=0.68)



(b) for LLy (ρ=0.67)

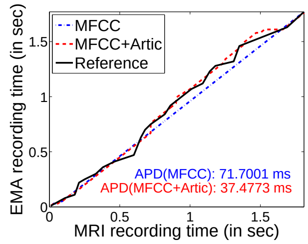


(c) for TTy (ρ=0.65)

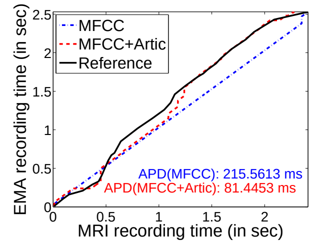


(d) for TBy (ρ=0.64)

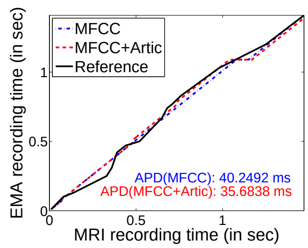
Fig. 2. Four examples of optimum MRI regions whose mean pixel intensities show highest correlation with corresponding sensor trajectories. Automatically selected pixel region is marked by a blue square box on each MRI image. 'x' or 'y' after sensor name, i.e., LI, indicates the direction of sensor movement (in the x or y axis).



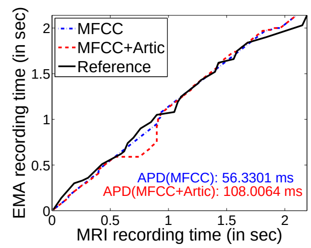
(a) Sentence 3



(b) Sentence 19



(c) Sentence 14



(d) Sentence 12

Fig. 3. Alignment maps of 4 example sentences with acoustic only (MFCC) and acoustic-articulatory features (MFCC+Artic). Reference is for manually corrected phoneme boundary (baseline). (a) and (b) are when JAATA performs better than only MFCC, (c) is when benefits from JAATA is minimal, and (d) is when JAATA performs worse than only MFCC.

nique does not rely much on spatial alignment. JAATA formulation does not use spatial alignment information directly. Even if we transform EMA sensor coordinates by spatial alignment before using them in JAATA, the temporal alignment performance may not change much. This is because the optimum spatial alignment parameter of rotation ($\theta^\star$) is small. However, the detailed information offered by spatial alignment, i.e., precise geometric relation between EMA sensor trajectories and the whole vocal tract in MRI could be beneficial for other speech production research problems.

Figure 3 shows that the temporal alignment of JAATA while promising, still has alignment error. Also, the temporal alignment of MRI and EMA recording using joint acoustic articulatory features improves APD for some sentences but decreases for others. This could be due to the temporal sparseness of articulatory information in rtMRI data. The frame resolution of rtMRI image is about 43 msec/frame, and the APD of temporal alignment using acoustic features is 50 msec. Therefore, the information gain for temporal alignment by incorporating articulatory features on top of acoustic feature might be limited. Error in manual phone boundary correction could be another possible reason for the limited performance of JAATA.

We have also investigated the benefit of using a subset of EMA sensors in temporal alignment using forward sensor selection approach. This was done by varying q (in eqn. (2)) over a subset of sensor indices instead of all 12 EMA trajectories. The APD value was used to select the best EMA sensor trajectory in each iteration of forward selection approach. The lowest value of APD (44.106 msec) was achieved with A=30 and ULx, ULy, LLx, LIy, TDy trajectories. Thus, there was no significant benefit in APD by using forward sensor selection compared to using all sensor trajectories.

Finally, we tested the spatio-temporal alignment performance using CTW [7] on our corpus. Identical to JAATA evaluation, CTW performance is also measured by APD for each sentence. Articulatory features used in CTW are the direct 12 EMA sensor trajectories and the MRI image pixels (in the blue bounding box in Figure 1(a) for feature reduction without loosing surface movements of articulators in the vocal tract). The mean ($\pm$SD) APD across all 20 sentences of CTW is 93.143 msec ($\pm$ 56.026 msec), when CTW is initialized with uniform time warping [16] (the default initialization method of CTW). For fair comparison with JAATA, we also initialized CTW by DTW with MFCC. The mean APD of DTW with MFCC is 50.101 msec ($\pm$40.659 msec). With MFCC based initialization, the APD of CTW with only articulatory data is 60.731 msec ($\pm$39.427 msec). It indicates that CTW with articulatory data does not improve temporal alignment on top of MFCC based initialization. When both MFCC and articulatory data are used in CTW, the mean APD becomes 50.229 msec ($\pm$40.617 msec). This result is worse than that of JAATA - 44.198 msec ($\pm$19.949 msec) - which uses MFCC and automatically extracted articulatory features. This performance benefit suggests that the proposed JAATA formulation results in better temporal alignment performance. Additional benefit of JAATA is that it provides "interpretable" EMA-like articulatory features from MRI video.

8. CONCLUSIONS AND FUTURE WORKS

The goal of this study is to obtain spatial and temporal alignments of multimodal speech production data, specifically MRI and EMA in order to gain the advantages of both types. For spatial alignment, we aligned the coordinates of EMA data to MRI images successfully by a grid search of estimated EMA palate tracking. For temporal alignment, we propose a novel algorithm, called JAATA, which combines DTW-based temporal alignment with optimum articulatory feature extraction from MRI video. This technique also generates the best MRI image regions from which the EMA-like articulatory features are extracted for optimum alignment. We observed the benefits of using this technique experimentally using data from MRI and EMA articulatory corpora of English TIMIT sentences spoken by the same talker. Experiment on 20 sentences' data shows that JAATA reduces mean APD value from 50.101 msec (acoustic only alignment) to 44.198 msec, which is 12% improvement. Although results are reported on 20 sentences, the alignment algorithm developed in this work can be readily applied on all the sentences from MRI TIMIT and EMA TIMIT corpora.

The temporal alignment of EMA TIMIT and MRI TIMIT still has room for improvement. For example, more flexible specifications (size, shape, numbers) of automatic pixel region selection might generate articulatory features leading to better alignment. These are part of our planned future work.

9. REFERENCES

[1] Shrikanth S. Narayanan, Erik Bresch, Prasanta Kumar Ghosh, Louis Goldstein, Athanasios Katsamanis, Yoon-Chul Kim, Adam Lammert, Michael I. Proctor, Vikram Ramanarayanan, and Yinghua Zhu, "A multimodal real-time mri articulatory corpus for speech research," in *Proceedings of Interspeech, Florence, Italy*, Aug 2011.

[2] Erik Bresch, Jon Nielsen, Krishna S. Nayak, and Shrikanth S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1791–1794, Oct 2006.

[3] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[4] Lucas Kovar and Michael Gleicher, "Automated extraction and parameterization of motions in large data sets," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 559–568, 2004.

[5] M.J. Ledesma-Carbayo, J. Kybic, M. Desco, A. Santos, M. Suhling, P. Hunziker, and M. Unser, "Spatio-temporal nonrigid registration for ultrasound cardiac motion estimation," *Medical Imaging, IEEE Transactions on*, vol. 24, no. 9, pp. 1113–1126, 2005.

[6] Stefan Kopp and Kirsten Bergmann, "Towards an architecture for aligned speech and gesture production," in *Proceedings of the 7th international conference on Intelligent Virtual Agents*, Berlin, Heidelberg, 2007, IVA, pp. 389–390, Springer-Verlag.

[7] Feng Zhou and Fernando De la Torre Frade, "Canonical time warping for alignment of human behavior," in *Advances in Neural Information Processing Systems Conference (NIPS)*, Dec 2009.

[8] J Westbury, "X-ray microbeam speech production database users handbook," *University of Wisconson at Madison*, 2005.

[9] M. Tiede, "Multi-channel visualization application for displaying dynamic sensor movements," *In development*, 2010.

[10] Alan A. Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *In Proceedings 5 th Seminar of Speech Production*, 2000, pp. 305–308.

[11] Jangwon Kim, Sungbok Lee, and Shrikanth Narayanan, "An exploratory study of the relations between perceived emotion strength and articulatory kinematics," in *INTERSPEECH*. 2011, pp. 2961–2964, ISCA.

[12] Prasanta Kumar Ghosh and Shrikanth S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2162–2172, 2010.

[13] Athanasios Katsamanis, Matthew Black, Panayiotis G. Georgiou, Louis Goldstein, and Shrikanth S. Narayanan, "Sailalign: Robust long speech-text alignment," in *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Jan 2011.

[14] Christina Hagedorn, Michael I. Proctor, and Louis Goldstein, "Automatic analysis of singleton and geminate consonant articulation using real-time magnetic resonance imaging," in *INTERSPEECH*. 2011, pp. 409–412, ISCA.

[15] Adam C. Lammert, Michael I. Proctor, Athanasios Katsamanis, and Shrikanth S. Narayanan, "Morphological variation in the adult vocal tract: A modeling study of its potential acoustic impact," in *INTERSPEECH*. 2011, pp. 2813–2816, ISCA.

[16] Ada Wai-Chee Fu, Eamonn Keogh, Leo Yung Lau, Chotirat Ann Ratanamahatana, and Raymond Chi-Wing Wong, "Scaling and time warping in time series querying," *The VLDB Journal*, vol. 17, no. 4, pp. 899–921, Jul 2008.