Spatio-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 mid sagittal 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 MOC HA speech with simultaneous recording of speech audio [1]. Using this frame rate of MRI images is 23.180 frames/sec, and spatial resolution [1], information available in http://sail.usc.edu/span/mri-timit/. The to be preserved in the inference.

\[
\begin{align*}
\{\delta_x^*, \delta_y^*, \theta^*\} &= \arg\max_{\delta_x, \delta_y, \theta} \sum_{p_{i,j} \in \text{palate trace}} \frac{p_{i,j-1}}{p_{i,j+1}} \tag{1}
\end{align*}
\]

where \(p_{i,j}\) is a pixel at \((i, j)\). The optimum translation and rotation is found to be \((\delta_x^* = 25.5, \delta_y^* = 24, \theta^* = -\pi/32\). \(\delta_x^*, \delta_y^*, \text{ and} \theta^*\) are found by maximizing the contrast across palate trace as follows:

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.

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 sentence \((1 \leq f \leq F)\) sentence has \(N_M\) and \(N_E\) frames in MRI and EMA recordings, respectively. Let \(X_{M,f} = [x_{1,M} \cdot \cdot \cdot x_{N_M,M}]\) denote the acoustic feature sequence matrix of MRI audio of the \(f\)-th sentence where \(x_{i,M}\) is the acoustic feature vector at the \(i\)-th frame. Similarly, let \(X_{E,f} = [x_{1,E} \cdot \cdot \cdot x_{N_E,E}]\) denote the acoustic feature sequence matrix of EMA audio. We vectorize MRI video in each frame, i.e., at \(i\)-th frame MRI video matrix \(V_{i,M} (68^2 \times 68)\) is converted to MRI video vector \(y_{i,M} (68^2 \times 1)\) such that \(y_{i,M}(68j + i) = V_{i,M}(i,j), 0 \leq i,j \leq 67\). Thus, for the \(f\)-th sentence, we obtain the MRI video feature matrix \(Y_{M,f} = [y_{1,M} \cdot \cdot \cdot y_{N_M,M}]\). The 12 EMA sensor trajectory matrix is denoted by \(Y_{E,f} = [y_{1,E} \cdot \cdot \cdot y_{N_E,E}] = [z_{E,f}^1 \cdot \cdot \cdot z_{E,f}^{12}]^T\), where \(y_{i,E} (12 \times 1)\) represents the 12 EMA sensor values at the \(i\)-th frame and \(z_{E,f}^q (N_E \times 1)\) 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:

$$J(\lambda, \{W_{M,f}, W_{E,f}\}, \{N_{q,M}, 1 \leq q \leq 12\}) = \sum_{f=1}^{F} J_f(\lambda, W_{M,f}, W_{E,f}, \{N_{q,M}, 1 \leq q \leq 12\})$$

$$= \sum_{f=1}^{F} \left\{ \lambda \left( \left\| X_{M,f}W_{M,f} - X_{E,f}W_{E,f} \right\|_F^2 \right) + (1 - \lambda) \left( \sum_{k=1}^{12} \left\| X_{M,f}A_{N_{q,M}}W_{M,f} - (z'_{E,f})^T W_{E,f} \right\|_F^2 \right) \right\}$$

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 $\lambda$ - 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. $||U||_F^2 = \text{Tr}(U^T U)$ designates the Frobenious norm. $W_{M,f}, W_{E,f}$ encode the time alignment path for $f$-th sentence (for details see [7]). $N_{q,M}$ is a masking matrix, whose non-zero elements selects a submatrix (of size $K \times L, K, L \in \mathbb{Z}$) from the MRI image matrix. Thus, $A_{N_{q,M}}Y_{M,f}$ 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 $N_{q,M}$ can take value of 0 or 1. Thus, $N_{q,M}^T 1 = A$, where $1$ 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 $W_{M,f}, W_{E,f}$ (time alignment matrices), $\{N_{q,M}, 1 \leq q \leq 12\}$ and $\lambda$. Hence we use an iterative approach comprising two main steps - 1) Optimize $W_{M,f}, W_{E,f}$ using DTW given $\{N_{q,M}, 1 \leq q \leq 12\}$ and $\lambda$. 2) Given $W_{M,f}, W_{E,f}$ $\forall f$ and $\lambda$, optimize $\{N_{q,M}\}$ sequentially $\forall q$ by searching over $K, L$ such that $KL = A$. $\lambda$ 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 $X_M$ and $X_E$ 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 $Y_{M}$. 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 $Y_{M}$. 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 $\lambda$ 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 $\lambda$ turns out to be 0.1. For different choices of $A$ APD averaged over all sentences reduces by $\sim 5$ 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 audio for $A = 21$ for four different EMA trajectories, namely Lix, Lly, 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 ($\rho$) between the two. $\rho$'s when averaged over all articulators, is 0.59 with a SD of 0.10. $\rho$ values for different articulators ranges from 0.36 (ULy) to 0.68 (Lix). $\rho$ values suggest that, on 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-
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).

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.

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


