Estimation of the air-tissue boundaries of the vocal tract in the mid-sagittal plane from electromagnetic articulograph data

Satyabrata Parida[1], *Pattem Ashok Kumar* [2], *Prasanta Kumar Ghosh*[2]

[1]Electrical Engineering, Indian Institute of Technology (IIT), Kharagpur-721302, India
[2]Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India

satyabrataparida@iitkgp.ac.in, ashokkumar.pattem@gmail.com, prasantg@ee.iisc.ernet.in

10.21437/Interspeech.2015-484

Abstract

Electromagnetic articulograph (EMA) provides movement data of sensors attached to a few flesh points on different speech articulators including lips, jaw, and tongue while a subject speaks. In this work, we quantify the amount of information these flesh points provide about the vocal tract (VT) shape in the mid-sagittal plane. VT shape is described by the air-tissue boundaries, which are obtained manually from the recordings by real-time magnetic resonance imaging (rtMRI) of a set of utterances spoken by a subject, from whom the EMA recordings of the same set of utterances are also available. We propose a two-stage approach for reconstructing the VT shape from the EMA data. The first stage involves a co-registration of the EMA data with the VT shape from the rtMRI frames. The second stage involves the estimation of the air-tissue boundaries from the co-registered EMA points. Co-registration is done by a spatio-temporal alignment of the VT shapes from the rtMRI frames and EMA sensor data, while radial basis function (RBF) network is used for estimating the air-tissue boundaries (ATBs). Experiments with the EMA and rtMRI recordings of five sentences spoken by one male and one female speakers show that the VT shape in the mid-sagittal plane can be recovered from the EMA flesh points with an average reconstruction error of 2.55 mm and 2.75 mm respectively.

Index Terms: Electromagnetic articulography, real time magnetic resonance imaging, RBF network, dynamic programming

1. Introduction

The study of time-varying vocal tract (VT) shapes during speaking is of great interest in speech production research with a wide range of applications including different accents learning [1], better understanding of speech production mechanism [2, 3, 4], and emotion analysis [5]. In the recent past, several studies have been carried out to understand the dynamics of VT shapes using different modalities such as X-Ray [6], Ultrasound [7] , Electropalatography [8], Electromagnetic Articulography (EMA) [9] and real-time magnetic resonance imaging (rtMRI) [10]. These modalities provide information about the time-varying VT shapes to different degrees. For example, rtMRI captures the complete view of the articulatory dynamics in the mid-sagittal plane imaged at a low temporal resolution [11]. The air-tissue boundaries (ATBs) in the rtMRI images provide rich information about the VT shapes while the subject speaks. On the other hand, EMA offers high temporal resolution [12], although it provides the movement data of only a few articulatory flesh points unlike the complete mid-sagittal view of rtMRI. In this work, we combine the complementary advantages of rtMRI and EMA data to reconstruct the time-varying VT shapes from the EMA data to quantify the information that the EMA sensors provide about ATBs. Several related works in the literature have been restricted to the reconstruction of tongue shapes from the EMA flesh points validated through alternative modalities such as ultrasound [13, 14, 15]. For tongue interpolation, linear [16, 17] and cubic [18] spline interpolation as well as data-driven reconstruction methods [13, 19, 15, 20] with varied number of flesh points have been reported. However, the accuracy of reconstructing the entire VT shape (ATBs) from the EMA flesh points remains to be investigated. Reconstructing the VT shape from the EMA data is challenging mainly due to the sparsity of the sensor points in the EMA recording.

In order to exploit the complementary nature of the rtMRI and EMA data, a simultaneous recording of both the modalities is desirable. However, a simultaneous recording of the articulatory movement using rtMRI and EMA remains infeasible due to technological limitations. So, we use the databases where EMA and rtMRI were recorded separately while multiple subjects are speaking identical sentences during both the acquisitions. The VT shape is represented by the ATBs in the rtMRI frames. Thus the proposed reconstruction of the VT shapes from EMA flesh points has two main steps – 1) co-registration of the VT shapes from rtMRI with the flesh points of the EMA data, 2) interpolation of the registered EMA points to the complete VT shape.

Co-registration of the VT shapes from the rtMRI and the EMA flesh points involves a spatio-temporal alignment. Spatio-temporal alignment algorithms have been used in various domains including multimedia [21] and medical imaging [22]. However, such alignment can not be readily applied to our multimodal data because of the difference in the spatio-temporal nature of rtMRI and EMA data. JAATA proposed by Kim et al [23], is an algorithm which addresses the spatio-temporal alignment of the features from the rtMRI frames and EMA flesh point movement. However, JAATA does not align rtMRI and EMA data for the best reconstruction of the vocal tract shape from sparse flesh points. JAATA performs spatial alignment using the upper palate trace in EMA and upper vocal airway surface in rtMRI. However, this transformation will be highly sensitive to the coordinates of the flesh points in the longitudinal direction since the upper palate has a smaller variation of position in the longitudinal coordinate axis as compared to the orthogonal mid-sagittal coordinate axis. The temporal alignment in JAATA can not be directly applied to the present work since the EMA flesh points have to be temporally aligned with optimal points on the ATBs.

In this paper, we have proposed a novel spatio-temporal alignment between the EMA flesh points and points of the ATBs in the rtMRI frames. The proposed method is based on the assumption that the optimal points on the rtMRI frame boundaries are affine transformation of the EMA flesh points in the mid-sagittal plane. This assumption holds good because both the boundaries in the rtMRI frame and the EMA sensor locations

would correspond to the morphology of a subject, recorded using each of these modalities. The spatial alignment determines the affine transformation parameters as well as the optimal points on the ATBs in the rtMRI frames. For temporal alignment, dynamic programming is used to correlate EMA frames optimally with the rtMRI frames using minimum mean squared error criterion. The affine transformed EMA flesh points are used to estimate the ATBs using Radial Basis Function (RBF) network. Experiments with the multi-modal data of one male and one female subject from USC-TIMIT [12] corpus reveal that the ATBs in the mid-sagittal plane can be reconstructed from the EMA flesh points to an accuracy of 2.55 mm and 2.75 mm respectively.

2. Dataset

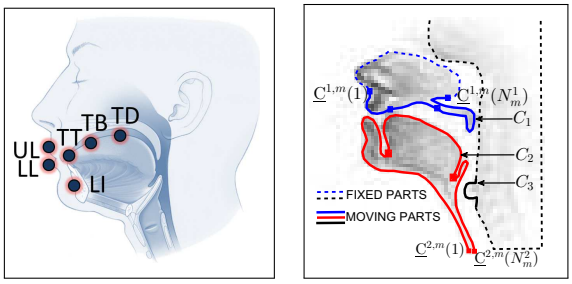
USC TIMIT [12] is a collection of MRI-TIMIT and EMA TIMIT. MRI TIMIT provides the recordings of rtMRI image sequences of the mid-sagittal upper airway imaged at 23.18 frame/sec having a spatial resolution of 68 $\times$ 68 pixels (2.9 mm $\times$ 2.9 mm) with synchronized audio at a sampling frequency of 20 kHz. EMA TIMIT provides the 3-D movement data of six flesh points namely tongue tip (TT), tongue blade (TB), tongue dorsum (TD), upper lip (UL), lower lip (LL), and lower incisor (LI) (as shown in Figure 1(a)) at a rate of 100 Hz with synchronous audio at 16 kHz. Both the MRI TIMIT and EMA TIMIT correspond to the same 460 sentences from MOCHA TIMIT corpus [24] recorded from the same subjects but at different times. A hidden Markov model (HMM) toolkit (HTK) [25] based time aligned phonetic transcription provides the phonetic boundaries in both recordings. The experiments in this paper are performed on the recordings of five sentences each from a Male subject (M1) and a Female subject (F1) of native American English from USC TIMIT. These correspond to 540 and 474 rtMRI frames for M1 and F1 respectively. The corresponding number of samples in the EMA data is 1821 and 1774 respectively. We use the Entropy of phonemes as a measure to rank sentences in the entire corpus based on their phonetic richness. The selected sentences in this work are the top five sentences from the ranked list.

The ATBs in the rtMRI frames are manually traced using a MATLAB based GUI. A sample rtMRI frame with manually drawn ATBs are shown in Figure 1(b). Since the ATBs in the mid-sagittal plane are reconstructed from EMA data, the 2-dimensional (X and Y) coordinates of the EMA sensor locations in the mid-sagittal plane are considered.

3. Estimation of ATBs from EMA points

Reconstruction of the ATBs from the EMA sensor data consists of two steps. At first the EMA sensor locations are transformed to the coordinates of the rtMRI frame by an affine transformation. Then the ATBs are estimated from these transformed EMA points using a RBF network. The affine transformation and the RBF network parameters are calculated using the manually drawn ATBs on the frames of the rtMRI recordings and the sensor data from the EMA recordings of an identical set of sentences spoken by the same subject.

The ATBs in an rtMRI frame are divided into three major segments as shown in Figure 1(b) – 1) $\mathcal{C}1$: nose, UL, hard and soft palate, velum, 2) $\mathcal{C}2$: jaw, LL, LI, tongue, epiglottis and 3) $\mathcal{C}3$: pharyngeal wall, glottis. Part of $\mathcal{C}1$ and $\mathcal{C}2$ (indicated by thick blue and red contours in Figure 1(b)) vary across rtMRI frames as the subject speaks; however, except glottis (thick black contour in Figure 1(b)), $\mathcal{C}3$ typically remains fixed. Parts of $\mathcal{C}1$ and $\mathcal{C}3$ together represents the upper ATB of the vocal tract tube. Similarly, parts of $\mathcal{C}2$ construct the lower ATB. Since EMA sensors are placed on UL, LL, LI and tongue (as shown in Figure 1(a)), it does not provide any information about $\mathcal{C}3$. Hence, in this work, we reconstruct the moving parts of $\mathcal{C}1$ and $\mathcal{C}2$ from the EMA sensor data. Both the co-registration and RBF network based reconstruction are described in the following sub-sections.



(a) EMA flesh points (b) ATBs in rtMRI

Figure 1: Schematic diagrams of EMA sensor locations and manually drawn ATBs in a rtMRI frame

3.1. Co-registration of the EMA data and ATBs in the rtMRI

The recordings of rtMRI and EMA are done separately although the same subject speaks an identical set of sentences in both recording setups. Thus, the duration of each sentence recorded in both the setups is not identical. Consequently, the articulatory movements in the rtMRI and EMA recordings are not synchronized even for the same sentence. On the other hand, the spatial resolution of the rtMRI is different from that of the EMA recording. The EMA sensors also may not be placed exactly in the plane corresponding to the rtMRI scan. Hence, the rtMRI and EMA recordings will be neither temporally nor spatially aligned. In order to align the EMA sensor data to the manually drawn ATBs in the rtMRI frames, we perform a joint spatio-temporal alignment between the two.

Suppose the moving parts of $\mathcal{C}1$ of the m-th ($1 \leq m \leq M$) rtMRI frame in the training data is defined as $\underline{C}^{1,m} = \{[C_x^{1,m}(n) \;\; C_y^{1,m}(n)]^T,\ 1 \leq n \leq N_m^1\}$ (as shown in Figure 1(b)), where N_m^1 is the number of points on the manually drawn $\mathcal{C}1$ in the m-th frame and $\underline{C}^{1,m}(n) = [C_x^{1,m}(n) \;\; C_y^{1,m}(n)]^T$ denotes the X,Y coordinates (in mm) of the n-th point in the m-th rtMRI frame. Similarly $\mathcal{C}2$ is defined as $\underline{C}^{2,m} = \{[C_x^{2,m}(n) \;\; C_y^{2,m}(n)]^T,\ 1 \leq n \leq N_m^2\}$ (as shown in Figure 1(b)). M denotes the total number of rtMRI frames corresponding to a set of sentences in the training data. Let $\underline{\xi}^S(r) = \{[x^S(r) \;\; y^S(r)]^T,\ 1 \leq r \leq R\}$ denotes the X,Y coordinates (in mm) of an EMA sensor $S \in \Gamma = \{\text{UL,LL,LI,TT,TB,TD}\}$ at the r-th sample where R is the total number of samples of EMA data when the same set of training sentences are recorded using EMA. Frame rate of EMA is higher than that of rtMRI and, hence, $R > M$.

EMA sensors are placed on the articulators in the mid-sagittal plane. Hence, we assume that six points on the ATBs (one on $\mathcal{C}1$ and five on $\mathcal{C}2$) correspond to the EMA sensor location except for an affine transformation which can be obtained by a spatial alignment. However, as $R > M$, a temporal alignment between ($\underline{C}^{1,m}$, $\underline{C}^{2,m}$, $1 \leq m \leq M$) and $\underline{\xi}^S(r)$, $1 \leq r \leq R\}$ is required to select a subset of R EMA samples which, when affinely transformed, best matches (in Euclidean sense) with the six selected points on the ATBs on the M rtMRI frames. We propose a joint spatio-temporal alignment for this purpose. Let the affine transformation parameters be denoted by A and $\underline{b}$, i.e., with affine

transformation we obtain $\hat{\underline{\xi}}^S(r) = A\underline{\xi}^S(r) + \underline{b}$. The goal of the spatio-temporal alignment is to obtain M samples from R EMA samples and six points on $\mathcal{C}1$ and $\mathcal{C}2$ such that the Euclidean distance between the selected six points and the affinely transformed sensor locations from the selected EMA samples is minimum as follows:

$$J\left(\left\{r_m^\star,\ 1 \le m \le M\right\}, \left\{n_S^\star,\ \forall S \in \Gamma\right\}, A^\star, \underline{b}^\star\right) = \underset{\{r_m\},\{n_S\},A,\underline{b}}{\arg\min}$$
$$\sum_{m=1}^{M}\left\{\left\|\underline{C}^{1,m}(n_{UL}) - \hat{\underline{\xi}}^{UL}(r_m)\right\|_2^2 + \sum_{S\in\Gamma\setminus\{UL\}}\left\|\underline{C}^{2,m}(n_S) - \hat{\underline{\xi}}^{S}(r_m)\right\|_2^2\right\}$$

$||\cdot||_2^2$ is the L2 norm of a vector. $\{r_m,\ 1 \le m \le M\}$ are the M sample indices, where r_m-th EMA sample corresponds to the m-th rtMRI frame. Since the rtMRI frames are temporally ordered, it is required to have $r_1 < r_2 < \cdots < r_M$. n_S is the index of a point on an ATB corresponding to the EMA sensor S. n_{UL} is the index of a point on $\mathcal{C}1$ since UL is on the upper ATB. The remaining indices $n_S,\ S \in \Gamma \setminus \{UL\}$ are for the points on $\mathcal{C}2$. Since LL, LI, TT, TB, TD are spatially ordered on $\mathcal{C}2$, it is required to have $n_{LL} < n_{LI} < n_{TT} < n_{TB} < n_{TD}$. Note that the objective function (eqn (1)) is not a convex function of the optimization parameters except for A and $\underline{b}$. Thus, in general, no closed form solution of eqn (1) is possible. We propose to minimize the objective function in an iterative fashion.

In every iteration (i), we use linear regression to solve $A^i, \underline{b}^i$ using $\{r_m^{i-1}\}$ and $\{n_S^{i-1}\}$ from the $(i$-1)-th iteration. Similarly, $\{r_m^{i}\}$ and $\{n_S^{i}\}$ are solved using $A^i, \underline{b}^i$. This is done by formulating the minimization problem using dynamic programming. This ensures that the objective function value decreases in every iteration as follows:

$$\begin{aligned}\cdots \ge J\left(\left\{r_m^{i-1},\ 1 \le m \le M\right\}, \left\{n_S^{i-1},\ \forall S \in \Gamma\right\}, A^{i-1}, \underline{b}^{i-1}\right)\\ \ge J\left(\left\{r_m^{i-1},\ 1 \le m \le M\right\}, \left\{n_S^{i-1},\ \forall S \in \Gamma\right\}, A^{i}, \underline{b}^{i}\right)\\ J\left(\left\{r_m^{i},\ 1 \le m \le M\right\}, \left\{n_S^{i},\ \forall S \in \Gamma\right\}, A^{i}, \underline{b}^{i}\right) \ge \cdots \quad (2)\end{aligned}$$

In order to solve $\{r_m^i\}$ and $\{n_S^i\}$ using dynamic programming, we first compute the index $(n_S^{r,m})$ of the point on the ATB of the m-th rtMRI frame which is closest to the affinely transformed EMA sensor location of the r-th sample as follows:

$$n_{UL}^{r,m} = \underset{n}{\arg\min}\left\|\underline{C}^{1,m}(n) - \hat{\underline{\xi}}^{UL}(r)\right\|_2 \quad (3)$$

$$\{n_S^{r,m},\ S \in \Gamma \setminus \{UL\}\} = \underset{\{n_S\}}{\arg\min}\sum_{S\in\Gamma\setminus\{UL\}}\left\|\underline{C}^{2,m}(n_S) - \hat{\underline{\xi}}^{S}(r)\right\|_2 \quad (4)$$

where $n_{LL} < n_{LI} < n_{TT} < n_{TB} < n_{TD}$, which is solve by using a dynamic programming formulation. This, in turn, provides the cost of aligning r-th EMA sample with the m-th rtMRI frame's ATBs given by $\zeta^{r,m} = \left\|\underline{C}^{1,m}(n_{UL}^{r,m}) - \hat{\underline{\xi}}^{UL}(r)\right\|_2^2 + \sum_{S\in\Gamma\setminus\{UL\}}\left\|\underline{C}^{2,m}(n_S^{r,m}) - \hat{\underline{\xi}}^{S}(r)\right\|_2^2$. Then $\{r_m^\star,\ 1 \le m \le M\}$ can be solved by a dynamic programming technique, where the cumulative cost $\bar{\zeta}^{r,m}$ of aligning r-the EMA sample with the m-th rtMRI frame can be computed recursively as follows:

$$\bar{\zeta}^{r,m} = \min_{1\le\rho<r}\left\{\bar{\zeta}^{\rho,m} + \zeta^{r,m}\right\}$$
$$\pi^{r,m} = \underset{1\le\rho<r}{\arg\min}\left\{\bar{\zeta}^{\rho,m} + \zeta^{r,m}\right\},\ \ \forall m, r \quad (5)$$

where $\pi\zeta^{r,m}$ is the backtracking pointer which can be used to trace back the best path resulting in $\{r_m^i\}$. Consequently, $\{n_S^i\} = n_S^{r_m^i,m}$. The iteration stops when the change in the objective function value $< 10^{-6}$. Then, the optimization variables are obtained as follows: $\{r_m^\star\} = \{r_m^i\}$, $\{n_S^\star\} = \{n_S^i\}$, $A^\star = A^i$, and $\underline{b}^\star = \underline{b}^i$.

3.2. Estimation of the ATBs

(1)We follow the RBF network-based predictive model used by Qin et al [15] to estimate the ATBs from the affinely transformed EMA sensor locations. We estimate the moving parts of $\mathcal{C}1$ and $\mathcal{C}2$ separately from the respective transformed EMA data. Since $\mathcal{C}1$ and $\mathcal{C}2$ are manually drawn, the number of points on these boundaries varies across rtMRI frames. For training the RBF network, we resample $\mathcal{C}1$ and $\mathcal{C}2$ in all rtMRI frames into a fixed set of N_{C1} and N_{C2} equally spaced points. Since the velum location is not available in the EMA data, we manually mark the velum in every rtMRI frame and assume the marked velum point to be the affinely transformed velum sensor location from EMA if it were available.

For the description of RBF network based reconstruction of the ATBs, let us assume there are K (K=1 for $\mathcal{C}1$ and K= 5 for $\mathcal{C}2$) transformed EMA points $\left(\mathbf{u} \in \mathbb{R}^{2K}\right)$ and corresponding P (K=N_{C1} for $\mathcal{C}1$ and K=N_{C2} for $\mathcal{C}2$) ATB points $\left(\mathbf{v} \in \mathbb{R}^{2P}\right)$. Using RBF network, we predict $\mathbf{v}$ from $\mathbf{u}$ using a mapping function $\mathbf{f}(\mathbf{u}) = \mathbf{v}$ that is estimated from the training set. $\mathbf{f}$ is represented using RBF network: $\mathbf{f}(\mathbf{u}) = \mathbf{W}\Phi(\mathbf{u})$, where the weight matrix $\mathbf{W}_{2P\times Q}$, where Q is the number of Gaussian basis functions $\phi_i(\mathbf{u}) = \exp\left(-\frac{1}{2}\left\|(\mathbf{u} - \mu_i)/\sigma\right\|^2\right)$ with center μ_i and spread σ. μ_i are learn from training data using vector quantization, and $\mathbf{W}$ is estimated by solving a linear least-squares problem.

4. Experiments and Results

4.1. Experimental setup

The ATBs are estimated from the EMA data for M1 and F1 separately in a leave-one-sentence-out cross-validation setup. Four sentences are used for training the affine transformation as well as RBF network parameters while the EMA data of the test sentence is used for evaluation. A block diagram of the experimental setup is shown in Figure 2. For the RBF prediction, we have chosen N_{C1}=300 and N_{C2}=500. Q is set to the number of rtMRI frames in the training set. σ is set to 64. For the spatio-temporal alignment, A and $\underline{b}$ is initialized as a 2×2 identity matrix and a zero vector respectively.

Since there is no ground truth ATBs corresponding to the test EMA samples, we perform a co-registration between the ATBs from rtMRI and EMA sensor locations for the test sentence. The EMA samples aligned with the test rtMRI frames are finally used to estimate the ATBs, which are evaluated against the manually drawn ATBs from the test rtMRI frames. In order to compare the original manually drawn ATBs with the estimated ones, dynamic time warping (DTW) [26] is used to spatially align both of them and the root mean squared error (RMSE) is reported on the

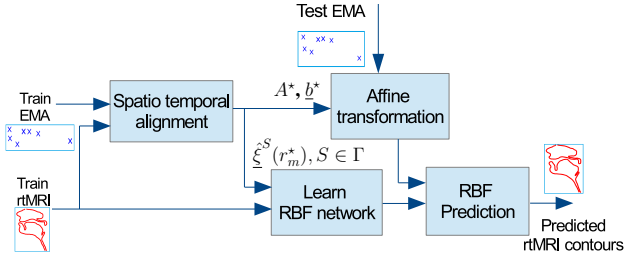


Figure 2: Experimental Setup

$\mathcal{C}1$ and $\mathcal{C}2$ separately. Note that, before DTW, the reconstructed ATBs in the m-th frame are resampled to N_m^1 and N_m^2 equally spaced points. We also report RMSE of the reconstructed ATBs in three segments for both $\mathcal{C}1$ (UL (S1), hard palate (S2) and velum (S3)) and $\mathcal{C}2$ (LL and jaw (S1), tongue (S2) and epiglottis (S3)) in order to examine the quality of reconstruction of the VT shape in articulator specific manner. The segment boundaries are illustrated by solid squares in Figure 1(b). As a baseline for comparison with RBF network, we have used spline interpolation technique.

4.2. Results and Discussions

Figure 3 illustrates the result of spatio-temporal alignment for four different phonetic frames, namely /b/, /k/, /d/ and /s/. Close match between the affinely transformed EMA points (blue cross) and selected points from manually drawn ATBs (red circle) indicates a good quality spatio-temporal alignment in different frames corresponding to various phonemes which requires different style of articulation. This also suggests that an affine transformation between the rtMRI and EMA coordinates represents their relationship well.

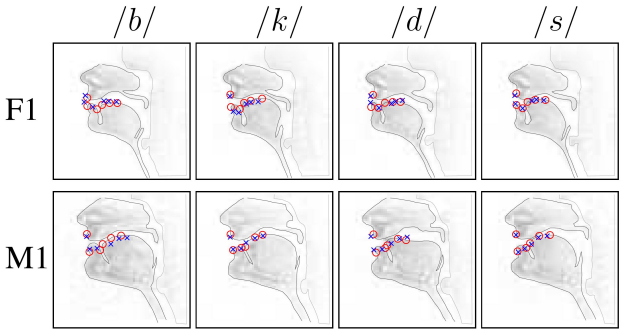


Figure 3: Illustration of the spatio-temporal alignment of the manually drawn ATBs with EMA sensors for frames corresponding to four different phonemes for each subject (F1 and M1).

Figure 4 illustrates the reconstructed ATBs (both $\mathcal{C}1$ and $\mathcal{C}2$) with the ground truth for different folds of both the subjects. Pink crosses indicate the affinely transformed EMA points, which are used for reconstructing the ATBs using RBF network. It should be noted that the reconstructed ATB does not pass through the pink crosses because the affinely transformed EMA points do not exactly match with the points on the manually drawn ATBs, which are used for training RBF network. It is clear from Figure 4 that the reconstructed ATBs closely match with the manually drawn ATBs although there are few regions where both do not match exactly. For example, in Fold5 for F1, the reconstructed velum does not match with the ground truth from rtMRI. Similarly in Fold2 for F1 and Fold3 for M1, the reconstructed epiglottis does not match with the ground truth.

Table 1: Average RMSE (in mm) in different segments (C1-S1, C1-S2, C1-S3) of the reconstructed ATBs for $\mathcal{C}1$ using RBF network for both subjects.

Subject	C1-S1	C1-S2	C1-S3
F1	2.51 ($\pm$0.64)	2.62 ($\pm$0.65)	3.1 ($\pm$0.53)
M1	3.23 ($\pm$0.57)	3.48 ($\pm$0.61)	3.38 ($\pm$0.66)

In order to obtain a quantitative measure of the quality of reconstruction, we report the RMSE values for individual ATB and their respective segments. For $\mathcal{C}1$, the RBF network based reconstruction results in average RMSE values of 2.4($\pm$0.9) mm and 2.2($\pm$0.6) mm for F1 and M1 respectively. These are significantly lower than those obtained by baseline scheme, 25.8($\pm$3.8) mm and 30.9($\pm$2.9) mm. This is true for $\mathcal{C}2$ too. The average RMSE values obtained by RBF network are 2.7($\pm$0.7) mm and 3.3($\pm$0.5) mm for F1 and M1 respectively while those for the baseline scheme are 23.7($\pm$3.7) mm and 33.9($\pm$5.2)mm. We find that, on average, the RMSE for $\mathcal{C}2$ is lower than that for $\mathcal{C}1$ for both F1 and M1. This could be because $\mathcal{C}2$ contains five sensor points as opposed to one sensor point on $\mathcal{C}1$. Table 1 and 2 provide average RMSE for three segments of $\mathcal{C}1$ and $\mathcal{C}2$ respectively for both F1 and M1 only for RBF based reconstruction. It is seen that the RMSE is similar across different segments indicating that the reconstruction quality is similar across segments.

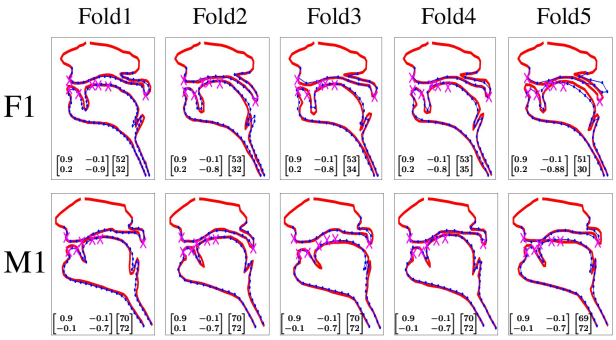


Figure 4: Comparison of the manual (red) and automatically reconstructed (blue) ATBs. One frame from each fold is shown for illustration for both subjects. $A^\star$ and $\underline{b}^\star$ for each fold are shown in the respective plots indicating a consistency in their values across folds and subjects.

Table 2: Average RMSE (in mm) in different segments (C2-S1, C2-S2, C2-S3) of the reconstructed ATBs for $\mathcal{C}2$ using RBF network for both subjects.

Subject	C2-S1	C2-S2	C2-S3
F1	2.43 ($\pm$0.61)	2.46 ($\pm$0.64)	2.43 ($\pm$0.63)
M1	3.1 ($\pm$0.57)	3.04 ($\pm$0.55)	3.02 ($\pm$0.56)

5. Conclusions

In this paper, we quantify the information provided by EMA sensor locations about the ATBs in the mid-sagittal plane. The reconstruction of the ATBs are done using a two-stage approach. The first stage involves a spatio-temporal alignment of the EMA data with the ATBs from the rtMRI frames and the second stage involves the estimation of the ATBs from the affine transformed EMA points using RBF network. We find that the average RMSE of the reconstructed ATBs varies from 2.4mm to 3.5mm for different articulatory regions for both the subjects considered for experiment in this work. The RMSE (averaged over different ATBs) are 2.55mm and 2.75mm for female and male subjects respectively. Although five phonetically rich sentences are used for the experiments in this study, manual ATBs of more sentences are required to show robustness of the proposed technique. This is part of our future work.

6. Acknowledgments

We thank Dr. C. Qin and Prof. M. Á. Carreira-Perpiñán for providing the code for RBF network based prediction. We also thank Department of Science and Technology (DST), Govt. of India for their support.

7. References

[1] A. Dowd, J. Smith, and J. Wolfe, "Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time," *Language and Speech*, vol. 41, no. 1, pp. 1–20, 1998.

[2] M. Stone, G. Stock, K. Bunin, K. Kumar, M. Epstein, C. Kambhamettu, M. Li, V. Parthasarathy, and J. Prince, "Comparison of speech production in upright and supine position," *The Journal of the Acoustical Society of America*, vol. 122, no. 1, pp. 532–541, 2007.

[3] J. Schroeter and M. M. Sondhi, "Techniques for Estimating Vocal-Tract Shapes from the Speech Signal," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 133–150, 1994.

[4] B. Denby and M. Stone, "Speech Synthesis from Real Time Ultrasound Imagery of the Tongue," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 2004*, vol. 1, pp. 685–688.

[5] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing Speech: Capturing Vocal Tract Shaping Using Real-Time Magnetic Resonance Imaging," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 123–132, 2008.

[6] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, "Generating vocal tract shapes from formant frequencies," *The Journal of the Acoustical Society of America*, vol. 64, no. 4, pp. 1027–1035, 1978.

[7] K. L. Watkin and J. M. Rubin, "Pseudo-three-dimensional reconstruction of ultrasonic images of the tongue," *The Journal of the Acoustical Society of America*, vol. 85, no. 1, pp. 496–499, 1989.

[8] M. Stone and A. Lundberg, "Three-dimensional tongue surface shapes of English consonants and vowels," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3728–3737, 1996.

[9] D. Maurer, B. Gröne, T. Landis, G. Hoch, and P. Schönle, "Re-examination of the relation between the vocal tract and the vowel sound with electromagnetic articulography (EMA) in vocalizations," *Clinical linguistics & phonetics*, vol. 7, no. 2, pp. 129–143, 1993.

[10] D. Demolin, T. Metens, and A. Soquet, "Real time MRI and articulatory coordinations in vowels," in *Proc. 5th Speech Production Seminar*, 2000, pp. 86–93.

[11] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. C. Lammert, M. I. Proctor, V. Ramanarayanan, Y. Zhu *et al.*, "A Multimodal Real-Time MRI Articulatory Corpus for Speech Research," in *INTERSPEECH*, 2011, pp. 837–840.

[12] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

[13] T. Kaburagi and M. Honda, "Determination of sagittal tongue shape from the positions of points on the tongue surface," *The Journal of the Acoustical Society of America*, vol. 96, no. 3, pp. 1356–1366, 1994.

[14] A. J. Lundberg and M. Stone, "Three-dimensional tongue surface reconstruction: Practical considerations for ultrasound data," *The Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2858–2867, 1999.

[15] C. Qin, M. A. Carreira-Perpinán, K. Richmond, A. Wrench, and S. Renals, "Predicting tongue shapes from a few landmark locations," in *INTERSPEECH*, 2008, pp. 2306–2309.

[16] J. R. Westbury, M. Hashi, and M. J. Lindstrom, "Differences among speakers in lingual articulation for American English /ɹ/," *Speech Communication*, vol. 26, no. 3, pp. 203–226, 1998.

[17] C. Qin and M. Á. Carreira-Perpiñán, "An Empirical Investigation of the Nonuniqueness in the Acoustic-to-Articulatory Mapping." in *INTERSPEECH*, 2007, pp. 74–77.

[18] A. Toutios, S. Ouni, and Y. Laprie, "Protocol for a Model-based Evaluation of a Dynamic Acoustic-to-Articulatory Inversion Method using Electromagnetic Articulography," in *The eighth International Seminar on Speech Production-ISSP'08*, 2008.

[19] P. Badin, E. Baricchi, and A. Vilain, "Determining tongue articulation: from discrete fleshpoints to continuous shadow," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[20] C. Qin and M. A. Carreira-Perpinán, "Reconstructing the full tongue contour from EMA/X-ray microbeam," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 2010*, pp. 4190–4193.

[21] M. Iino, Y. F. Day, and A. Ghafoor, "An Object-Oriented Model for Spatio-Temporal Synchronization of Multimedia Information," in *Proceedings of the International Conference on Multimedia Computing and Systems*, 1994, pp. 110–119.

[22] D. Perperidis, R. H. Mohiaddin, and D. Rueckert, "Spatio-temporal free-form registration of cardiac MR image sequences," *Medical Image Analysis*, vol. 9, no. 5, pp. 441–456, 2005.

[23] J. Kim, A. C. Lammert, P. K. Ghosh, and S. S. Narayanan, "Co-registration of speech production datasets from electromagnetic articulography and real-time magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. EL115–EL121, 2014.

[24] A. A. Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *Proceedings 5 th Seminar of Speech Production*, 2000.

[25] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The HTK book (for HTK version 3.2)," *Cambridge University Engineering Department*, 2004.

[26] E. J. Keogh and M. J. Pazzani, "Scaling up Dynamic Time Warping for Datamining Applications," in *Proceedings of the Sixth ACM International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 285–289.