A SUPERVISED AIR-TISSUE BOUNDARY SEGMENTATION TECHNIQUE IN REAL-TIME MAGNETIC RESONANCE IMAGING VIDEO USING A NOVEL MEASURE OF CONTRAST AND DYNAMIC PROGRAMMING

Advait Koparkar[1] *Prasanta Kumar Ghosh*[2]

[1]Electrical and Electronics Engineering, Birla Institute of Technology and Science, Goa-403726, India
[2]Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India

ABSTRACT

This paper introduces a technique for the supervised segmentation of Air-Tissue Boundaries (ATBs) in the upper airway of the vocal tract in the real time magnetic resonance imaging (rtMRI) videos. The proposed technique uses a novel measure of contrast across a boundary using Fisher discriminant function. ATBs in all frames of an rtMRI video are jointly estimated by maximizing the proposed measure of contrast around the predicted ATBs and incorporating a smoothness constraint to ensure the ATBs in consecutive frames do not change drastically. Dynamic programming is used for this purpose. The accuracy of the proposed technique is evaluated separately for the upper and lower ATBs using the Dynamic Time Warping distance between the predicted and the ground truth contours. Experiments with rtMRI videos from four subjects show that the error in ATB prediction using the proposed technique is 8.99% less than that using a semi-supervised grid based segmentation approach. A key feature of the proposed approach is that it can reliably predict the ATB outside the vocal tract unlike those with the existing methods.

***Index Terms*—** real-time magnetic resonance imaging, air-tissue boundary segmentation, Fisher discriminant, dynamic programming

1. INTRODUCTION

Real time magnetic resonance imaging (rtMRI) of the vocal tract in the midsagittal plane while speaking is an invaluable tool for studying human speech production. By providing images of the entire vocal tract in a non-invasive manner [1], rtMRI proves itself to be more effective than other available methods including Electromagnetic Articulography (EMA) [2], X-Ray [3] and Ultrasound [4]. The spatio-temporal information about various speech articulators obtained from rtMRI not only offers insights into speech articulation and acoustics but also sheds light on how speech production can be modelled [5]. The spatio-temporal information about the various speech articulators present in the vocal tract can be extracted by segmenting the upper airway of vocal tract in each frame of the rtMRI videos. This is done by finding the set of points which represent the boundary between the tissue and the air cavity in the vocal tract. Air-Tissue Boundaries (ATBs) in the upper airway can be described as contours which separate the regions of high pixel intensity (corresponding to the tissue) from the regions with relatively lower pixel intensity (corresponding to the airway cavity in the vocal tract). This paper proposes a technique for accurately segmenting rtMRI videos to obtain ATBs. The importance of accurately segmenting the upper airway of the vocal tract stems from the need to study the time evolution of the vocal tract cross-sectional area [6] which often forms the basis for speech processing applications. For example, Patil et al. [7] compares the articulatory control of beat-boxers using rtMRI data to gain an insight into ways in which articulators can be trained and used to achieve acoustic goals. Studies involving the analysis of vocal tract movements [8] and morphological structures of the vocal tract [9] require segmentation of rtMRI frames as a pre-processing step. Toutios [10] also uses the estimated ATBs in the rtMRI videos of the mid sagittal plane as the first step in developing a text-to-speech synthesis system. Thus, it is clear that rtMRI videos require ATB segmentation before analyses on the dynamics of the vocal tract and different articulators can be carried out [11, 12, 13, 14].

Several works in the past have addressed the problem of ATB prediction in rtMRI video frames using a number of techniques. There are several robust methods [15, 16, 17, 18] for prediction of ATB of the vocal tract by using a composite analysis grid lines superimposed on each MR image. A Region of Interest (ROI) based approach has also been proposed by Lammert et al. [19]. Asadiabadi and Erzin [20] presented a statistical approach for segmentation based on appearance and shape models for the human vocal tract. Somandepalli et al. [21] tackled the problem of boundary tracking in rtMRI frames as a pixel labelling problem and obtained contours using a greedy search of the probability maps. Lammert et al. [22] also presented a data-driven approach to the segmentation problem based on average intensities of pixels. The approach applied by Toutios [23] and Sorensen [24] used factor analysis to derive compact representations of vocal-tract outlines. Multi-directional Sobel operators were used in the tongue region to construct a boundary intensity map by Zhang et al. [25]. Although unsupervised, semi-automatic approaches such as those presented in [15, 18, 20, 22] have their advantages. However, a more accurate boundary can be obtained using a supervised technique where boundary shapes can be learned from training data rather than estimating in an unsupervised manner. This may result in a more reliable prediction of the ATBs in the upper airway of the vocal tract.

In this work, we propose a supervised approach for accurate and nuanced segmentation as well as tracking of the ATBs in rtMRI videos. The proposed approach offers several advantages over other methods: (1) It reliably predicts contours by overcoming the imaging artifact and grainy noise which could be challenging for unsupervised or naive low-level gradient based approaches, (2) It also results in a more accurate and realistic prediction of boundaries because it accounts for global contrast features in the ATB rather than local gradients which is not guaranteed in unsupervised methods, (3) It exploits the slowly varying nature of vocal tract morphology and predicts the ATB jointly across multiple video frames unlike a frame-by-frame segmentation in the existing methods.

Predicting ATB in rtMRI images can be viewed as a problem of finding the boundary corresponding to the contour of maximal con-

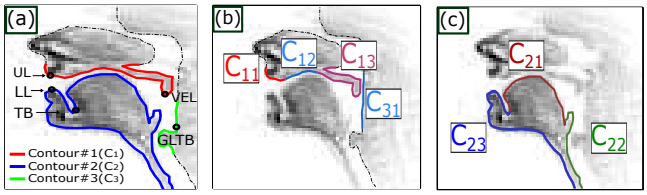

Fig. 1. (a) Illustrative air-tissue boundaries (C_1, C_2, C_3) in an rtMRI frame, (b,c) parts of the different contours used in this work

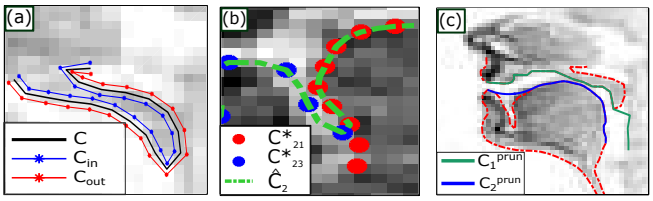

Fig. 2. (a) C_{in} and C_{out}, (b) Contour stitching, (c) Contour Pruning

trast. The proposed method uses a novel measure of contrast based on Fisher Discriminant Measure (FDM) along the contour for predicting the ATB. The proposed segmentation scheme also imposes a temporal continuity constraint using Dynamic Programming (DP) so that the predicted contours in consecutive frames do not vary erratically. We begin with the description of the dataset used in this work.

2. DATASET

USC-TIMIT [26] is a rich database of the rtMRI videos of the upper airway in the midsagittal plane with a spatial resolution of 68×68 pixels (2.9 mm $\times$2.9 mm) at 23.18 frame/sec. The USC-TIMIT rtMRI database contains data from five male and five female subjects speaking a set of 460 sentences taken from the MOCHA-TIMIT corpus [27]. The experiments in this work use rtMRI data for 10 sentences each from two male subjects (M1,M2) and two female subjects (F1,F2). The selected ten sentences correspond to 856, 753, 987 and 779 rtMRI frames for F1, F2, M1 and M2 respectively.

A MATLAB based GUI was used to manually trace the ATB of the rtMRI frames, the details of which is available in [28]. Fig. 1(a) shows the three major manually drawn contours representing the complete ATBs in a typical rtMRI frame. Upper lip (UL), lower lip (LL), tongue base (TB), velum tip (VEL) and glottis begin (GLTB) were also marked for each frame using the GUI. For the ATB segmentation in this work as shown in Fig. 1(b) and Fig. 1(c) , Contour1 (C_1) is divided into three parts - C_{11} corresponding to the upper lip (UL), C_{12} corresponding to the hard palate whose position and shape is manually chosen and kept fixed across video frames for a subject and C_{13} corresponding to the Velum (VEL). Similarly, Contour2 (C_2) is divided into three parts - C_{21} which covers the lower lip and jaw till the tongue base (TB), C_{22} which extends from the tongue base (TB) along the tongue blade till the epiglottis (determined by the location of the groove in the epiglottis region) and C_{23} extending below the epiglottis. The contour C_{31} which marks the pharyngeal wall till GLTB, remains fixed across all the video frames for a subject.

3. PROPOSED AIR-TISSUE BOUNDARY SEGMENTATION

A measure of contrast along a given contour and a measure of proximity between two contours of different lengths are required to describe the proposed segmentation technique.

3.1. Fisher Discriminant: Measure of Contrast

The proposed ATB segmentation approach uses the Fisher discriminant function to quantify the contrast between the pixel intensities on either sides of a given contour. A contour consisting of M points is defined as: $C \triangleq \{(x_i, y_i), 1 \leq i \leq M\}$, where x_i and y_i denote the X and Y coordinates of the i^{th} point on the contour. x_i and y_i align with the column and row indices respectively starting from the top left corner of an rtMRI frame. In order to find a measure of contrast along a given contour, the inner contour C_{in} and the outer contour C_{out} are constructed from C. Each point of C_{in} and C_{out} is found by drawing a normal to the corresponding point in C. The distance between the corresponding points of C_{in} & C and C_{out} & C is equal to the Euclidean distance between two successive points of C.

For each point on C_{in} and C_{out} the corresponding pixel value of the image is found using bicubic interpolation [29]. Thus the collection of pixel intensities along C_{in} (I_{in}) and along C_{out} (I_{out}) are denoted as: $I_{in} = \{I(x_i, y_i) \mid (x_i, y_i) \in C_{in}\}$ and $I_{out} = \{I(x_i, y_i) \mid (x_i, y_i) \in C_{out}\}$, where I denotes an image. The Fisher Discriminant Measure (FDM) for a given contour C and an image I is defined as:

$$\mathcal{D}_F(C, I) = \frac{(\overline{I_{in}} - \overline{I_{out}})^2}{\sigma^2{}_{I_{in}} + \sigma^2{}_{I_{out}}} \tag{1}$$

where $\sigma^2{}_{I_{in}}$ and $\sigma^2{}_{I_{out}}$ are the variances of pixel intensities of I_{in} and I_{out} respectively and $\overline{I_{in}}$ and $\overline{I_{out}}$ denote the sample average of their respective pixel intensities. A high FDM results from not only a large difference between the average pixel intensities from the inner and outer regions but also the *uniformity* (low variance) of pixel intensities in each region. The FDM value reflects the contrast along the entire contour.

3.2. Measure of Proximity Between Two Contours

The alignment of any two given contours is measured using the DTW distance [30]. Consider two contours $C_a = \{(x_i^a, y_i^a) \mid 1 \leq i \leq M_a\}$ and $C_b = \{(x_i^b, y_i^b) \mid 1 \leq i \leq M_b\}$ such that $C_a(i) \in \mathbb{R}^2$ and $C_b(j) \in \mathbb{R}^2$ represent the i^{th} and the j^{th} points' co-ordinates in C_a and C_b respectively. In order to find an optimal alignment map $\{(m_a(l), m_b(l)) \mid 1 \leq l \leq L, 1 \leq m_a(l) \leq M_a$ and $1 \leq m_b(l) \leq M_b\}$ between the points of C_a and C_b, the following optimization is performed:

$$\{(m_a(l), m_b(l)), 1 \leq l \leq L\} = \underset{\substack{1 \leq m'_a(l) \leq M_a, \\ 1 \leq m'_b(l) \leq M_b}}{\operatorname{argmin}} \sum_{l=1}^{L} ||C_a(m'_a(l)) - C_b(m'_b(l))||_2 \tag{2}$$

The DTW distance between two contours C_a and C_b is defined as:

$$\mathcal{D}_D(C_a, C_b) \triangleq \frac{1}{L} \sum_{l=1}^{L} ||C_a(m_a(l)) - C_b(m_b(l))||_2 \tag{3}$$

$\mathcal{D}_D(C_a, C_b)$ is less if two contours C_a and C_b have similar shape and located close to each other. From the above equations, it can be seen that the value of L is dependent on the lengths of the contours C_a and C_b (M_a and M_b respectively). The distance measures $\mathcal{D}_F$ and $\mathcal{D}_D$ can be computed irrespective of the lengths of the contours (M_a and M_b).

The steps in the proposed ATB segmentation approach are summarized in Fig. 3. Following pre-processing of the input test rtMRI video, ATBs of different parts of C_1 and C_2 are predicted. Note that C_{12} and C_{31} are not predicted, rather are fixed to a manually chosen contour as these parts do not move during speaking. The predicted contours are finally stitched and pruned to obtain the upper airway

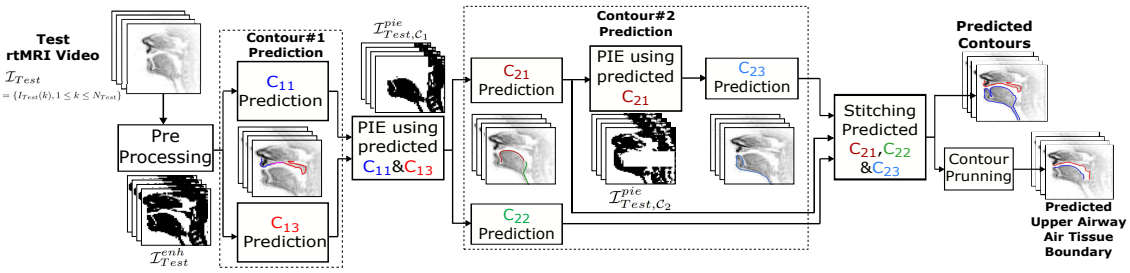

Fig. 3. Illustration of the steps in the proposed FDM-based approach

ATBs. The details of the steps are described in the following subsections.

3.3. Pre-Processing

Each frame of a test rtMRI video is enhanced using the technique used in [15] to reduce the rtMRI artifact for better predictions of the ATBs. Let a test rtMRI video containing N_{Test} frames be represented by $\mathcal{I}_{Test}$ such that $\mathcal{I}_{Test}(k)$ represents the k^{th} rtMRI frame of the video. Following pre-processing, the enhanced video is represented by $\mathcal{I}^{enh}_{Test}$.

3.4. Air-Tissue Boundary Prediction

The ATBs $C_{11}, C_{13}, C_{21}, C_{22}, C_{23}$ are predicted in a specific order. First the C_{11} and C_{13} are predicted using $\mathcal{I}^{enh}_{Test}$ which is followed by partial image erosion (PIE) using the predicted C_1. The image sequence after PIE is used as the input to the prediction of C_{21} and C_{22}. Finally C_{23} is predicted after PIE using the predicted C_{21}. As different image sequences are used as the input for prediction of different parts, we describe the boundary prediction using a generic symbol for image sequence and the contour, namely $\mathcal{I}$ and $\mathcal{C}$ respectively.

Let $\mathcal{I}$ represent an image sequence in an rtMRI video of length N, where k^{th} image is denoted by $\mathcal{I}(k)$. Let $\mathcal{C} = \{\mathcal{C}(k), 1 \leq k \leq N\}$ denote the set of the contours of interest for boundary prediction in N different images where $\mathcal{C}(k)$ denotes the contour in the k^{th} image. Let $\mathcal{C}^{Tr}$ be the respective set of N_{Tr} training contours. The boundaries in N images are predicted by selecting the best contour from the training set in each image such that the predicted contour sequence varies smoothly as well as maximizes the overall FDM. For this, the objective function $J(\mathcal{C}, \mathcal{I})$ is defined as:

$$J(\mathcal{C}, \mathcal{I}) = \sum_{k=2}^{N} \mathcal{D}_F(\mathcal{C}(k), \mathcal{I}(k)) - \lambda \mathcal{D}_D(\mathcal{C}(k), \mathcal{C}(k-1)) \quad (4)$$

$\mathcal{D}_F$ and $\mathcal{D}_D$ are defined in Eq. 1 and 3. The sequence of predicted contours for all the frames of $\mathcal{I}$ is obtained as:

$$\mathcal{C}^* = \{\mathcal{C}^*(k), 1 \leq k \leq N\} = \underset{\mathcal{C} \in \{\mathcal{C}^{Tr}(i), 1 \leq i \leq N_{Tr}\}}{\operatorname{argmax}} J(\mathcal{C}, \mathcal{I}) \quad (5)$$

The optimization problem above is solved using DP. The constant λ in Eq. 4 is the temporal stiffness factor. The optimal value of λ for every contour part is obtained separately using a development set. These are denoted by $\lambda_{\mathcal{C}_{11}}, \lambda_{\mathcal{C}_{13}}, \lambda_{\mathcal{C}_{21}}, \lambda_{\mathcal{C}_{22}}, \lambda_{\mathcal{C}_{23}}$ for $\mathcal{C}_{11}, \mathcal{C}_{13}$, $\mathcal{C}_{21}, \mathcal{C}_{22}, \mathcal{C}_{23}$ respectively.

As C_{12} is kept fixed during the optimization process, the continuity of the contours at boundary points of $\mathcal{C}_{11}$ & $\mathcal{C}_{12}$ and $\mathcal{C}_{12}$ & $\mathcal{C}_{13}$ is maintained by appending the extreme points of the hard palate ($\mathcal{C}_{12}$) to the training contours $\mathcal{C}^{Tr}_{11}$ and $\mathcal{C}^{Tr}_{13}$ respectively. Thus the complete set of predicted upper contours $\hat{\mathcal{C}}_1$ for the image sequence $\mathcal{I}_{Test}$ is constructed by concatenating $\mathcal{C}^*_{11}$, $\mathcal{C}^*_{12}$ and $\mathcal{C}^*_{13}$ and removing the duplicate boundary points.

Similarly, $\hat{\mathcal{C}}_2$ is obtained by concatenating $\mathcal{C}^*_{21}$, $\mathcal{C}^*_{22}$,, $\mathcal{C}^*_{23}$. In order to avoid $\hat{\mathcal{C}}_2$ intersecting $\hat{\mathcal{C}}_1$, we perform PIE of $\mathcal{I}^{enh}_{Test}$ using $\hat{\mathcal{C}}_1$. Details of PIE are described in the next subsection. It should be noted that the part $\mathcal{C}_{31}$ of contour $\mathcal{C}_3$ is not predicted rather $\hat{\mathcal{C}}_{31}$ is kept fixed to a manually chosen contour for all rtMRI frames.

3.5. Partial Image Erosion

Partial Image Erosion (PIE) is performed in order to ensure that $\hat{\mathcal{C}}_2$ is below the predicted upper ATB ($\hat{\mathcal{C}}_1$) and to improve the accuracy of $\mathcal{C}^*_{23}$. PIE is performed twice in the proposed ATB prediction - (1) before predicting $\mathcal{C}^*_{21}$ and (2) before predicting $\mathcal{C}^*_{23}$. Before the prediction of $\mathcal{C}^*_{21}$, all the training contours $\mathcal{C}^{Tr}_{21}$ which intersect with $\hat{\mathcal{C}}_1$ for the respective rtMRI frames are removed. Then the set of pixels in a column with a row index lesser than the respective $\hat{\mathcal{C}}_1$ is made zero in each frame of $\mathcal{I}^{enh}_{Test}$. The modified sequence of rtMRI frames, thus obtained, is represented by $\mathcal{I}^{pie}_{Test,C_1}$. After $\mathcal{C}^*_{21}$ is obtained, to prevent $\mathcal{C}^*_{23}$ from intersecting $\mathcal{C}^*_{21}$, the collection of pixels in a row of $\mathcal{I}^{enh}_{Test}$ with column indices more than those in the points in $\mathcal{C}^*_{21}$ are made zero. The sequence obtained from this operation is represented by $\mathcal{I}^{pie}_{Test,C_2}$.

PIE before predicting $\mathcal{C}^*_{21}$ and $\mathcal{C}^*_{23}$ ensures that the solution of the respective optimization (Eq. 5) comes from a subset of $\mathcal{C}^{Tr}$ which do not intersect with $\hat{\mathcal{C}}_1$ and $\mathcal{C}^*_{21}$ thus resulting in a more accurate ATB prediction.

3.6. Contour Stitching

Because the lower ATB prediction is done in three separate parts, a simple concatenation of $\mathcal{C}^*_{21}$, $\mathcal{C}^*_{22}$ and $\mathcal{C}^*_{23}$ does not ensure a smooth $\hat{\mathcal{C}}_2$. In order to prevent erratic and jagged contours at the junctions of $\mathcal{C}^*_{21}$ & $\mathcal{C}^*_{22}$ and $\mathcal{C}^*_{21}$ & $\mathcal{C}^*_{23}$, contour stitching is performed. This is done by considering the end parts of the two contours at the junctions and trimming the end of the contour with higher row index till it matches with the row index of the end point of the other contour. To illustrate the contour stitching, a zoomed in view of the part of the rtMRI frame in the pink box in Fig. 1(b) is shown in Fig. 2(b), where the red and blue points correspond to the contour $\mathcal{C}^*_{21}$ and $\mathcal{C}^*_{23}$ respectively. $\mathcal{C}^*_{21}$ is trimmed to obtain a smooth lower ATB $\hat{\mathcal{C}}_2$ (shown in green).

3.7. Contour Pruning

The predicted ATBs as described in the section 3.4 span regions both inside and outside the vocal tract. In order to obtain boundaries within the vocal tract, we use two different strategies for upper ($\hat{\mathcal{C}}_1$) and lower ($\hat{\mathcal{C}}_2$) ATBs. For pruning $\hat{\mathcal{C}}_1$, at first the velum tip is automatically detected by finding out the index for change in the direction of row values in $\mathcal{C}^*_{13}$. Following this, $\hat{\mathcal{C}}_1$ is segmented from UL to VEL tip and concatenated with $\mathcal{C}^*_{31}$ till GLTB to obtain $\hat{\mathcal{C}}^{prun}_1$. Note that the point corresponding to GLTB is a part of $\mathcal{C}^*_{31}$ and thus remains fixed for all the frames of the test rtMRI video.

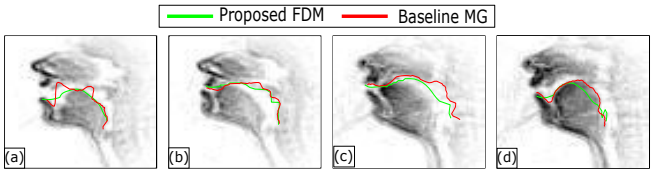

Fig. 4. Illustration of upper airway ATBs using MG and FDM schemes.

Similarly, the $\hat{\mathcal{C}}_2$ is pruned from LL to GLTB. However, the segment of $\hat{\mathcal{C}}_2$ near tongue base (near the junction between $\mathcal{C}^*_{21}$ and $\mathcal{C}^*_{23}$) does not reflect the actual vocal tract cross sectional area due to the presence of lower teeth. In order to obtain a smooth boundary in this region, at first, the point (C^1_{tb}) with the lowest row index in $\mathcal{C}^*_{23}$ (typically near LL) is identified and the point (C^2_{tb}) on the $\mathcal{C}^*_{21}$ with this row index is selected. A segment of length N_{tb} in $\hat{\mathcal{C}}_2$ from C^1_{tb} to C^2_{tb} is denoted by $\mathcal{C}_{tb} = \{(x_i^{tb}, y_i^{tb}), 1 \leq i \leq N_{tb}\}$. $\mathcal{C}_{tb}$ in $\hat{\mathcal{C}}_2$ is replaced with $\mathcal{C}_{sm} = \{(x_i^{tb}, y_i^{sm}), 1 \leq i \leq N_{tb}\}$, where $y_i^{sm} \triangleq a_0 + a_1 x_i^{tb} + a_2 (x_i^{tb})^2$. Coefficients of the polynomial are obtained as follows:

$$\{a_0, a_1, a_2\} = \underset{\alpha,\beta,\gamma}{\operatorname{argmin}} \sum_{i=1}^{N_{tb}} (y_i^{tb} - (\alpha + \beta x_i^{tb} + \gamma (x_i^{tb})^2))^2 \quad (6)$$
$$subject\ to\ \alpha + \beta x_i^{tb} + \gamma (x_i^{tb})^2 \leq y_i^{tb}, \forall i$$

After $\mathcal{C}_{tb}$ is replaced with $\mathcal{C}_{sm}$, the pruned predicted lower ATB is denoted by $\hat{\mathcal{C}}_2^{prun}$.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

The ATBs are estimated from the rtMRI data for each subject (F1, F2, M1 and M2) separately using a five-fold cross-validation setup. In each fold, eight training and two test rtMRI videos are used in a round-robin fashion. Among the eight training rtMRI videos, five are used for training and the remaining three videos are used as the development set. The training contours corresponding to different parts of $\mathcal{C}_1$ and $\mathcal{C}_2$ are obtained from the manually traced boundaries as illustrated in Fig. 1(b) and 1(c).

The evaluation of the predicted contours is done using the DTW Distance $\mathcal{D}_D$ between the manually traced and predicted ATBs (Eq. 3). The DTW Distances have the unit of pixel. In this work, we have performed two kinds of evaluations: (1) evaluation of the ATBs within the vocal tract ($\hat{\mathcal{C}}_1^{prun}$, $\hat{\mathcal{C}}_2^{prun}$) predicted using FDM. A Maeda Grid (MG) based approach [15] is used as a baseline for comparison, (2) evaluation of the complete predicted contours $\hat{\mathcal{C}}_1$, $\hat{\mathcal{C}}_2$, and $\mathcal{C}_3$. To obtain the ground truth contour for evaluation of $\hat{\mathcal{C}}_1^{prun}$ and $\hat{\mathcal{C}}_2^{prun}$, we have pruned the upper and lower manually traced ATBs within vocal tract following the steps outlined in section 3.7. The pruned manually traced boundaries are denoted by $\mathcal{C}_1^{prun}$ and $\mathcal{C}_2^{prun}$. The evaluation of MG and FDM based approaches was done by comparing the predictions of each approach with the corresponding hand-annotated ground truth ATBs.

4.2. Results and Discussion

Table 1 shows the average ($\pm$ standard deviation) $\mathcal{D}_D(\mathcal{C}_1^{prun}, \hat{\mathcal{C}}_1^{prun})$ and $\mathcal{D}_D(\mathcal{C}_2^{prun}, \hat{\mathcal{C}}_2^{prun})$ (in pixels) using both MG and the proposed FDM schemes. It is clear from the table that the proposed FDM approach, on average, results in a lower DTW distance compared to the baseline MG scheme. The average error of the lower and upper ATBs across the four subjects from the FDM approach is

	Lower ATB		Upper ATB	
Sub	***MG***	***FDM***	***MG***	***FDM***
F1	1.09 ± 0.22	1.02 ± 0.24	1.00 ± 0.17	0.95 ± 0.17
F2	1.28 ± 0.29	1.27 ± 0.26	1.42 ± 0.35	1.20 ± 0.22
M1	1.31 ± 0.57	1.25 ± 0.26	1.18 ± 0.19	1.10 ± 0.20
M2	1.38 ± 0.31	1.17 ± 0.28	1.37 ± 0.23	1.17 ± 0.24

Table 1. ATB prediction error in pixels (average $\pm$ standard deviation) using MG and FDM schemes

8.99% lower than the average error in the predicted contours obtained from the baseline MG scheme. Fig. 4(a) and (b) show two sample rtMRI frames for which the ATBs obtained by the proposed FDM approach are more accurate than the baseline MG scheme. The superior performance using the FDM scheme could be due to the fact that the FDM (Eq. 1) is robust to local rtMRI artifact. The temporal constraint used in the optimization (Eq. 4) also prevents the proposed FDM approach from predicting jagged contours and yields smoothly varying contours across rtMRI frames.

Fig. 4(c) and Fig. 4(d) illustrate two frames where the MG approach yields more accurate boundaries than the FDM approach. This happens because the training contours of the subject do not have a velum contour (C_13) as observed in the test case. The predicted ATB in Fig. 4(d) is not as accurate as the one obtained from the MG scheme. This happens because a significant length of the velum tissue is in contact with the tongue dorsal causing FDM value to drop for the actual ATB contour.

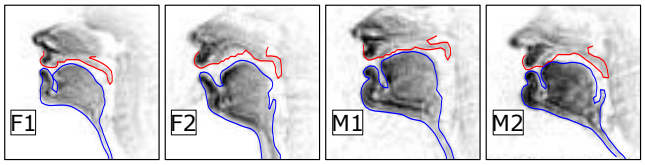

Fig. 5. Illustrations of full contour prediction using FDM scheme

In addition to predicting the pruned ATBs inside the vocal tract, complete contours $\hat{C}_1$ and $\hat{C}_2$ are also predicted as shown in Fig. 5 using one example frame for each of the four subjects. The evaluation of the full predicted contours $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$ was done separately. The average ($\pm$ standard deviation) DTW distances (in pixels) between $\hat{\mathcal{C}}_1$ and the ground truth for F1, F2, M1 and M2 are 0.92 ± 0.12, 1.09 ± 0.19, 1.13 ± 0.18 and 1.17 ± 0.25 respectively. Similarly, the average ($\pm$ standard deviation) DTW distances (in pixels) between $\hat{\mathcal{C}}_2$ and the ground truth for F1, F2, M1 and M2 are 0.83 ± 0.13, 0.99 ± 0.17, 0.98 ± 0.16 and 0.98 ± 0.18 respectively. Thus it is clear that the proposed FDM scheme reliably predicts the complete ATB in both inside and outside the vocal tract.

5. CONCLUSION

In this work, we propose a supervised approach for ATB prediction in the midsagittal rtMRI videos. As the ATB shapes are learned from the training data, the proposed method performs well across four subjects considered in this work. This robust performance of the proposed scheme is due to the proposed measure of contrast and joint prediction of ATBs across all frames in a video ensuring temporal continuity unlike frame-by-frame ATB prediction in existing methods. The proposed scheme could be further improved by developing a deformation model of the contour to deform a training contour to better fit a given frame.

6. ACKNOWLEDGEMENTS

The authors thank the Pratiksha Trust for their support and Anisha Banerjee for helping in the marking of boundaries in rtMRI videos.

7. REFERENCES

[1] Erik Bresch, Yoon-Chul Kim, Krishna Nayak, Dani Byrd, and Shrikanth Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," *in IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 123–132, 2008.

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

[3] Peter Ladefoged, Richard Harshman, Louis Goldstein, and Lloyd Rice, "Generating vocal tract shapes from formant frequencies," *in The Journal of the Acoustical Society of America*, vol. 64, no. 4, pp. 1027–1035, 1978.

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

[5] Shrikanth Narayanan, Krishna Nayak, Sungbok Lee, Abhinav Sethy, and Dani Byrd, "An approach to real-time magnetic resonance imaging for speech production," *in The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, 2004.

[6] Brad H Story, Ingo R Titze, and Eric A Hoffman, "Vocal tract area functions from magnetic resonance imaging," *in The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 537–554, 1996.

[7] Nimisha Patil, Timothy Greer, Reed Blaylock, and Shrikanth Narayanan, "Comparison of basic beatboxing articulations between expert and novice artists using real-time magnetic resonance imaging," *in Interspeech*, pp. 2277–2281, 2017.

[8] Vikram Ramanarayanan, Louis Goldstein, Dani Byrd, and Shrikanth S Narayanan, "An investigation of articulatory setting using real-time magnetic resonance imaging," *in The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 510–519, 2013.

[9] Adam Lammert, Michael Proctor, and Shrikanth Narayanan, "Interspeaker variability in hard palate morphology and vowel production," *in Journal of Speech, Language, and Hearing Research*, vol. 56, no. 6, pp. S1924–S1933, 2013.

[10] Asterios Toutios, Tanner Sorensen, Krishna Somandepalli, Rachel Alexander, and Shrikanth S Narayanan, "Articulatory synthesis based on real-time magnetic resonance imaging data.," *in Interspeech*, pp. 1492–1496, 2016.

[11] Benjamin Parrell and SS Narayanan, "Interaction between general prosodic factors and language-specific articulatory patterns underlies divergent outcomes of coronal stop reduction," *in International Seminar on Speech Production (ISSP) Cologne, Germany*, pp. 308–311, 2014.

[12] Fang-Ying Hsieh, Louis Goldstein, Dani Byrd, and Shrikanth Narayanan, "Truncation of pharyngeal gesture in english diphthong [ał].," *in Interspeech*, pp. 968–972, 2013.

[13] Abhay Prasad, Vijitha Periyasamy, and Prasanta Kumar Ghosh, "Estimation of the invariant and variant characteristics in speech articulation and its application to speaker identification," *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4265–4269, 2015.

[14] Ming Li, Jangwon Kim, Adam Lammert, Prasanta Kumar Ghosh, Vikram Ramanarayanan, and Shrikanth Narayanan, "Speaker verification based on the fusion of speech acoustics and inverted articulatory signals," *in Computer speech & language*, vol. 36, pp. 196–211, 2016.

[15] Jangwon Kim, Naveen Kumar, Sungbok Lee, and Shrikanth Narayanan, "Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data," *in International Seminar on Speech Production, ISSP*, pp. 222–225, 2014.

[16] Sven EG Öhman, "Numerical model of coarticulation," *in The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 310–320, 1967.

[17] Shinji Maeda, "Une modele articulatoire de la tongue avec des composantes lineaires," *in JEP, GALF*, vol. 10, pp. 152–164, 1979.

[18] Michael I. Proctor, Danny Bone, Nassos Katsamanis, and Shrikanth Narayanan, "Rapid semiautomatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis," in *Proceedings of the International Conference on Speech Communication and Technology*, 2010.

[19] Adam C Lammert, Vikram Ramanarayanan, Michael I Proctor, Shrikanth Narayanan, et al., "Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis.," pp. 959–962, 2013.

[20] Sasan Asadiabadi and Engin Erzin, "Vocal tract airway tissue boundary tracking for rtMRI using shape and appearance priors," *in Interspeech*, pp. 636–640, 2017.

[21] Krishna Somandepalli, Asterios Toutios, and Shrikanth S Narayanan, "Semantic edge detection for tracking vocal tract air-tissue boundaries in real-time magnetic resonance images," *in Interspeech*, pp. 631–635, 2017.

[22] Adam C Lammert, Michael I Proctor, and Shrikanth S Narayanan, "Data-driven analysis of realtime vocal tract MRI using correlated image regions," *in Interspeech*, 2010.

[23] Asterios Toutios and Shrikanth S Narayanan, "Factor analysis of vocal-tract outlines derived from real-time magnetic resonance imaging data," *in International Congress of Phonetic Sciences (ICPhS), Glasgow, UK*, 2015.

[24] Tanner Sorensen, Asterios Toutios, Louis Goldstein, and SS Narayanan, "Characterizing vocal tract dynamics with real-time MRI," *in 15th Conference on Laboratory Phonology, Ithaca, NY*, 2016.

[25] Dawei Zhang, Minghao Yang, Jianhua Tao, Yang Wang, Bin Liu, and Danish Bukhari, "Extraction of tongue contour in real-time magnetic resonance imaging sequences," *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 937–941, 2016.

[26] Shrikanth Narayanan, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee, Krishna Nayak, Yoon-Chul Kim, Yinghua Zhu, Louis Goldstein, et al., "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)," *in The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

[27] 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.

[28] Ashok Kumar Pattem, Aravind Illa, Amber Afshan, and Prasanta Kumar Ghosh, "Optimal sensor placement in electromagnetic articulography recording for speech production study," *in Computer speech & language*, vol. 47, pp. 157–174, 2018.

[29] Robert Keys, "Cubic convolution interpolation for digital image processing," *in IEEE transactions on acoustics, speech, and signal processing*, vol. ASSP-29, no. 6, pp. 1153–1160, 1981.

[30] Donald J Berndt and James Clifford, "Using dynamic time warping to find patterns in time series," *in KDD workshop*, vol. 10, no. 16, pp. 359–370, 1994.