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\textsuperscript{1} Prasanta Kumar Ghosh\textsuperscript{2}

\textsuperscript{1}Electrical and Electronics Engineering, Birla Institute of Technology and Science, Goa-403726, India
\textsuperscript{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 \cite{1}, rtMRI proves itself to be more effective than other available methods including Electromagnetic Articulography (EMA) \cite{2}, X-Ray \cite{3} and Ultrasound \cite{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 \cite{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 \cite{6} which often forms the basis for speech processing applications. For example, Patil et al. \cite{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 \cite{8} and morphological structures of the vocal tract \cite{9} require segmentation of rtMRI frames as a pre-processing step. Toutios \cite{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 \cite{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 \cite{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. \cite{19}. Asadi-badi and Erzin \cite{20} presented a statistical approach for segmentation based on appearance and shape models for the human vocal tract. Somandepalli et al. \cite{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. \cite{22} also presented a data-driven approach to the segmentation problem based on average intensities of pixels. The approach applied by Toutios \cite{23} and Sorensen \cite{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. \cite{25}. Although unsupervised, semi-automatic approaches such as those presented in \cite{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-
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 × 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 (C1) were also marked for each frame using the GUI. For the ATB segmentation in this work, the distance between the corresponding points of C1n & C2n and Cout & C is equal to the Euclidean distance between two successive points of C.

For each point on C1n and Cout, the corresponding pixel value of the image is found using bicubic interpolation [29]. Thus the collection of pixel intensities along C1n (I1n) and along Cout (Iout) are denoted as: I1n = {I(x1, y1) | (x1, y1) ∈ C1n} and Iout = {I(x1, y1) | (x1, y1) ∈ Cout}, where I denotes an image. The Fisher Discriminant Measure (FDM) for a given contour C and an image I is defined as:

$$D_F(C, I) = \frac{(I_{out} - I_{in})^2}{\sigma^2_{in} + \sigma^2_{out}}$$

where $\sigma^2_{in}$ and $\sigma^2_{out}$ are the variances of pixel intensities of I1n and Iout respectively and I1n and Iout 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.

3.3. Measure of Proximity Between Two Contours

The alignment of any two given contours is measured using the DTW distance [30]. Consider two contours $C_a = \{(x^a_i, y^a_i) \mid 1 \leq i \leq M_a\}$ and $C_b = \{(x^b_i, y^b_i) \mid 1 \leq i \leq M_b\}$ such that $C_a(i) \in \mathbb{R}^2$ and $C_b(j) \in \mathbb{R}^2$ represent the ith and the jth 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:

$\{ (m_a(l), m_b(l)) \mid 1 \leq l \leq L \} = \text{argmin}_{1 \leq m_a(l) \leq M_a, 1 \leq m_b(l) \leq M_b} \sum_{l=1}^L ||C_a(m_a(l)) - C_b(m_b(l))||_2$\)

(2)

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

$$D_D(C_a, C_b) = \frac{1}{L} \sum_{l=1}^L ||C_a(m_a(l)) - C_b(m_b(l))||_2$$

(3)

$D_D(C_a, C_b)$ 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 $D_F$ and $D_D$ 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.
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 $I_{Test}(k)$ such that $I_{Test}(k)$ represents the $k^{th}$ rtMRI frame of the video. Following pre-processing, the enhanced video is represented by $I_{Pre}^{enh}(k)$.

### 3.4. Air-Tissue Boundary Prediction

The ATBs $C11, C13, C21, C22, C23$ are predicted in a specific order. First the $C11$ and $C13$ are predicted using $I_{Pre}^{enh}(k)$ which is followed by partial image erosion (PIE) using the predicted $C1$. The image sequence after PIE is used as the input to the prediction of $C21$ and $C22$. Finally $C21$ is predicted after PIE using the predicted $C21$. 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 $I$ and $C$ respectively.

Let $I$ represent an image sequence in an rtMRI video of length $N$, where $k^{th}$ image is denoted by $I(k)$. Let $C = \{C(k), 1 \leq k \leq N\}$ denote the set of the contours of interest for boundary prediction in $N$ different images where $C(k)$ denotes the contour in the $k^{th}$ image. Let $C^{Tr}$ 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 $J(C, I)$ is defined as:

$$J(C, I) = \sum_{k=2}^{N} D_{F}(C(k), I(k)) - \lambda D_{D}(C(k), C(k-1))$$

(4)

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

$$C^* = \{C^*(k), 1 \leq k \leq N\} = \arg\max_{C \in \{C^{Tr}(k), 1 \leq k \leq N_{Tr}\}} J(C, I)$$

(5)

The optimization problem above is solved using DP. The constant $\lambda$ in Eq. 4 is the temporal stiffness factor. The optimal value of $\lambda$ for every contour part is obtained separately using a development set. These are denoted by $\lambda_{C11}, \lambda_{C13}, \lambda_{C21}, \lambda_{C22}, \lambda_{C23}$ for $C11, C13, C21, C22, C23$ respectively.

As $C_{12}$ is kept fixed during the optimization process, the continuity of the contours at boundary points of $C_{11}$ and $C_{12}$ and $C_{12}$ and $C_{13}$ is maintained by appending the extreme points of the hard palate ($C_{12}$) to the training contours $C^{Tr}_{11}$ and $C^{Tr}_{13}$ respectively. Thus the complete set of predicted upper contours $C_{1}$ for the image sequence $I_{Test}$ is constructed by concatenating $C_{11}, C_{12}$ and $C_{13}$ and removing the duplicate boundary points.

Similarly, $C_{2}$ is obtained by concatenating $C^{*}_{21}, C^{*}_{22}, C^{*}_{23}$. In order to avoid $C_{2}$ intersecting $C_{1}$, we perform PIE of $I_{Pre}^{enh}$ using $C_{1}$. Details of PIE are described in the next subsection. It should be noted that the part $C_{31}$ of contour $C_{3}$ is not predicted rather $C_{31}$ 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 $C_{2}$ is below the predicted upper ATB ($C_{1}$) and to improve the accuracy of $C^{*}_{21}$. PIE is performed twice in the proposed ATB prediction - (1) before predicting $C_{21}$ and (2) before predicting $C_{22}$. Before the prediction of $C_{21}$, all the training contours $C^{Tr}_{21}$ which intersect with $C_{1}$ for the respective rtMRI frames are removed. Then the set of pixels in a column with a row index lesser than the respective $C_{1}$ is made zero in each frame of $I_{Pre}^{enh}$. The modified sequence of rtMRI frames, thus obtained, is represented by $I_{Pre}^{pie, C_{1}}$. After $C_{21}$ is obtained, to prevent $C_{22}$ from intersecting $C_{21}$, the collection of pixels in a row of $I_{Pre}^{enh}$, with column indices more than those in the points in $C_{21}$ are made zero. The sequence obtained from this operation is represented by $I_{Pre}^{pie, C_{1}, C_{21}}$.

 PIE before predicting $C_{21}$ and $C_{22}$ ensures that the solution of the respective optimization (Eq. 5) comes from a subset of $C^{Tr}$ which do not intersect with $C_{1}$ and $C_{21}$ 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 $C_{21}^{*}, C_{22}^{*}$ and $C_{23}^{*}$ does not ensure a smooth $C_{2}$. In order to prevent erratic and jagged contours at the junctions of $C_{21}^{*}$ and $C_{22}^{*}$ and $C_{21}^{*}$ and $C_{23}^{*}$, 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 $C_{21}^{*}$ and $C_{23}^{*}$ respectively. $C_{21}$ is trimmed to obtain a smooth lower ATB $C_{2}$ (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 ($C_{1}$) and lower ($C_{2}$) ATBs. For pruning $C_{1}$, at first the velum tip is automatically detected by finding out the index for change in the direction of row values in $C_{11}^{*}$. Following this, $C_{1}$ is segmented from UL to VEL tip and concatenated with $C_{11}$ till GLTB to obtain $C_{11}^{prun}$. Note that the point corresponding to GLTB is a part of $C_{11}$ and thus remains fixed for all the frames of the test rtMRI video.
Similarly, the $C_2$ is pruned from LL to GLTB. However, the segment of $C_2$ near tongue base (near the junction between $C_{21}$ and $C_{23}$) 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 $(C^b_{ib})$ with the lowest row index in $C^b_{21}$ (typically near LL) is identified and the point $(C^b_{ib})$ on the $C^b_{21}$ with this row index is selected. A segment of length $N_{ib}$ in $C_2$ from $C^b_{ib}$ to $C^b_{ib}$ is denoted by $C_{ib} = \{(x_{ib}^b, y_{ib}^b), 1 \leq i \leq N_{ib}\}$. $C_{ib}$ in $C_2$ is replaced with $C_{sm} = \{(x_{ib}^{sm}, y_{ib}^{sm}), 1 \leq i \leq N_{ib}\}$, where $y_{ib}^{sm} \leq a_0 + a_1 x_{ib}^b + a_2 (x_{ib}^b)^2$. Coefficients of the polynomial are obtained as follows:

$$\begin{align*}
\{a_0, a_1, a_2\} &= \arg\min_{\alpha, \beta, \gamma} \sum_{i=1}^{N_{ib}} \left( y_{ib}^b - (\alpha + \beta x_{ib}^b + \gamma (x_{ib}^b)^2) \right)^2 \\
&\quad \text{subject to } \alpha + \beta x_{ib}^b + \gamma (x_{ib}^b)^2 \leq y_{ib}^b, \forall i
\end{align*}$$

After $C_{ib}$ is replaced with $C_{sm}$, the pruned predicted lower ATB is denoted by $C_2^{prun}$.

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 correspond to different parts of $C_1$ and $C_2$ are obtained from the manually traced boundaries as illustrated in Fig. (b) and (c).

The evaluation of the predicted contours is done using the DTW Distance $D_D$ 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 ($C_1^{prun}$, $C_2^{prun}$) predicted using FDM. A Maeda Grid (MG) based approach [15] is used as a baseline for comparison, (2) evaluation of the complete predicted contours $C_1$, $C_2$, and $C_3$. To obtain the ground truth contour for evaluation of $C_1^{prun}$ and $C_2^{prun}$, 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 $C_1^{prun}$ and $C_2^{prun}$. 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 (± standard deviation) $D_D(C_1^{prun}, C_2^{prun})$ and $D_D(C_1^{prun}, C_2^{prun})$ (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 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 ($C_{21}$) 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.

In addition to predicting the pruned ATBs inside the vocal tract, complete contours $C_1$ and $C_2$ 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 $C_1$ and $C_2$ was done separately. The average (± standard deviation) DTW distances (in pixels) between $C_1$ 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 (± standard deviation) DTW distances (in pixels) between $C_2$ 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


