A deep neural network based correction scheme for improved air-tissue boundary prediction in real-time magnetic resonance imaging video

Renuka Mannem*, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, Karnataka, India

ARTICLE INFO

Article History:
Received 20 February 2020
Revised 22 September 2020
Accepted 23 September 2020
Available online 28 September 2020

ABSTRACT

The real-time Magnetic Resonance Imaging (rtMRI) video captures the vocal tract movements in the mid-sagittal plane during speech. Air tissue boundaries (ATBs) are contours that trace the transition between the high-intensity tissue corresponding to the speech articulators and the low-intensity airway cavity in the rtMRI video. The ATB segmentation in an rtMRI video is a common preprocessing step which is used for many speech production and speech processing applications. However, ATB segmentation is very challenging due to the low resolution and low signal-to-noise ratio of the rtMRI images. Several works have been proposed in the literature for accurate ATB segmentation. However, every ATB segmentation technique, be it knowledge-based or data-driven, has its own limitations due to model assumption or data quality. The errors in the predicted ATBs from a typical ATB segmentation approach can be corrected in a data-driven manner as a post-processing step. In this work, we propose a deep neural network (DNN) based correction scheme for improving the ATB segmentation. In the DNN based correction approach, the correction of each point on a predicted ATB is done using a pattern of intensity variation in the direction of the normal to the predicted ATB at that point. For this, inputs and target outputs needed for DNN training are generated using a normal-grid based method. Experimental results show that the proposed DNN based correction yields more accurate ATBs in terms of Dynamic Time Warping (DTW) distance compared to the ATB segmentation approaches it is applied on. Thus, the DNN based correction could be used as a post-processing step to improve the accuracy of the predicted ATBs from any segmentation scheme.

© 2020 Elsevier Ltd. All rights reserved.

Keywords:
Real-time magnetic resonance imaging video
Air tissue boundary segmentation
Deep neural network
Error correction

1. Introduction

The real-time Magnetic Resonance Imaging (rtMRI) is used to capture the vocal tract dynamics in the mid-sagittal plane during speech (Bresch et al., 2008). In particular, an rtMRI video provides a complete view of the vocal tract including pharyngeal structures in a non-invasive manner (Bresch et al., 2008). As it is safe and free of health hazards, it is often used as a pre-eminent tool for speech production research. There have been attempts to capture vocal tract movement using other methods such as Electromagnetic articulography (Maurer et al., 1993), ultrasound (Watkin and Rubin, 1989), and X-ray (Wold, 1985). However, compared to other methods, the rtMRI has the particular advantage of capturing the entire vocal tract in the mid-sagittal plane. An rtMRI video provides the spatiotemporal information of the movements of speech articulators including lips, lower and upper incisors, jaw, palate, soft palate, velum, tongue, epiglottis which are useful for speech production modeling. The tissues

*Corresponding author.
E-mail address: mannemrenuka@iisc.ac.in (R. Mannem).

https://doi.org/10.1016/j.csl.2020.101160
0885-2308/© 2020 Elsevier Ltd. All rights reserved.
corresponding to these speech articulators and pharyngeal wall in an rtMRI video frame have a contrastive gray value compared to the air around them. A common pre-processing step for using an rtMRI video is to perform Air Tissue Boundary (ATB) segmentation to identify different speech articulators in an rtMRI frame. Hence, it is important to have an accurate ATB segmentation in the rtMRI video.

1.1. Applications using ATBs

ATB segmentation in an rtMRI video has been used for a number of speech applications in the past. For example, (Toutios et al., 2016) used the predicted ATBs in the rtMRI video to develop a text-to-speech synthesis system. An accurate ATB segmentation is also needed to study the articulators and dynamics of the vocal tract (Parrell and Narayananan, 2014; Hsieh et al., 2013; Prasad et al., 2015; Li et al., 2016). ATBs from the rtMRI video have been used for automatic visual augmentation for spoken language training (S. et al., 2018). The ATBs have also been used in the estimation of the variant and invariant characteristics in speech articulation for speaker identification (Prasad et al., 2015). In Mannem et al. (2019), the ATBs have been used as articulatory representations to estimate the speech rate. In Lammert et al. (2013) and (Ramanarayanan et al., 2013), the ATB segmentation is used as a preprocessing step to analyze the vocal tract’s morphological structure and movement. Ashok et al. (Pattem et al., 2018) employed ATBs for determining optimal sensor placement in electromagnetic articulography recording, which is useful in the speech production analysis. The predicted ATB is utilized to study the time evolution of the vocal tract cross-sectional area (Story et al., 1996) which is employed in most of the speech processing applications. Thus, an accurate ATB segmentation in the rtMRI video of the upper airway of the vocal tract is essential for many speech processing and speech production applications.

1.2. Related work for ATB prediction

Several unsupervised and supervised approaches have been used in the past for the ATB segmentation in an rtMRI video. Among the unsupervised approaches, Proctor et al. (2010) proposed a semi-automatic approach for ATB segmentation in which the ATBs are typically located in the vicinity of the steepest change in the pixel intensities on a grid line. Lammert et al. (2010) proposed a data-driven approach where the intensity variation in the correlated regions was considered for ATB prediction. Toutios and Narayananan (2015) used a factor analysis approach to derive the vocal tract outlines. The ATB segmentation has also been done using a spatial frequency domain representation of the rtMRI image (Bresch and Narayanan, 2009). Zhang et al. (2016) used multi-directional Sobel operators and constructed a boundary intensity map to extract the tongue contour. Kim et al. (2014) proposed a composite analysis grid line approach using an enhanced rtMRI frame including pixel intensity correction and grainy noise reduction. This Maeda Grid (MG) based technique (Kim et al., 2014) is advantageous over other unsupervised and semi-automatic approaches due to the image enhancement and grid-based analysis. However, more accurate ATBs are predicted using supervised methods (Koparkar and Ghosh, 2018; CA et al., 2018; Valliappan et al., 2019; Somandepalli et al., 2017). For example, Somandepalli et al. (2017) used a convolutional neural network based semantic segmentation for ATB prediction. Koparkar and Ghosh (2018) proposed a Fisher Discriminant Measure (FDM) based approach utilizing dynamic programming to ensure the temporal smoothness in the ATBs across the rtMRI video frames. The FDM method performs better than the MG baseline approach due to its novel contrast measure and temporal smoothness criterion. Given a new test rtMRI image, an ATB is predicted as a combination of ATBs from the training set which maximizes the FDM based objective function. Due to this, the predicted ATB is not smooth and its dynamics are limited by ATBs from the training set. Avoiding these limitations, Valliappan et al. proposed a method of ATB prediction based on the segmented rtMRI images (consisting of air and tissue classes) using the Fully Convolutional Neural network (FCN) (CA et al., 2018) and a deep convolutional encoder-decoder network (SegNet) (Valliappan et al., 2019). The encoder part of the SegNet follows VGG16 architecture (Simonyan and Zisserman, 2014). The SegNet and FCN approaches provide smooth and reliable boundaries compared to the FDM approach. The SegNet approach results in boundaries more accurate than FCN due to a high pixel classification accuracy provided by SegNet for semantic segmentation of rtMRI images. In FCN and SegNet based approaches, misclassification of the pixels in some regions like epiglottis and velum (due to low resolution and low Signal-to-Noise Ratio (SNR) of the rtMRI images) leads to erroneous ATBs. Likewise, every ATB segmentation approach, be it knowledge-based or data-driven, has its own limitations in ATB prediction.

1.3. Need of ATB correction

Each prediction algorithm has its limitation due to model assumption or data quality, which, in turn, limits its accuracy in predicting the ATBs. Fig. 1 illustrates the ground truth ATBs and the nature of the predicted ATBs from MG, FDM and SegNet approaches for a typical rtMRI image. In particular, the blue box in each sub-figure highlights regions with significant errors. Comparing the ground truth and predicted ATBs, it is observed that the error occurs in different regions in MG, FDM, and SegNet approaches. In the MG approach, the error is observed near the velum region and in the FDM approach, it is near the hard palate region. In the SegNet approach, the error occurs in both upper and lower ATBs near the hard palate and tongue base regions, respectively. The predicted ATBs using the SegNet approach are smooth and reliable compared to those using the MG and FDM approaches. From these observations, it can be stated that the nature and characteristics of ATB prediction error vary across algorithms. The error patterns in the predicted ATBs by an algorithm could be corrected in a data-driven manner as a post-processing step. Thus, an ATB correction approach could further improve the quality of the predicted ATBs. In this paper, we propose a Deep
Neural Network (DNN) based correction scheme for an accurate ATB segmentation. For this purpose, we use a feed forward neural network with three layers as DNN.

1.4. Related work for error correction

Although there was no prior work on error correction for ATB segmentation in the literature, various error correction algorithms in other areas of signal processing exist. For example, Huang et al. (2018) proposed an error correction method for semantic segmentation of an image using convolutional neural networks. The correction approach predicts dense semantic labels using both the input image and an initial estimate of the segmented image. In a similar way, Gidaris and Komodakis (2017) proposed a deep structured prediction approach using the initial estimates and the input images for dense disparity estimation. An error detection and correction framework was proposed by Zung et al. for 3D reconstruction of neurons from electron microscopic images (Zung et al., 2017). Bassil and Alwani (2012) proposed an error correction algorithm to correct misrecognized words from a typical automatic speech recognition (ASR) system. Fusayasu et al. (2015) used confusion networks for the ASR word-error correction based on normalized relevance distance. Liu (2002) proposed a simplified self-correction algorithm to improve the accuracy of the approximate algorithms in digital signal processing and parameter estimation. Likewise, different error correction algorithms were used in the literature to improve the accuracy of the prediction algorithms.

1.5. Motivation for the proposed ATB correction approach

In this work, we propose a DNN based ATB correction approach which uses a normal-grid based feature extraction method which is a contribution of our work. In an rtMRI image, the pixel intensity from tissue to airway often drops by a large margin reflecting the ATB. Although the goal of any ATB segmentation technique is to predict such changes in the pixel intensity, it sometimes fails to predict accurate ATBs due to low SNR of the rtMRI images. In the literature, the dynamic time warped (DTW) distances between the predicted and ground truth ATBs lie in the range of 2.5 – 3.5 mm (Kim et al., 2014; Koparkar and Ghosh, 2018; CA et al., 2018; Valliappan et al., 2019). From Fig. 1, it is observed that the predicted ATBs, although erroneous, remain in the vicinity of the ground truth ATBs. Based on this observation, we hypothesize that correction of each point on a predicted ATB can be done using the pattern of intensity variation in the direction of normal at that point. Fig. 2(a) and (b) illustrate the normal...
drawn at the $j^{th}$ predicted point ($x_j, y_j$). From Fig. 2(c), it is observed that ($x_j, y_j$), indicated by blue dashed line, lies in the constant intensity region whereas ($x_g_j, y_g_j$), indicated by green dashed line, lies in a region where there is large change in the intensity profile. Thus, the slope of the intensity profile at the ground truth ATB point is higher than that at the predicted ATB point. Hence, the goal of the DNN based correction is to trace the point of the maximal change in the pixel intensity profile on the normal to a predicted contour. To achieve this, the inputs and target outputs required for the DNN training are generated using features extracted from a normal-grid based method. The input feature is a set of intensities (grid), representing an intensity profile, in the direction of normal at a predicted point and the corresponding target output is the normalized distance which indicates the location of the ground truth point (obtained as the intersection point between the ground truth ATB and the normal) with respect to the input grid.

Fig. 2. Illustration of (a) ground truth ($C_{gr}$) and predicted lower ATBs ($C$) and normal ($N_j$) drawn at the $j^{th}$ point ($j \approx 34$) of $C$ in an rtMRI image where the pixel intensity varies from 0 (white) to 1 (black) indicating the transition between airway cavity and tissue regions. (b) region indicated by black box in (a), and (c) intensity variation at each point on the normal as a function of distance where $d$ and $d_g$ are distances corresponding to ($x_j, y_j$), ($x_g_j, y_g_j$), respectively. ($x_j, y_j$) is the $j^{th}$ predicted point and ($x_g_j, y_g_j$) is the intersection point between $N_j$ and $C_{gr}$).

Fig. 3(a) and (b) illustrate the kernel density estimates (KDEs) of the slope of the intensity profiles for the upper and lower ATBs, respectively. KDEs are shown for both the ground truth as well as predicted boundaries. It is observed that the slope of the intensity profile is primarily positive for lower ATBs and is negative for upper ATBs. This is because, for the upper ATB, the pixel intensities on the normal vary from 0 (airway cavity) to 1 (tissue) and for the lower ATB, the pixel intensities on the normal vary from 0 (airway cavity) to 1 (tissue) which can be observed in Fig. 4(b). From Fig. 3, it is observed that the average slope of the intensity curve at the ground truth ATB is higher compared to that at the predicted ATBs for both upper and lower boundaries. Since the ground truth ATB lies in the region of high-intensity variation, it is assumed that the target output lies in the high-intensity variation region for a given input set of intensities. Hence, for a given test input set of intensities, the trained DNN is expected to generate the output in the vicinity of a region with high intensity variation. However, the ground truth ATB may not always lie in the region of high intensity variation. Fig. 5 illustrates the ground truth upper and lower complete ATBs for four subjects. The regions in the black boxes do not correspond to a region with high intensity variation when different articulators come in contact with each other to produce specific phonemes (e.g., tongue touches hard palate to produce /t/). The regions in the black boxes occur mainly due to constriction between 1) the hard palate and the tongue base 2) the velum and the pharyngeal wall 3) the glottal base and the tongue dorsum 4) the upper lip and the lower lip 5) the velum and the tongue base. During manual annotation, these regions are annotated based on the knowledge of the morphology of the vocal tract. Thus, the manually annotated ATBs may not always lie in the region of high intensity variation. Hence, for ATB correction, the DNN is expected to learn factors beyond high intensity variation in a data-driven manner from the knowledge of an annotator. In this paper, DNN corrects each point in a predicted ATB separately instead of correcting them jointly. Due to this, the corrected ATB becomes jagged which...
requires smoothing to obtain a realistic ATB. Thus, a local regression lines based smoothing technique (Tolga Birdal, 2020) is used to obtain realistic ATBs post correction.

1.6. How the objective function of correction differs from that of prediction

Although every ATB segmentation algorithm predicts the ATB using intensity patterns from a given input image, they use different objective functions which are optimized for accurate ATB prediction. For example, in the MG approach, the ATBs are predicted using an optimal airway path by minimizing a score which considers the first order derivatives of the pixel intensities in the region of interest with a temporal continuity constraint. In the FDM approach, the ATB for a test rtMRI image is chosen from the set of train ATBs which maximizes an objective function that considers the contrast (which is a function of the test image’s intensities) around an ATB and a temporal continuity constraint. In the SegNet approach, the ATB is predicted based on a semantically segmented image using an edge detection algorithm. For semantic segmentation, the binary cross entropy loss function is minimized such that the pixels in the tissue and airway cavity regions are classified as 1 and 0, respectively. Likewise, the MG, FDM and SegNet approaches use different objective functions based on intensity patterns from a given image. However, in the proposed DNN based approach, to correct a predicted ATB point, we consider the pattern of intensity variation in the direction of normal at that point without considering the intensity profile of the entire image. During DNN training, the distance between the corrected and ground truth ATB points is minimized with respect to the given intensity profile. Thus, the DNN based correction performs ATB segmentation directly from a given set of intensity profiles along the normals on the predicted ATB points unlike the entire rtMRI image in the case of MG, FDM, and SegNet approaches. The complexity (number of layers and nodes) of the DNN is chosen based on the number of training features which depends on the length of (i.e., number of points on) the predicted ATBs from a typical ATB segmentation approach. The performance of the proposed approach is evaluated using the DTW distance between the corrected and ground truth ATBs. The DNN based correction is done for two types of predicted ATBs: 1) Complete ATBs (as shown in Fig. 5), 2) ATBs within the vocal tract (as shown in Fig. 4). In this paper, ATBs predicted using SegNet, FDM, and MG approaches are corrected using DNN.

Experimental results show that the DNN based correction yields more accurate ATBs in terms of the DTW distance compared to those from the ATB segmentation scheme it is applied on. The DNN based correction could be used as a post-processing step to improve the accuracy of the predicted ATBs from any segmentation scheme. The KDEs of the slope of the intensity curve at each point on the predicted and corrected ATBs are obtained for the test data. It is observed that the mean of the slope corresponding
to the corrected ATB is more than that of the absolute value of the slope corresponding to the predicted ATB. This change in the mean of the slope indicates that the intensity variation around the ATB points post-correction is, on average, higher compared to the intensity variation around the predicted ATB points.

2. Dataset

In this work, USC-TIMIT corpus (Narayanan et al., 2014) is used which is a rich source of rtMRI videos of the upper airway in the midsagittal plane. The database contains videos of five female and five male subjects each speaking 460 sentences from MOCHA-TIMIT (Wrench, 2000) database. The rtMRI video is recorded at 23.18 frames/sec. Each frame of the rtMRI video has a spatial resolution of 68 × 68 pixels and each pixel has a dimension of 2.9 mm × 2.9 mm. In this work, we consider 16 videos corresponding to 16 sentences from each of the four subjects, namely, F1, F2, M1, and M2. The 16 videos are chosen due to their highest phonetic richness which is explained in the work by Pattem et al. (2018). The chosen 16 videos have a total of 1462, 1270, 1399 frames from two female (F1 and F2) and two male (M1 and M2) subjects, respectively. For ATB prediction, we use a four fold cross validation setup. In each fold, eight training, four development, and four test videos are used in a round-robin fashion from each subject. For the DNN based correction, the predicted ATBs from the 75% and 25% of the development data (corresponding to ATB prediction) are used as training and development data, respectively. The test set is same for both ATB prediction and ATB correction.

The manual annotation of the complete ATBs is done using a MATLAB based GUI. Along with the ATBs, upper lip (UL), lower lip (LL), tongue base (TB), velum tip (VEL) and glottis begin (GLTB) are also annotated. The procedure of the manual annotation is explained in the work by Pattem et al. (2018). Fig. 6 illustrates the three contours representing the complete ATBs (C1, C2, and C3) and the five points (UL, LL, TB, VEL, GLTB) for a typical rtMRI frame. The contours C1, C2, and C3, on average, have lengths of 3.50 cm, 4.55 cm, 1.47 cm with 55, 64, 15 points on them respectively across all the frames. As shown in Fig. 6, the contour C1 is a closed contour which starts from UL, traces through the hard palate and joins the VEL and goes around the fixed nasal tract. The contour C2 is also a closed contour which traces the jawline, LL, tongue blade and extends below the epiglottis. The contour C3 marks the pharyngeal wall. Fig. 7 illustrates the fixed (black) and movable parts (red, blue, green) in C1, C2, and C3 ATBs separately for F1, F2, M1, and M2 subjects. From Fig. 7, it is observed that the fixed and movable ATB parts vary across subjects which

![Fig. 5. Illustration of rtMRI images (for four subjects) with ground truth upper (red) and lower (green) ATBs and the regions (black box) where the ground truth ATBs do not correspond to a high intensity variation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)](image-url)
depends on the vocal tract morphology of a subject. The fixed parts for each subject are annotated once and used in all frames. Thus, for every frame, the annotation is done only for the movable parts. Fig. 8 illustrates the blurry regions (velum, tongue tip, epiglottis and glottis) in the rtMRI images. From the regions indicated by red circles, it can be observed that the change in gray values from the tissue to the airway cavity is not prominent, which, in turn, creates ambiguity in ATB annotation. Thus, in this case, the ATBs are annotated based on the knowledge of the morphology of the subject. The manual annotations are verified and corrected in multiple steps.

3. Methodology

The steps followed in the proposed DNN based ATB correction scheme are illustrated in Fig. 9. Given a test rtMRI image and the corresponding predicted ATB from a typical ATB segmentation approach, the input test features are generated using the normal-grid based feature extraction method using the optimum parameters learnt during training. The generated input test grids are given as input to the trained DNN model for correction. The outputs of the DNN model are processed to obtain the coordinate points on the corrected ATB. Then the corrected ATB is smoothed using the local regression lines based smoothing technique.
3.1. Feature extraction

The ATBs, in an rtMRI image, correspond to the boundaries that typically have high pixel intensity variation across them. Therefore, in the proposed approach, the input features of DNN are generated by using the pixel intensities around the given predicted ATB. The feature extraction is done in two steps: (1) Normal Generation (2) Grid Generation. These two steps are explained in detail below.

3.1.1. Normal generation

A predicted contour in an rtMRI image (I) consisting of \( K \) points is defined as: \( C = \{ (x_j, y_j), 1 \leq j \leq K \} \), where \( x_j \) and \( y_j \) denote the X and Y coordinates of the contour. The midpoint of the normal \( N \) is identical to the predicted contour point \( (x_j, y_j) \). The normal \( (\text{having } 2 \times P \text{ pixel length with } 2N+1 \text{ number of equally spaced points}) \) corresponding to the \( f^{th} \) point on the given predicted contour \( C \) is defined as:

\[
N_j(x_j, y_j), (x_j+1, y_j), (x_j+1, y_j+1), P, N) \end{eqnarray}
\)

\[\begin{aligned}
N^x_j & = \{ x_j + \Delta x \cdot l, \quad -N \leq l \leq N \}, \\
N^y_j & = \{ y_j + m_j(p - x_j), \quad l_k \in N^x_j, \quad m_j = \frac{y_{j+1} - y_j}{x_{j+1} - x_j}, \quad m_{j+1} = \frac{y_{j+1} - y_j}{x_{j+1} - x_j}, \quad \Delta x = \frac{P}{2N} \}
\end{aligned}\]

The equation of the normal is defined as:

\[f_j(z) = y_j + m_j(z - x_j)\]

where \( N^x_j \) and \( N^y_j \) are the starting and ending points on the normal. The normal is generated at the predicted contour point \( (x_j, y_j) \) using the following and previous neighbouring points \( ((x_{j+1}, y_{j+1}), (x_j+1, y_j), (x_j, y_{j+1})) \) on the given predicted contour. The midpoint of the normal \( (N^x_j(N+1), N^y_j(N+1)) \) is identical to the predicted contour point \( (x_j, y_j) \). For extreme points on the contour \( C \) i.e., \( j = 1 \) and \( j = K \), the normal is generated as \( N_1((x_1, y_1), (x_1, y_1), (x_2, y_2), P, N) \) and \( N_2((x_{K-1}, y_{K-1}), (x_K, y_K), (x_K, y_K), (x_K, y_{K+1}), P, N) \), respectively. Let \( \mathcal{I}N_j = \{ \mathcal{I}N_j(k), 1 \leq k \leq 2N+1 \} \) where \( \mathcal{I}N_j(k) \) is the bilinearly interpolated intensity value at the point \( (N^x_j(k), N^y_j(k)) \) on the normal \( N_j \).

Thus, \( \mathcal{I}N_j \) is a set of intensity values on the normal \( N_j \) and it depends on \( I \) and \( N_j \).
Likewise, $K$ normals corresponding to $K$ points on the predicted contour $\mathcal{C}$ are generated for the given rtMRI image (I). Fig. 10(b) illustrates the generated normals at the points on a predicted contour $\mathcal{C}$ in the region marked with $B$.

### 3.1.2. Grid generation

For the training data, the grids are generated based on the intersection point between the normal and the ground truth contour. The intensity profile around the ground truth ATB varies depending on the articulators and the articulatory configuration based on what sound is being produced. Thus, an intensity profile around the ground truth ATB would be needed to learn the complex relationship between the ATB location and intensity profile. For this, multiple grids are generated from a normal corresponding to the predicted and ground truth ATB points. Each grid from the normal has a set of intensities ensuring that the ground truth point lies within the corresponding set of points on the grid. Likewise, multiple grids are generated from a normal to increase the training data size so that the trained DNN learns better mapping between the input set of intensities and the target output, the ground truth point.

Given a normal $\mathcal{N}_j$ (defined as in Eq. 1) and corresponding $\mathcal{I}\mathcal{N}_j$ at the $j^{th}$ point on a predicted train contour $\mathcal{C}_{tr}$, the grids are generated based on the intersection point $(x^j, y^j)$ between the normal $(\mathcal{N}_j)$ and the corresponding ground truth contour $\mathcal{C}_{gr}$. Let the point on the normal $\mathcal{N}_j$ which is nearest to $(x^j, y^j)$ is denoted as $(\mathcal{N}_j^x(x^j), \mathcal{N}_j^y(y^j))$ where:

$$r_j^* = \arg \min_{1 \leq r \leq 2N+1} \| (x^j, y^j) - (\mathcal{N}_j^x(r), \mathcal{N}_j^y(r)) \| \quad \text{where} \quad \| \| \text{indicates the L2 norm}. \quad (3)$$

Based on the value of $r_j^*$, the possible combinations of $N+1$ successive points are considered such that each combination includes $(\mathcal{N}_j^x(r_j^*), \mathcal{N}_j^y(r_j^*))$. Likewise, for each possible combination from the given normal $\mathcal{N}_j$, the corresponding intensity values are considered from $\mathcal{I}\mathcal{N}_j$. Each combination of the intensities from $\mathcal{I}\mathcal{N}_j$ is referred to as a grid. Let $G_j$ be the set of grids corresponding to the normal $\mathcal{N}_j$ such that the cardinality of the set $|G_j|$ depends on the value of $r_j^*$. 

Fig. 9. Illustration of the steps in the proposed DNN based ATB correction scheme.

Fig. 10. Illustration of (a) a predicted contour (C) in an rtMRI image and (b) the normals (with $P = 3.71$ and $N = 12$) to the predicted contour $\mathcal{C}$ in the region marked with $B$. 

Increases, the number of grids reduces since the intersection point which is the highest possible number of grids on a normal. As the distance between the predicted and ground truth point coordinate points (on the normal \(G_j\)) lies within the starting and ending points of the normal. From Eq. 4, it is observed that the possible number of grids varies for a given normal based on the value of \(r_j\). Each element of the set \(G_j\) is referred as a grid which is a \(N+1\) dimensional vector of intensities. \(C_{tr}\) is the midpoint of the normal \(N_j\) such that \(\text{sgn} \left( \left( x_j^r - N_j^x(1) \right) \times \left( x_j^r - N_j^x(2N+1) \right) \right) < 0^1\) otherwise \(|G_j|\) is considered as zero. In the latter case, the normal \(N_j\) corresponding to a predicted point on the contour \(C_{tr}\) does not contribute to grid generation since the intersection point between the normal and ground truth contour does not lie within the starting and ending points of the normal.

Fig. 11 illustrates the grid generation for different values of \(r_j\). In Fig. 11, the green lines indicate the set of points considered on the normal for each grid. Fig. 11(b), (c), (d) illustrate the grid generation corresponding to the regions B1, B2 and B3, shown in Fig. 11(a). In B1 and B2 regions, the intersection point \((x_j^r, y_j^r)\) lies above \(\| N_j^x(1) - x_j^r, y_j^r \| < P\) and below \(\| N_j^x(1), N_j^x(1) - [x_j^r, y_j^r] \| > P\) the midpoint of the normal \(N_j^x(N+1), N_j^x(N+1)\), respectively. In B1 region, \(r_j^r = 4\). Hence, \(|G_j| = 4\) from Eq. 4. Thus, in Fig. 11(b), we observe 4 grids. In B2 region, \(r_j^r = 12\), then \(|G_j| = 6\) from Eq. 4. Thus, in Fig. 11(c), we observe 6 grids. It is observed that the value \(\| N_j^x(1), N_j^x(1) - [x_j^r, y_j^r] \| \) (which indicates the distance between the predicted and ground truth points) is more for B1 region compared to B2 region. Thus, the \(|G_j|\) value is less for B1 region compared to B2 region. In B3 region, \(r_j^r = N+1\). which indicates that the predicted and ground truth points coincide (Fig. 11(d)). In this case, \(|G_j| = N+1 = 9\) which is the highest possible number of grids on a normal. As the distance between the predicted and ground truth point increases, the number of grids reduces since the intersection point \((x_j^r, y_j^r)\) moves away from the midpoint of the normal (refer Fig. 11(b) and (c)).

The DNN based correction is done in two different methods (denoted as DNN\(^d\) and DNN\(^i\)) based on the target output used for the given input grid: 1) DNN\(^d\) uses the target output as the distance between the predicted and the ground truth points with respect to the input grid and 2) DNN\(^i\) uses the target output as the index of the ground truth point with respect to the input grid.

From Eq. 4, the \(r_j^r\) element of the \(G_j\) is defined as: \(G_j(r) = \lfloor r N_j^x(k) \rfloor. r + N_i \leq k \leq r + N_1 + N\) where \(N_i\) depends on \(r_j^r\). \(G_j(r)\) is the \(N+1\) dimensional input grid to the DNN. Consider \(G_j(r) = \lfloor N_j^x(k). r + N_1 \leq k \leq r + N_i + N\\) which is a vector with elements as coordinate points (on the normal \(N_j\) corresponding to the intensities of the grid \(G_j\)). In DNN\(^d\) and DNN\(^i\), the target output \(d_{tr}\) for the input grid \(G_j(r)\) is generated as follows:

1) DNN\(^d\): Distance between the intersection point \((x_j^r, y_j^r)\) and the starting point of \(C_j(\bar{r})\) divided by the total pixel length of \(C_j(\bar{r})\).

\[
d_{kr} = \frac{\| x_j^r, y_j^r - N_j^x(r+N_1), N_j^x(r+N_1) \|}{\eta_d} \quad \text{where} \quad \eta_d = P.
\]  

2) DNN\(^i\): Index of the intersection point \((r_j^r)\) divided by total number of points on the grid \((N+1)\).

\[
d_{ki} = \frac{r_j^r - (r + N_1) + 1}{\eta_i} \quad \text{where} \quad \eta_i = N+1
\]

\[1\]

\[\text{sgn}(x) = \begin{cases} +1 & \text{if} \quad x > 0 \\ -1 & \text{if} \quad x < 0 \end{cases}\]
From Eq. 5 and 6, it is observed that the target outputs in DNN\textsuperscript{d} and DNN\textsuperscript{i} schemes are normalized to lie in the range 0 to 1 using the normalization factors \(\eta_\text{d}\) and \(\eta_\text{i}\), respectively. \(\eta_\text{d}\) and \(\eta_\text{i}\) are constants for all the grids and they depend only on \(P\) and \(N\). Due to the normalization, the target output values range from 0 to 1. Thus, the DNN is expected to obtain the outputs within this range with respect to the input grid.

The ground truth point is not available for a test rtMRI image. Hence, for a given test rtMRI image with the corresponding predicted test contour, the set of test grids is generated in a manner different from that of a training rtMRI image as follows. This difference in the grid generation for the train and test data does not affect the correction because the train and test grids are generated as the set of intensities corresponding to the points on the normal obtained using the predicted ATB points. Given a normal \(N_j\) (defined as in Eq. 1) corresponding to the \(j\)th point on a predicted test contour \(c_{\text{le}}\), a grid is generated using the set of \(N_j\) values corresponding to the center points on the normal. Given a normal \(N_j\) and the corresponding \(N_j\) at the \(j\)th point on \(c_{\text{le}}\), the center grid \((g_{c})\) is defined as follows:

\[
g_{c} = \left[ \frac{1}{2}, N \leq k < N - \frac{1}{2} \right] \quad (N \text{ is assumed to be even})
\]  

(7)

Let \(a_\text{d}\) and \(o_\text{i}\) be the offset values defined for DNN\textsuperscript{d} and DNN\textsuperscript{i} approaches respectively as follows:

\[
a_\text{d} = \frac{P}{2} \quad \text{and} \quad o_\text{i} = \frac{N}{2}
\]  

(8)

The offset values for DNN\textsuperscript{d} and DNN\textsuperscript{i} are kept fixed for all the test grids and they depend only on \(P\) and \(N\). The offset values \((a_\text{d}, o_\text{i})\) and the normalization factors \((\eta_\text{d}, \eta_\text{i})\) are used to generate the ATB points from the DNN output. Likewise, DNN\textsuperscript{d} and DNN\textsuperscript{i} methods are used to analyse the performance of the DNN based correction when the distance and index of the ground truth ATB point with respect to the input grid are considered as the target outputs, respectively.

### 3.2. DNN based correction

The DNN based correction technique can be applied on a predicted ATB obtained from any ATB segmentation approach. In this work, the DNN based correction is done on the output of three segmentation approaches, namely, MG, FDM and SegNet which show diversity in their prediction strategies (as explained in Section 1).

In the literature, two types of ATBs are predicted from a typical ATB segmentation approach: 1) ATBs within the vocal tract 2) Complete ATBs which trace regions both inside and outside of the vocal tract. To evaluate the predicted complete upper ATB, \(C_1\) contour is used as ground truth which is obtained considering the non-fixed points from the manually annotated complete ATB \((C_1)\) as shown in Fig. 12(b). To evaluate the predicted complete lower ATB, manually annotated complete lower ATB \((C_2)\) is used as the ground truth as shown in Fig. 12(b). To evaluate the predicted ATBs within the vocal tract, the upper and lower ground truth complete ATBs \((C_1, C_2)\) are pruned to obtain the ATBs within the vocal tract using the contour pruning approach outlined in the work by Koparkar and Ghosh (2018). The pruned upper and lower ground truth ATBs within the vocal tract are represented by \(C_1^\text{pr}, C_2^\text{pr}\), respectively, and they are illustrated in Fig. 12(c).

The FDM and SegNet approaches predict ATBs within the vocal tract as well as complete ATBs whereas the MG approach predicts ATBs within the vocal tract only. The predicted complete upper and lower ATBs are denoted as \(C_1^\text{pr}\) and \(C_2^\text{pr}\), respectively. And the predicted upper and lower ATBs within the vocal tract are represented as \(C_1^\text{pr}\) and \(C_2^\text{pr}\), respectively. Fig. 13(a), (b) and (c) illustrate within vocal tract ATBs predicted using MG, FDM, and SegNet, respectively. Fig. 14(a) and (b) illustrate the complete ATBs predicted using FDM and SegNet, respectively. The DNN based correction is done separately for \(C_1^\text{pr}, C_2^\text{pr}\), and \(C_1^\text{pr}\).

The MG and FDM methods use a preprocessing technique to enhance the given rtMRI images (Kim et al., 2014). Thus, the preprocessed rtMRI images are considered to generate the grids, that are used as inputs for the DNN based correction of predicted ATBs from MG and FDM. On the other hand, the original rtMRI images without any preprocessing are used for DNN based correction of ATBs predicted by SegNet. The complexity of DNN is decided based on the training grid data size, i.e., the number of grids available for training. The complexity of the network is determined by the number of hidden layers and nodes in each layer of.
the DNN. From Section 3.1, it is observed that the training grid data size depends on $P$, $N$ and number of points in the predicted contours $Cv_1$, $Cv_2$, $Cf_1$ and $Cf_2$. We use a three layer feed forward neural network as DNN, where the number of nodes in each layer is decided based on the training data size as described in Table 1. A dropout of 0.15 and ReLu activation function is used after every dense layer. Mean squared error loss and Adam optimizer are used for the DNN training. The DNN is trained for a maximum of 100 epochs with early stopping criterion based on the validation data loss.

3.3. Finding optimum $P$ and $N$ values

Given the rtMRI images and the corresponding predicted ATBs from a typical ATB segmentation approach, the optimum values of $P$, $N$ are decided based on the performance on the validation data. The following steps are used to obtain the optimum $P$ and $N$ values:

1) The normals are generated (as explained in Section 3.1.1) with $P = 5$ and $N = 14$ for the given rtMRI images and predicted ATBs in the training data. From KDEs of the distance between the upper and the lower ATBs computed on the normal at all the points on the predicted ATBs shown in Fig. 15, it is observed that for 98% of the images the distance between the upper and the lower vocal tract boundaries does not cross 10 pixels. Thus, the normals are generated with a length of $2 \times P = 10$ pixels for all four subjects. This ensures that the distance between the upper and the lower vocal tract boundaries is mostly covered by the normal on a point on ATBs.

2) Consider a normal $N_j$ (defined as in Eq. 1) at the $j$th point on a predicted train contour $C_{tr}$ with corresponding ground truth contour $C_{gr}$. Let $(x'_j, y'_j)$ be the intersection point between $N_j$ and $C_{gr}$. Consider $d_j = ||(x'_j, y'_j) - (\frac{P}{2}, (N+1))||$ which is defined as the distance (in units of pixel) between the predicted point (midpoint of the normal) and the intersection point between the normal and the corresponding ground truth contour. Likewise, for each point on all the training contours, the corresponding $d_j$ values are calculated. Let $\mathcal{D}$ be the set of all such $d_j$ values.

<table>
<thead>
<tr>
<th>Training grid data size range (in millions)</th>
<th>$T \leq 0.05$</th>
<th>$0.05 &lt; T \leq 0.1$</th>
<th>$0.1 &lt; T \leq 0.2$</th>
<th>$0.2 &lt; T \leq 0.3$</th>
<th>$0.3 &lt; T \leq 0.4$</th>
<th>$0.4 &lt; T \leq 0.5$</th>
<th>$0.5 &lt; T \leq 0.6$</th>
<th>$T &gt; 0.6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of nodes</td>
<td>64, 64, 64</td>
<td>128, 128, 128</td>
<td>256, 128, 128</td>
<td>256, 256, 128</td>
<td>256, 256, 256</td>
<td>512, 256, 256</td>
<td>512, 512, 256</td>
<td>512, 512, 512</td>
</tr>
</tbody>
</table>

Fig. 13. Illustration of ATBs within vocal tract ($Cv_1$, $Cv_2$) from (a) MG (b) FDM (c) SegNet approaches.

Fig. 14. Illustration of complete ATBs ($Cf_1$, $Cf_2$) from (a) FDM (b) SegNet approaches.

Table 1

Number of nodes in three layers of DNN corresponding to the training grid data size ($T$).
3) Consider the set \( R = \{x, 70 \leq x < 99\} \cup \{99.25, 99.5, 99.75, 100\} \) which consists of the percentile ranks corresponding to the set \( D \). For each percentile rank from \( R \), the corresponding quantile value of \( D \) is obtained following the method described in Dunning and Ertl (2019). Let \( P_D \) be the set of all the quantiles corresponding to the percentile ranks in \( R \). Consider each value from \( P_D \) as \( P \) (with \( N = F(3 \times P^2) \)) and using which training grid features (denoted as \( Trg \)) are generated following Eq. 4.

4) In a similar way, the validation grid features are also generated based on the value of \( P \). But, the validation grid features are generated in two ways: i) Following Eq. 4 to generate all the possible grids ii) Following Eq. 7 to generate the center grids. The sets of validation grid features generated using i and ii are denoted as \( V_g \) and \( V_c \), respectively.

5) The corresponding target values for \( Trg, V_g \) and \( V_c \) are generated separately for the two DNN correction strategies \( DNN_d \) and \( DNN_i \) as explained in Section 3.1.2.

6) Consider \( DNN_d^R \) and \( DNN_i^R \) which are defined as follows:

\[
DNN_d^R = \{DNN_d^P : P \in P_D\} \quad \text{and} \quad DNN_i^R = \{DNN_i^P : P \in P_D\}
\]

where \( DNN_d^P \) and \( DNN_i^P \) are the \( DNN_d \) and \( DNN_i \) based correction strategies, respectively, which are trained using \( Trg \) generated using \( P \) that is chosen from \( P_D \). Likewise, for each element of the set \( P_D \), the corresponding \( Trg, V_g \) and \( V_c \) are generated and DNNs with appropriate complexity (as summarized in Table 1) are trained and validated using these features as explained in Section 3.2.

7) Early stopping is done based on \( V_g \) data loss. The predicted ATBs of validation data are corrected using the output values of the trained DNN obtained for the given input \( V_g \) data.

The correction strategy based on the DNN outputs corresponding to the center grids is explained later in Sections 3.4 and 3.5.

8) For each DNN from the sets \( DNN_d^R \) and \( DNN_i^R \), the corrected ATBs of validation data are obtained and the average DTW distance between the corrected and ground truth ATBs is computed. Let \( DNN_d^* \) and \( DNN_i^* \) be the trained DNNs corresponding to the least average DTW distance from the sets \( DNN_d^R \) and \( DNN_i^R \), respectively. The \( P \) (and corresponding \( N \)) values corresponding to \( DNN_d^* \) and \( DNN_i^* \) are considered as the optimum and are denoted as \( P_d^* \) and \( P_i^* \) (corresponding \( N_d^* \) and \( N_i^* \)), respectively.

9) Given a new test rtMRI image and predicted ATB, the test grid feature is generated (based on Eq. 7) using the optimum parameters and the predicted ATBs are corrected using the trained models (\( DNN_d^* \) and \( DNN_i^* \)) which yield the best performance on the validation data.

Thus, based on the validation data performance, we find the optimum parameters for feature generation and the optimized DNN model whose model complexity is chosen based on the size of training grid features.

3.4. Transformation of DNN outputs to ATB points

Given a test rtMRI image and predicted ATB, the corrected ATB points from the DNN outputs are obtained in different ways for \( DNN_d \) and \( DNN_i \) as the DNNs in these two approaches are trained using different target outputs:

\( DNN_d^\): For the \( j^{th} \) point on the predicted ATB, the center grid \((g_{jc})\) is obtained using the normal \( N_j \) which is generated using \( P = P_d^* \) and \( N = N_d^* \). \( g_{jc} \) is given as an input to the \( DNN_d^* \) and the generated output is denoted as \( d_d^j \) which is the normalized distance of the \( j^{th} \)

\[ F(x) \text{ rounds } x \text{ to nearest even integer} \]
corrected ATB point with respect to the grid. \( \tilde{d}_i^d \) is denormalized and offset is added to obtain the corrected ATB point’s distance with respect to the normal which is denoted as \( \tilde{d}_i^d \). Hence, \( \tilde{d}_i^d = \tilde{d}_i^d \times \eta_d + o_d \) where \( \eta_d \) and \( o_d \) are defined in Eq. 5 and 8, respectively. From \( \tilde{d}_i^d \), the corrected ATB point’s coordinates \((\tilde{x}_i, \tilde{y}_i)\) are obtained as follows:

\[
(\tilde{x}_i, \tilde{y}_i) = \left( (N_x^d(1) + \tilde{d}_i^d \times \cos(\theta_i)), (N_y^d(1) + \tilde{d}_i^d \times \sin(\theta_i)) \right)
\]

where \( \theta_i = \arctan(m_i) \) and \((N_x^d(1), N_y^d(1))\). \( m_i \) are defined in Equation 2.

DNN:

For the \( j^{th} \) point on the predicted ATB, the center grid \((g_{jc})\) is obtained using the normal \( N_i \) which is generated using \( P = P^* \) and \( N = N^* \); \( g_{jc} \) is given as an input to DNN\(^*\) and the generated output is denoted as \( \hat{d}_i^1 \) which is the normalized index of the \( j^{th} \) corrected ATB point with respect to the grid. \( \hat{d}_i^1 \) is denormalized and offset is added to obtain the corrected ATB point’s index with respect to the normal which is denoted as \( \hat{d}_i^1 \). Hence, \( \hat{d}_i^1 = [\hat{d}_i^1 \times \eta_i] + o_i^1 \) where \( \eta_i \) and \( o_i \) are defined as in Eq. 6 and 8, respectively. The value \( \hat{d}_i^1 \) is an integer in the range of 1 to \((2N^* + 1)\) which indicates a point on the normal. Thus, the point \((N_x^d(1), N_y^d(1))\) is the corrected ATB point corresponding to the input predicted point \((N_x^d(N + 1), N_y^d(N + 1))\).

As explained in Section 3.1.2, for a given test image, the number of grids is equal to the number of points on the respective predicted test contour so that the trained DNN outputs \( \mathcal{L} \) values for the \( \mathcal{L} \) points on the predicted test contour. The \( \mathcal{L} \) outputs are further processed as explained above to obtain the corrected ATB from the input predicted ATB. The corrected ATBs corresponding to the predicted ATBs \( C_Y^1, C_Y^2, C_Y^1, \) and \( C_Y^2 \) are represented as \( \tilde{C}_Y^1, \tilde{C}_Y^2, \tilde{C}_Y^1, \) and \( \tilde{C}_Y^2 \), respectively. The corrected full ATBs \( \tilde{C}_2^1 \) and \( \tilde{C}_2^2 \) are pruned to obtain the ATBs within the vocal tract using the contour pruning method outlined in Koparkar and Ghosh (2018). These pruned contours from \( \tilde{C}_2^1 \) and \( \tilde{C}_2^2 \) are denoted as \( \tilde{C}_2^1 \) and \( \tilde{C}_2^2 \), respectively.

The DNN\(^i\) based schemes using the predicted contours from MG, FDM, SegNet approaches are represented as DNN\(^{iM}\) and DNN\(^{iF}\), respectively, whereas the DNN\(^i\) based schemes using the predicted contours from MG, FDM, SegNet approaches are represented as DNN\(^{iM}\), DNN\(^{iF}\), and DNN\(^{iS}\), respectively. Likewise, for each prediction approach (MG, FDM, and SegNet), and for each predicted ATB \((C_1^1, C_2^1, C_1^2\) and \( C_2^2 \)), the DNN based correction is done separately using the optimized DNN models DNN\(^{i*}\) and DNN\(^{i*}\).

3.5. Smoothing using local regression lines approach

The DNN does pointwise correction of the predicted contours in which each point on the predicted contour is corrected independently without considering the neighbouring points. Thus, the contours corrected by DNN are jagged. Smoothing is needed to obtain the reliable contours. In this work, smoothing is done using local regression lines which are constructed using the

\[ \lceil \cdot \rceil \text{: round to nearest integer} \]
weighted least squares approach. This is a typical smoothing technique which does smoothing of the contour without disturbing the shape of the contour. Thus, we have used the local regression lines based smoothing instead of using a moving average filter which over-smooths the contour. The corrected contours from DNN are smoothed by projecting the contour points onto the local regression lines. For each point on a given contour, the local regression line is obtained using the $2 \times Q + 1$ points considering Q neighbouring points on either side of the point. To avoid over-smoothing, Gaussian weighted least squares fit is used to construct the linear regression line. The implementation of the smoothing technique is provided in Tolga Birdal (2020). The smoothness of a contour depends on the value of Q. In this work, the value of Q is decided based on the performance on the validation data. The corrected ATBs $\hat{C}_1, \hat{C}_2, \hat{C}_3, \hat{C}_4, \hat{C}_5, \hat{C}_6$ after smoothing are represented by $\hat{C}_1, \hat{C}_2, \hat{C}_3, \hat{C}_4, \hat{C}_5, \hat{C}_6$ respectively.

The proposed DNN based correction can be treated as a post-processing technique on the predicted ATBs. However, the DNN based correction can also be used as a direct ATB segmentation algorithm without using predicted ATBs from a typical ATB segmentation approach. In this case, the DNN based correction is done on the average ATB which is obtained as the mean of the training ATBs. For each train rtMRI image, the average ATB is considered as the predicted ATB and DNN is trained using the grid features as explained in Sections 3.1.2 and 3.2. The average ground truth upper and lower ATBs within the vocal tract are denoted as $\overline{C}_1$ and $\overline{C}_2$, respectively. Fig. 16(a) and (b) illustrate $\overline{C}_1$ and $\overline{C}_2$ and the ground truth ATBs ($C_1^{\text{gt}}$ and $C_2^{\text{gt}}$) within the vocal tract for two train rtMRI images from F1 subject. It is observed that the average ATBs, although not accurate, lie in the vicinity of the ground truth ATBs as observed in the case of ATBs predicted from a typical ATB segmentation approach. Thus, the DNN based correction is also performed on the average ATBs. Although the average ATBs lie in the vicinity of the ground truth ATBs, they do not capture the vocal tract dynamics as good as the predicted ATBs since $C_1^{\text{gt}}$ and $C_2^{\text{gt}}$ are identical across all the rtMRI images. Thus, in this case, the corrected ATBs may not be as accurate as the corrected ATBs obtained using the DNN based correction on the predicted ATBs from a typical segmentation approach. The DNN based correction schemes using the average ATB are done following DNNd and DNNi which are denoted by DNNd and DNNi respectively. The optimized DNN model and the optimum parameters for feature extraction are obtained following steps explained in Section 3.3.

4. Experiments

4.1. Experimental setup

In this work, we consider MG, FDM, and SegNet as the baseline ATB prediction schemes. As the DNN based correction scheme works on ATBs from a ATB prediction approach, ATB prediction is first carried out using MG, FDM and SegNet schemes. While MG is an unsupervised approach, FDM and SegNet are not. Hence, in this work, for ATB prediction using the FDM and SegNet approaches, the experiments are done using 16 videos from the four subjects (F1, F2, M1, and M2) in a four-fold cross-validation setup. In each fold, eight training (Trp), four development (Devp), and four test (TeP) videos are used in a round-robin fashion from each subject. Thus, a total of 32 training, 16 development, and 16 test videos are used for ATB prediction. Each fold, on average, contains ~2900 training, ~1443 development, and test images from all the four subjects. The training and validation of the FDM and SegNet approaches are done as explained in Koparkar and Ghosh (2018) and Valliappan et al. (2019), respectively. For the DNN based correction, the predicted ATBs from the 75% and 25% of the Devp are used as the training (Trc) and the development (Devc) data, respectively. The Devc data is used for training and validation of the DNN since it is unseen to training similar to the test data. Thus, the predicted ATBs of Devc helps in better training of the DNN based correction on the test set. The predicted ATBs from TeP are used as the test data (TeP) for correction. However, the predicted ATBs from Devp and TeP are also obtained using the MG approach for DNN based correction. Likewise, the DNN based correction is done for the predicted ATBs from MG, FDM, and SegNet approaches in the four-fold cross-validation setup. In each fold, the DNN is trained separately for each subject. Each fold, on average, consists of ~270 training, ~90 development, and ~360 test images for DNN based correction.

Many image segmentation approaches using deep convolutional networks have been shown to benefit from the pre-trained weights (Russakovsky et al., 2015; Wu et al., 2019; Iglovikov and Shvets, 2018). In these approaches, the encoder network is typically initialized with the pre-trained weights which are learned using the large ImageNet object classification dataset (Russakovsky et al., 2015). Similar to these approaches (Russakovsky et al., 2015; Wu et al., 2019; Iglovikov and Shvets, 2018), the encoder network of the SegNet model can also be initialized with pre-trained weights which may improve the accuracy of the predicted ATBs. The SegNet based ATB prediction approach which uses the pre-trained weights is denoted as SegNet-pre. We perform the experiments using SegNet-pre approach similar to those of the SegNet based ATB prediction approach. Table 2 shows the approximate average grid data sizes of Trc, Devc, TeP using SegNet, FDM, MG predicted ATBs ($C_1^{\text{gt}}, C_2^{\text{gt}}, C_3^{\text{gt}}, C_4^{\text{gt}}$) for the DNN based correction across all the folds of four subjects. The training grid data sizes and the DNN architectures in Table 2 correspond to the optimum P, N which are chosen following steps explained in Section 3.2. The optimum P and N values in different folds of four subjects range from 0.900 to 4.995 and 2 to 14, respectively. The experimental code for the proposed DNN based approach can be found in the link: https://github.com/mannemrenuka/DNN-based-correction.

4.2. Evaluation metric

The proposed DNN based correction scheme is evaluated using the DTW distance which measures the alignment between the corrected and the ground truth contour (Berndt and Clifford, 1994). The DTW distance has a unit of pixel.
The value of \( D(C_p, C_q) \) is less if \( C_p \) and \( C_q \) have similar shapes and located close to each other. From Eq. 11 and 12, it is observed that the value of \( L \) depends on the lengths of the given contours \( L_p \) and \( L_q \).

In this work, two types of performance evaluations are done: (1) evaluation of corrected complete ATBs \( (\tilde{C}_1, \tilde{C}_2, \tilde{C}_1, \tilde{C}_2) \) using the ground truth complete ATBs \( (C_1^p, C_2^p, C_1^c, C_2^c) \) (2) evaluation of the corrected ATBs within the vocal tract \( (\tilde{C}_1, \tilde{C}_2, \tilde{C}_1, \tilde{C}_2) \) using the ground truth ATBs within the vocal tract \( (C_1^v, C_2^v, C_1^c, C_2^c) \).

5. Results and discussions

Table 3 shows the average (± standard deviation) of the DTW distances for the ATBs within the vocal tract across all the subjects. It is observed that the DNN based correction scheme improves the accuracy of ATB segmentation with and without smoothing compared to that using all the baselines. It is also observed that DNN\(_d\) and DNN\(_i\) show improvements over the average ATBs.
As explained in Section 3, the DNNdA and DNNiA approaches can be treated as ATB prediction approaches. Comparing DNNdA and DNNiA performance with those from SegNet, FDM, and MG prediction approaches, it is observed that DNNdA and DNNiA, on average, perform better than the MG approach for both the upper and the lower ATBs whereas they perform better than the FDM approach for the lower ATB only. Similar to the DNN based prediction approach, the MG approach also uses the pixel intensities in the region of interest. However, the MG approach considers only the first order derivatives (which indicate the high intensity variation) of the pixel intensities to trace the ATBs. Hence, considering only the first order derivatives does not provide accurate ATBs compared to DNNdA and DNNiA which use a better mapping function between the intensity profile and ATBs. Among all the DNN based correction schemes, DNNiA and DNNdA provide the highest and lowest percentage improvements of 6.19% and 2.68%, respectively, for the corrected upper ATB whereas DNNdA and DNNiA provide the highest and lowest percentage improvements of 26.17% and 7.09%, respectively, for the corrected lower ATB compared to the predicted ATBs. Table 4 shows the average (± standard deviation) of the DTW distances (in pixels) for complete ATBs obtained using different ATB prediction and correction schemes across all the subjects.

Table 4  
Average (± standard deviation) of DTW distances (in pixels) for the ATBs within vocal tract obtained from the complete ATBs obtained using different ATB prediction and correction schemes across all the subjects.

<table>
<thead>
<tr>
<th>ATBs Approach</th>
<th>Upper ATB within vocal tract obtained from complete ATB</th>
<th>Lower ATB within vocal tract obtained from complete ATB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cfi Cfi</td>
<td>1.00 ± 0.18</td>
<td>1.02 ± 0.22</td>
</tr>
<tr>
<td>SegNet</td>
<td>1.10 ± 0.20</td>
<td>1.12 ± 0.27</td>
</tr>
<tr>
<td>SegNet-pre</td>
<td>1.10 ± 0.20</td>
<td>1.12 ± 0.27</td>
</tr>
<tr>
<td>FDM</td>
<td>1.08 ± 0.20</td>
<td>1.14 ± 0.29</td>
</tr>
<tr>
<td>MG</td>
<td>1.13 ± 0.23</td>
<td>1.27 ± 0.36</td>
</tr>
<tr>
<td>DNNdA</td>
<td>0.93 ± 0.19</td>
<td>0.97 ± 0.22</td>
</tr>
<tr>
<td>DNNiA</td>
<td>1.03 ± 0.20</td>
<td>1.06 ± 0.28</td>
</tr>
<tr>
<td>DNNfA</td>
<td>0.92 ± 0.18</td>
<td>0.94 ± 0.23</td>
</tr>
<tr>
<td>DNNiA + smoothing</td>
<td>1.03 ± 0.19</td>
<td>1.06 ± 0.28</td>
</tr>
<tr>
<td>DNNfA + smoothing</td>
<td>0.93 ± 0.20</td>
<td>0.92 ± 0.20</td>
</tr>
<tr>
<td>DNNiA + smoothing</td>
<td>1.02 ± 0.20</td>
<td>1.05 ± 0.28</td>
</tr>
<tr>
<td>DNNfA + smoothing</td>
<td>0.92 ± 0.23</td>
<td></td>
</tr>
</tbody>
</table>

Table 5  
Average (± standard deviation) of DTW distances (in pixels) for complete ATBs obtained using different ATB prediction and correction schemes across all the subjects.

<table>
<thead>
<tr>
<th>ATBs Approach</th>
<th>Complete upper ATB</th>
<th>Complete lower ATB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cfi Cfi</td>
<td>0.96 ± 0.15</td>
<td>0.90 ± 0.16</td>
</tr>
<tr>
<td>SegNet</td>
<td>1.11 ± 0.21</td>
<td>0.94 ± 0.20</td>
</tr>
<tr>
<td>SegNet-pre</td>
<td>1.03 ± 0.18</td>
<td>0.94 ± 0.19</td>
</tr>
<tr>
<td>FDM</td>
<td>0.89 ± 0.15</td>
<td>0.83 ± 0.15</td>
</tr>
<tr>
<td>DNNdA</td>
<td>1.00 ± 0.17</td>
<td>0.89 ± 0.18</td>
</tr>
<tr>
<td>DNNiA</td>
<td>0.89 ± 0.14</td>
<td>0.82 ± 0.15</td>
</tr>
<tr>
<td>DNNfA</td>
<td>1.00 ± 0.18</td>
<td>0.89 ± 0.18</td>
</tr>
<tr>
<td>DNNiA + smoothing</td>
<td>0.88 ± 0.14</td>
<td>0.83 ± 0.15</td>
</tr>
<tr>
<td>DNNfA + smoothing</td>
<td>0.99 ± 0.17</td>
<td>0.88 ± 0.15</td>
</tr>
<tr>
<td>DNNiA + smoothing</td>
<td>0.88 ± 0.14</td>
<td>0.81 ± 0.15</td>
</tr>
<tr>
<td>DNNfA + smoothing</td>
<td>0.99 ± 0.17</td>
<td>0.88 ± 0.15</td>
</tr>
</tbody>
</table>

From Table 3, 4, and 5, it is observed that the SegNet-pre approach does not perform better than the SegNet based approach for ATB prediction unlike the works presented in Russakovskiy et al. (2015); Wu et al. (2019); Iglovikov and Shvets (2018) which have been shown to benefit from the pre-trained weights. The pre-trained weights are learned using ImageNet dataset which includes the natural images which are very different from the rMRI images used in this work. The rMRI images are gray scale images with low resolution and low SNR. On the other hand, images from ImageNet are mainly high resolution RGB images.
at high SNR. Thus, the pre-trained weights may not provide the features optimal for ATB segmentation in low resolution rtMRI images.

5.1. Performance analysis of DNN based correction

As discussed in Section 1.5, the DNN is expected to generate the output in the vicinity of a region with high intensity variation. Thus, the intensity variation at the corrected ATB point could be higher than that of predicted ATB point and closer to that of the ground truth ATB point. To provide a quantitative analysis for this, the KDEs of the slopes of the intensity curves at the ground truth, predicted, and corrected ATB points are obtained for the test upper and lower ATBs within the vocal tract.

Given a ground truth ATB, the normal for each point on the ground truth ATB is generated with \( \frac{\partial}{\partial x} \) where \( \partial / \partial x \) is the set of intensities on the normal which is defined in Section 3.1.1. Let \( \mathcal{N}_p \) be the normal corresponding to the corrected ATB point, which is obtained using the corrected ATBs from DNN approach, and \( \mathcal{N}_p \) be the normal for each point on the ground truth ATB.

Consider a third order polynomial \( f(t) = at^3 + bt^2 + ct + d \) where \( a, b, c, d \) are the coefficients, \( f'(t) = 3at^2 + 2bt + c \) is the first order derivative of the polynomial \( f(t) \). Let \( f_{jp}(t) \) be the polynomial \( f(t) \) where the coefficients \( a, b, c, d \) are estimated using the least squares fit of the data

\[
\{ (d_{jp}^1, \mathcal{N}_p), (d_{jp}^2, \mathcal{N}_p), (d_{jp}^3, \mathcal{N}_p), (d_{jp}^4, \mathcal{N}_p), (d_{jp}^5, \mathcal{N}_p), (d_{jp}^6, \mathcal{N}_p), (d_{jp}^7, \mathcal{N}_p) \}
\]

where \( \mathcal{N}_j \) is the set of intensities on the normal which is defined in Section 3.1.1. Let \( f_{jp}(t) \) be the polynomial \( f(t) \) where the coefficients \( a, b, c, d \) are estimated using the least squares fit of the data

\[
\{ (d_{jp}^1, \mathcal{N}_j), (d_{jp}^2, \mathcal{N}_j), (d_{jp}^3, \mathcal{N}_j), (d_{jp}^4, \mathcal{N}_j), (d_{jp}^5, \mathcal{N}_j), (d_{jp}^6, \mathcal{N}_j), (d_{jp}^7, \mathcal{N}_j) \}
\]

Likewise \( f_{jp}(t) \), \( f_{jp}(t) \), and \( f_{jp}(t) \) are obtained using two back and forth points of the predicted, corrected, and ground truth points, respectively. Let \( f_{jp}(t) \), \( f_{jp}(t) \), and \( f_{jp}(t) \) be the polynomials which are obtained as the first order differentiation of \( f_{jp}(t) \), \( f_{jp}(t) \), and \( f_{jp}(t) \), respectively. For the \( j \)th predicted ATB point, \( f_{jp}(d_{jp}^1), f_{jp}(d_{jp}^2), f_{jp}(d_{jp}^3) \), and \( f_{jp}(d_{jp}^4) \) are obtained which are the slopes of the intensity curve on the normal at the predicted, corrected and ground truth ATB points, respectively. Consider \( F_p, F_p, \) and \( F_p \) as the set of all such \( f_{jp}(N+1), f_{jp}(d_{jp}^1), f_{jp}(d_{jp}^2), f_{jp}(d_{jp}^3) \), and \( f_{jp}(d_{jp}^4) \) which are obtained for each point on the test ground truth ATBs and corresponding predicted and corrected ATBs, respectively. Let \( K_g \) be the KDE of the elements of \( F_g \). Let \( K_{mp}, K_{cp}, K_{ap} \) and \( K_{np} \) be the KDEs of the data \( F_p \) which is obtained using the predicted ATBs from MG, FDM, and SegNet approaches and average ATBs, respectively. Let \( K_{mp}, K_{cp}, K_{ap} \) and \( K_{np} \) be the KDEs of the data \( F_c \) which is obtained using the corrected ATBs from DNN and DNN approaches, respectively. The KDEs \( K_g \), \( K_{mp}, K_{cp}, K_{ap} \) and \( K_{np} \) represent the set of all such KDEs of the data.

\[ \text{Fig. 17. Illustration of kernel density estimates of the slopes of intensity curves for ground truth and predicted upper ATBs within vocal tract for F1, F2, M1 and M2 subjects.} \]
and $K_{ac}$ are generated for the lower and upper ATBs within vocal tract. Let $F_{ng}$ be a set of slopes of intensity curves at each point on the normals corresponding to the test ground truth ATBs. Thus, $F_{ng}$ unlike $F_g$ consists of slopes of the intensity curves at each point on the normal and not only at the ground truth point. Consider $K_{ng}$ as the KDE of $F_{ng}$ which is also obtained for both upper and lower test ground truth ATBs within vocal tract. In all these cases, the KDE is obtained using the normal kernel function with an optimal bandwidth which is estimated by following the method described in Bowman and Azzalini (1997).

Fig. 17 and 19 illustrate $K_{ng}$, $K_g$, $K_{mp}$, $K_{fp}$, $K_{sp}$, and $K_{ap}$ for upper and lower test ATBs within the vocal tract, respectively. Fig. 18 and 20 illustrate $K_{ng}$, $K_g$, $K_{mc}$, $K_{fc}$, $K_{sc}$, and $K_{ac}$ for upper and lower ATBs within the vocal tract, respectively. It is observed that the slope of the intensity curve is primarily positive for lower ATBs and is negative for upper ATBs (the average values are mentioned in the legend of the respective figure). $K_{ng}$ for both upper and lower ATBs is almost equally spread in positive and negative slope regions since the slope of the intensity curve is obtained at every point on the normal. From Fig. 17 and 19, it is observed that the absolute value of the slope of the intensity curve at the ground truth ATB is higher compared to that at the predicted ATBs for both upper and lower ATBs. From Fig. 17 and 18, it is observed that the means of $K_{mp}$, $K_{fp}$, $K_{sp}$, $K_{ap}$, $K_{mc}$, $K_{fc}$, $K_{sc}$, and $K_{ac}$ are more negative compared to the means of $K_{ng}$, $K_g$, $K_{mp}$, $K_{fp}$, $K_{sp}$, $K_{ap}$, respectively. From Fig. 19 and 20, it is observed that the means of $K_{mp}$, $K_{fp}$, $K_{sp}$, $K_{ap}$, $K_{mc}$, $K_{fc}$, $K_{sc}$, and $K_{ac}$ are more positive compared to the means of $K_{ng}$, $K_g$, $K_{mp}$, $K_{fp}$, $K_{sp}$, $K_{ap}$, respectively. Thus, the proposed DNN based correction provides accurate ATBs which primarily trace the region of high intensity variation.

Fig. 21 illustrates the predicted and corrected ATBs within the vocal tract using $\text{DNN}_{1}^d$, $\text{DNN}_{2}^d$, $\text{DNN}_{3}^d$, $\text{DNN}_{4}^d$ schemes. As explained in Section 3, for $\text{DNN}_{3}^d$ approach, $C^1_{1,}\text{gv}$ and $C^2_{1,}\text{gv}$ are considered as predicted upper and lower ATBs ($C^1_1$ and $C^2_1$).
respectively. From Fig. 21, it is observed that the DNN based correction provides ATBs by more accurately tracing the transition region between the airway cavity and tissue regions compared to the predicted ATBs. The DNN based correction shows improvement in the regions where the predicted ATB lies in constant intensity (either in air cavity or in tissue) regions. However, DNN does not perform correction in the regions where the predicted ATB is already in a high intensity variation region. Thus, DNN does correction in the regions where the predicted ATB is erroneous without disturbing the regions where the predicted ATB is accurate. The DNN based correction approach is very effective in tongue base, tongue tip and tongue dorsum, hard palate and

**Fig. 20.** Illustration of kernel density estimates of the slopes of intensity curves for ground truth and corrected lower ATBs within vocal tract for F1, F2, M1 and M2 subjects.

**Fig. 21.** Illustration of predicted and corrected upper and lower ATBs within the vocal tract for (a) DNNdM, (b) DNNdF, (c) DNNdS, (d) DNNdAs schemes.
velum regions. However, the DNN based correction does not show improvement in the epiglottis region even though the predicted ATBs are erroneous which can be clearly seen in Fig. 21(c). The corrected ATBs are jagged in some regions compared to the predicted ATBs since the DNN based correction is done for each predicted ATB point independently without considering the neighbouring ATB points. Although, smoothing is done to avoid the jaggedness, it does not remove jaggedness completely. For smoothing, the parameter $Q$ is chosen based on the validation data performance to avoid over-smoothness which may lead to inaccurate ATBs. The optimum value of $Q$ ranges from 2 to 20 across different folds and subjects.

Fig. 22 illustrates the predicted and corrected completed ATBs for four subjects. The regions in the black boxes do not have the high intensity variation in the vicinity of the predicted ATBs.

Fig. 23. Illustration of ground truth, predicted and corrected complete upper ATBs in M2 and F1 subjects.

Fig. 22. Illustration of predicted and corrected complete ATBs from (a) F1 (b) F2 (c) M1 and (d) M2 subjects for SegNet (a,b) and FDM (c,d) approaches respectively. The regions in the black boxes do not have the high intensity variation in the vicinity of the predicted ATBs.
intensity variation. Thus, DNN does ATB correction effectively by learning the factors beyond high intensity variation based on annotator’s knowledge available in the training data. From Fig. 22, it is observed that DNN does correction in the following overlapped regions: 1) hard palate and tongue base 2) upper lip and lower lip 3) velum and tongue base. However, DNN does not do correction in the following overlapped regions: 1) velum and pharyngeal wall 2) glottal base and tongue dorsum. Fig. 23 illustrates the corrected ATBs which have higher DTW distance compared to the predicted ATBs. It is observed that the corrected ATBs are erroneous and jagged in velum and lip regions. In particular, considering the case illustrated in Fig. 23(a), the performance of the DNN based correction scheme is limited by the accuracy of the predicted contour. In other words, a highly inaccurate contour cannot be corrected by the proposed DNN based correction scheme.

Conclusion

In this paper, we propose a DNN based ATB correction approach using a normal-grid based feature extraction method. The DNN based correction improves the accuracy of the predicted ATBs obtained from MG, FDM and SegNet approaches in terms of the average DTW distance. The DNN based correction approach is also used as a direct ATB prediction approach in which the average of the ground truth ATBs in the training set is considered as the predicted ATB. The DNN based prediction approach performs better than the MG approach in terms of the average DTW distance. The DNN based correction approach provides accurate ATBs which trace the region of high intensity variation. DNN does correction in the regions where the predicted ATB is erroneous without disturbing the regions where the predicted ATB is accurate. DNN also performs corrections in the overlapped regions where the high intensity variation is not observed. Thus, DNN does ATB correction effectively by learning the factors beyond the high intensity variation from the training data based on annotator’s knowledge.

The DNN based correction fails to show improvement in epiglottis region, regions where velum and pharyngeal wall overlap and regions where glottal base and tongue dorsum overlap. The corrected ATBs are jagged in some regions compared to the predicted ATBs since the DNN based correction is done for each predicted ATB point without considering the neighbouring ATB points. DNN does not do correction if the predicted ATB is highly erroneous since it does not lie in the vicinity of the ground truth ATB. Thus, the performance of the DNN based correction is limited by the accuracy of the predicted ATB.

In this work, the DNN is trained and validated in seen subject conditions for the given predicted ATBs from a typical ATB segmentation approach. Hence, the DNN based correction approach needs to be analysed for unseen subjects and for unseen prediction approach. To avoid the jaggedness in the corrected ATB, the predicted ATB can be jointly corrected instead of correcting each predicted ATB point independently. A different feature extraction method is needed to improve the accuracy of the corrected ATBs in velum and epiglottis regions. An attention mechanism can be used to consider the weighted average of the outputs corresponding to each grid on the normal for a test predicted ATB point.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References


Li, M., Kim, J., Lammert, A., Ghosh, P.K., Ramanarayanan, V., Narayanan, S., 2016. Speaker verification based on the fusion of speech acoustics and inverted articularatory signals. in Computer Speech and Language 36, 196–211.


Toutios, A., Narayanan, S., 2015. Factor analysis of vocal-tract outlines derived from real-time magnetic resonance imaging data. 18th International Congress of Phonetic Sciences, ICPhS.


