Automatic segmentation of vocal tract articulators in real-time magnetic resonance imaging

Vinicius Ribeiro, Karyna Isaieva, Justine Leclere, Jacques Felblinger, Pierre-André Vuissoz, Yves Laprie

Automatic Segmentation of Vocal Tract Articulators in Real-Time Magnetic Resonance Imaging

Vinicius Ribeiro[a], Karyna Isaieva[b], Justine Leclere[b,c], Jacques Felblinger[b,d], Pierre-André Vuissoz[b], Yves Laprie[a]

[a] *Universite de Lorraine CNRS Inria LORIA Nancy, F-54000, France*
[b] *Universite de Lorraine INSERM U1254 IADI Nancy, F-54000, France*
[c] *Service de Medecine Bucco-dentaire Hopital Maison Blanche Reims, F-51100, France*
[d] *CIC-IT 1433 INSERM CHRU Nancy, F-54000, France*

Abstract

Background and Objectives: The characterization of the vocal tract geometry during speech interests various research topics, including speech production modeling, motor control analysis, and speech therapy design. Real-time MRI is a reliable and non-invasive tool for this purpose. In most cases, it is necessary to know the contours of the individual articulators from the glottis to the lips. Several techniques have been proposed for segmenting vocal tract articulators, but most are limited to specific applications. Moreover, they often do not provide individualized contours for all soft-tissue articulators in a multi-speaker configuration.

Methods: A Mask R-CNN network was trained to detect and segment the vocal tract articulator contours in two real-time MRI (RT-MRI) datasets with speech recordings of multiple speakers. Two post-processing algorithms were then proposed to convert the network's outputs into geometrical curves. Nine articulators were considered: the two lips, tongue, soft palate, pharynx, arytenoid cartilage, epiglottis, thyroid cartilage, and vocal folds. A leave-one-out cross-validation protocol was used to evaluate inter-speaker generalization. The evaluation metrics were the point-to-closest-point distance and the Jaccard index (for articulators annotated as closed contours).

Results: The proposed method accurately segmented the vocal tract articulators, with an average root mean square point-to-closest-point distance of less than 2.2 mm for all the

articulators in the leave-one-out cross-validation setting. The minimum $\texttt{P2CP}_{\texttt{RMS}}$ was 0.91 mm for the upper lip, and the maximum was 2.18 mm for the tongue. The Jaccard indices for the thyroid cartilage and vocal folds were 0.60 and 0.61, respectively. Additionally, the method adapted to a new subject with only ten annotated samples.

Conclusions: Our research introduced a method for individually segmenting nine non-rigid vocal tract articulators in real-time MRI movies. The software is openly available as an installable package to the speech community. It is designed to develop speech applications and clinical and non-clinical research in fields that require vocal tract geometry, such as speech, singing, and human beatboxing.

Keywords:

Segmentation, Articulation, Vocal Tract, MRI, Speaker-Independent

1. Introduction

The characterization of the complete vocal tract geometry is essential for many aspects of speech research, including the analysis of the speech sounds [1], modeling the relationships between sounds and gestures [2], developing articulatory models of speech production [3, 4, 5], studying critical articulators [6], compensatory effects [7], speech motor control [8], and others.

Many techniques exist for achieving this goal, but each has limitations. X-ray cineradiography was one of the first approaches [9, 10]; however, X-rays are ionizing radiation, which is dangerous for the subjects and was abandoned after having been used for the recording of a few sentences per speaker up to the 1980s [11]. Electromagnetic articulography (EMA) is a widespread technique. It consists of attaching small sensors to the articulators and monitoring their spatial positions over time with a high sampling frequency [12]. The disadvantages are the small number of sensors that can be used without disrupting speech and the inability to track articulators in the back of the vocal tract and the pharynx.

The X-ray and EMA weaknesses explain the vast use of magnetic resonance imaging

(MRI) in the speech community [3, 13, 6]. MRI is safe for the subject, enabling the recording of large datasets and permitting the acquisition of high-resolution volumetric images to visualize the complete vocal tract shape. Static volumetric MRI was extensively used to record sustained phonemes; however, the long acquisition duration per volume limits their use for recording dynamic processes such as speech, even though attempts can be found in the literature [14, 15]. Real-time magnetic resonance imaging (RT-MRI) overcomes the temporal resolution problem at the expense of being limited to one plane – usually the mid-sagittal – and a smaller spatial resolution – generally, between 1.5 to 2 mm. Lingala et al. [16] present guidelines, technical considerations, and recommendations for RT-MRI for studying speech. Yet, MRI presents disadvantages. First, it is substantially more expensive than the other techniques. Second, MRI technology is only sensitive to structures containing water; thus, bones, with short T2*, are indistinguishable from air in the images. Third, the MRI machine is claustrophobic, prohibitive for some subjects. Fourth, the subjects cannot have any ferromagnetic material in their bodies, including prostheses, dental braces, and pacemakers. Fifth, the subjects are constrained to the supine position, which differs from natural speech postures. Finally, the machine produces such intense noise that it requires destructive denoising algorithms to be used. As a result, the recorded speech signal has poor quality. Still, RT-MRI is state-of-the-art for vocal tract observation and has been widely used to study swallowing, speech, singing [17], blowing wind instruments [18], and human beatboxing [13].

Although RT-MRI displays the entire vocal tract, the images alone are insufficient for many speech research applications. The reason is that the exact geometry of the vocal tract air column, determined by the contours of the articulators from the glottis to the lips, is often needed. This problem has received much attention from the articulatory speech community. Raeesy et al. [19] proposed a method of automatic landmark tagging in which a recursive boundary subdivision algorithm [20] extracts a set of landmarks corresponding to the vocal tract contours. Then, the oriented active shape model [21] recognizes and delineates the vocal

tract shapes. However, the small dataset (25 images from five speakers) limits the significance of this work. Alternatively, Silva and Teixeira [22] proposed an unsupervised method based on a modified version of the active appearance model [23]. The technique takes advantage of the low inter-frame differences and is based on 26 vocal tract landmarks manually marked per frame in a training database of 51 images. Moving towards machine learning algorithms, Labrunie et al. [24] explored a large RT-MRI corpus in French for training three supervised segmentation methods: Multiple Linear Regression (using pixel intensities), a modified version of the Active Shape Model (mASM), and Shape Particle Filtering (using more elaborate image features). In a leave-one-out cross-validation scheme, the three methods were compared on several articulators using the point-to-closest-point distance (`P2CP`). The results showed that the mASM outperforms the other two for all articulators.

Meanwhile, deep neural networks have become the standard for computer vision and medical image processing [25]. Ca et al. [26] segmented air-tissue boundaries (ATB) using a Fully Convolutional Network (FCN) [27] followed by a canny edge detection algorithm to output smooth and realistic ATBs. Following a strategy similar to Fasel and Berry [28], Jaumard-Hakoun et al. [29] trained a deep neural network based on the stacking of Restricted Boltzmann Machines [30] with the contours extracted by an automatic algorithm that uses block-matching to enforce the frame-to-frame similarity. Eslami et al. [31] explored the segmentation of the jaw, given by the lower incisor profile, tongue, and vocal tract air cavity on static mid-sagittal MRI for ten subjects sustaining 62 articulations. The method uses a modified version of the pix2pix algorithm [32], taking advantage of the conditional generative adversarial networks.

While most of the approaches presented so far focus on a single-frame prediction, Asadiabadi and Erzin [33] proposed a sequence-to-sequence Deep Temporal Regression Network to estimate the coordinates of the vocal tract and the points separating the articulators, providing individualized contours for each articulator, which is essential to study their contributions during speech. As Hebbar et al. [34] show, using temporal information instead of

making single-frame predictions improves the segmentation when the articulators are in contact. Ruthven et al. [35] segmented six groups of articulators and the vocal tract in RT-MRI using a U-Net-like FCN. The most significant contribution of this work is to use the velopharyngeal closure as a metric for assessing the model's performances together with traditional segmentation metrics such as the Dice coefficient and the Hausdorff distance. As expected, the head segmentation had the best results. In contrast, the soft palate and the incisor had the lowest – the latter is probably explained by the teeth's short T2*. Unlike other works, Ruthven et al. [35] focused on the segmentation mask instead of contours, which explains why no post-processing of the network's outputs was proposed. Finally, Isaieva et al. [36] offered an alternative in which only the pixels in the tongue's edges are segmented. The method uses a U-Net [37] for image segmentation, followed by a graph-based algorithm (discussed later in the paper) to convert the soft probabilities in the network's output into the tongue contour.

Our state of the art review revealed several gaps in the literature. Few papers provide individual contours of all non-rigid articulators. Additionally, only some studies demonstrate generalization across multiple speakers. Finally, none provide an open repository for public usage and evaluation of the proposed methods. Our work aims to fill these gaps by presenting a robust speaker-independent approach to segment the vocal tract articulator contours in RT-MRI movies. We also investigate a speaker adaptation approach to enhance the performance of a target subject. This work proposes a single deep convolutional neural network (DCNN) that can automatically annotate the boundaries of nine vocal tract articulators in RT-MRI frames. The DCNN takes RT-MRI frames as input and outputs a probability map over the pixels belonging to the articulator's boundaries. Additional post-processing is then used to obtain a curve giving the exact shape of each articulator. This work is intended for articulatory speech research, including but not limited to articulation and articulatory speech synthesis. The main contributions of this paper are:

- The coverage of the main non-rigid articulators necessary for speech production;

- The assessment of inter-speaker generalization through leave-one-out cross-validation (LOOCV) protocol;
- The processing of vast RT-MRI corpora with a low error in comparison to human annotations;
- The public availability of the segmentation system[1], allowing it to be tested and audited by the scientific community.

2. Materials

2.1. Datasets

The corpus for this research comprises two real-time MRI datasets of French speakers, `ArtSpeech Database 1` (`ASD1`) and `ArtSpeech Database 2` (`ASD2`). The `ASD1` is a part of the database described in Isaieva et al. [38]. The `ASD1` corpus included 77 sentences which were constructed to provide a good coverage of the phonetic contexts of French vowels. While the published database contains ten subjects, only seven (denoted in this text as S1-S7) were used in this study because the larynx was not visible in the images of the other three. The `ASD1` contains a total of 365 400 frames uniformly distributed between the subjects. The same protocol was used for `ASD2`. This dataset contains 320 000 frames of a single subject and covers a larger speech corpus, ensuring better coverage of the French phonetic context. The subject participating in `ASD2` acquisition was the last subject S7 from the dataset `ASD1`. To distinguish these two sets, the data of S7 from `ASD1` is denoted as S7.1, and its data from `ASD2` is marked as S7.2.

Both datasets were collected using state-of-the-art protocols and recommendations at the Centre Hospitalier Régional Universitaire de Nancy, France. All participants provided written informed consent, and the data were recorded under the approved ethical protocol

[1] https://gitlab.inria.fr/multispeech/vt/vt_tracker

Table 1: Parameters of the MRI acquisition.

Parameter	Value
TR	2.22 ms
TE	1.47 ms
FOV	22.0 cm × 22.0 cm
Pixel Spacing	1.62 mm/pixel
Flip Angle	5 degrees
Slice Thickness	8 mm
Num. of Radial Encoding Lines per Frame	9
Pixel Bandwidth	1 670 Hz/pixel
Image Resolution	136 × 136 pixels

"METHODO" (ClinicalTrials.gov Identifier: NCT02887053). The study was approved by the institutional ethics review board (CPP EST-III, 08.10.01).

The images were acquired with a Siemens Prisma 3T scanner (Siemens, Erlangen, Germany). The radial RF-spoiled FLASH sequence [39] was used, with the parameters listed in Table 1. The films were recorded at a frame rate of 50 fps and reconstructed with the algorithm presented in Uecker et al. [39]. During the acquisitions, the speakers were asked to read out loud sentences that were projected to them. Each acquisition took about 80 seconds.

Due to the difficulty of annotating all of the images in the datasets, samples with a good representation of variability were selected. Initially, this selection was conducted independently for datasets `ASD1` and `ASD2`, by different co-authors of this study, using slightly different methodologies. Later the collected and annotated data were merged to increase the database. In Isaieva et al. [36], it was shown that several hundreds images were sufficient to train a good tongue segmentation algorithm. Therefore, sample sizes of both datasets was selected to be of this order. To ensure the best variability coverage, the images for manual annotation were selected with the k-means algorithm. For `ASD1` a $k = 100$ was selected. The k-means algorithms was applied independently for each subject, and only the closest to the cluster centers images were kept, resulting in 700 images in total. For `ASD2` the algorithm was applied with $k = 10$. The clusters were evenly sampled, resulting in 427 images.

Table 2: Number of annotated samples per subject per dataset split.

Dataset	Subject ID	ID on Isaieva et al. [38]	Gender	Train Images	Validation Images	Test Images
ASD1	S1	P1	Male	71	9	20
	S2	P3	Male	71	9	20
	S3	P5	Male	71	9	20
	S4	P6	Male	71	9	20
	S5	P7	Female	71	9	20
	S6	P9	Female	71	9	20
	S7.1	P10	Female	50	0	50
ASD2	S7.2		Female	310	54	63

The two datasets were split into train, validation, and test as described in Table 2. The data was divided at the complete sequences level, so all samples from the same acquisition were placed at the same split. The reason is that adjacent images are very similar, and putting them in separate sets would introduce bias into the train-test scheme. The validation set for S7.1 is empty because it is the same subject as S7.2, which already has a sizeable validation set.

2.2. Annotation Procedure

We performed semi-automatic annotations. Our previous segmentation system [5] was used to track the upper vocal tract cavity, including the two lips, tongue, soft palate, and pharynx. For the larynx articulators, which were not included in the previous study, we trained a Mask R-CNN network with the 427 samples from ASD2 described in Table 2 to produce a first guess of the contours for each articulator. The models were then used to automatically annotate the unlabeled images in the dataset. The automatic annotations were then carefully reviewed and manually corrected. This semi-automatic procedure allowed us to complete the annotation protocol with limited resources. The image annotations were made as follows:

- **Arytenoid Cartilage:** Through their vocal process, these cartilages are the siege of the vocal cord attachment to the posterior part of the larynx (represented by the

cricoid cartilage). The complete extension of the arytenoid cartilages were annotated, covering a vertical range of about two vertebrae (at the level of the $5^{th}/6^{th}$ cervical vertebrae). The annotation started at point `A` in Figure 1a and continued to point `B`, passing through point `C`.

- **Epiglottis:** The epiglottis is a thin and elongated cartilaginous structure describing the upper-anterior part of the larynx. Given the reduced thickness of this cartilage, we chose to annotate the epiglottis center line, starting from the anterior part of the larynx to the epiglottis posterior extremity. Laprie et al. [40] provided an algorithm for reconstructing the epiglottis from the center line. The annotation started at point `D` in Figure 1a and continued to point `E`.

- **Lower Lip:** This part begins from the anterior part of the mandible (at the lower bottom of the gingiva vestibule) to the external lip hem. The contour of the lower lip was annotated from point `F` to point `G` in Figure 1a.

- **Posterior limit of the pharynx:** This area was annotated from the posterior part of the nasal cavity to the cricoid cartilage (behind the arytenoid cartilage). The annotation started at point `H` in Figure 1a and continued to point `A`.

- **Soft Palate:** The soft palate appears as an elongated structure in the mid-sagittal plane, similar to the epiglottis. For the same reason, we only annotated the center line of this area, from the posterior limit of the hard palate (motionless) to the posterior extremity of the moving part. Like the epiglottis, the algorithm from Laprie et al. [40] can reconstruct the soft palate from the center line. The annotation started at point `I` in Figure 1a and continued to point `J`.

- **Tongue**: The complete extension of the tongue (apex, dorsal part, and root) was annotated, starting at the frenulum on the mouth floor and ending the tongue at the root below the hyoid bone. The annotation started at point `K` in Figure 1a and

continued to point L. The sublingual cavity was marked when visible. Since the MRI is 8 mm thick and the frame rate is low compared to the speed of tongue movements, the uncertainty related to the partial volume effect [41] and blurring is specially marked. In both cases, we decided to annotate the most visible contour.

- **Thyroid Cartilage:** This part was annotated as a closed contour, starting at the anterior limit of the epiglottis (at the level of the $3^{rd}/4^{th}$ cervical vertebrae) and ending below the vocal folds (approximately at the level of the 7^{th} cervical vertebrae). This annotation is drawn as the oval shape passing through points M, D, and N in Figure 1a. The position of the thyroid cartilage is more important than its precise shape for confirming the position of the glottis.

- **Upper Lip:** This upper part was drawn from the anterior nasal spine (at the upper bottom of the gingiva vestibule) to the external upper lip hem ("cupid's bow"). The complete contour of the upper lip was annotated from point O to point P in Figure 1a.

- **Vocal Folds:** The vocal folds are not entirely visible in the MRI frame. Only the negative of the glottis is observable between the thyroid cartilage and the arytenoid cartilage. The vocal folds are marked as an oval passing through D and C in Figure 1a.

Figure 1 shows an MRI frame with the landmarks used to reproduce the annotation procedure and three MRI samples with superimposed annotations for each articulator. A dental surgeon with seven years of experience validated the annotation procedure.

3. Methods

The strategy used to track the shapes of the articulators is similar to that used by Isaieva et al. [36] for the tongue. It comprises two phases:

1. A DCNN is trained to estimate the probability that a pixel belongs to the articulator's contour;

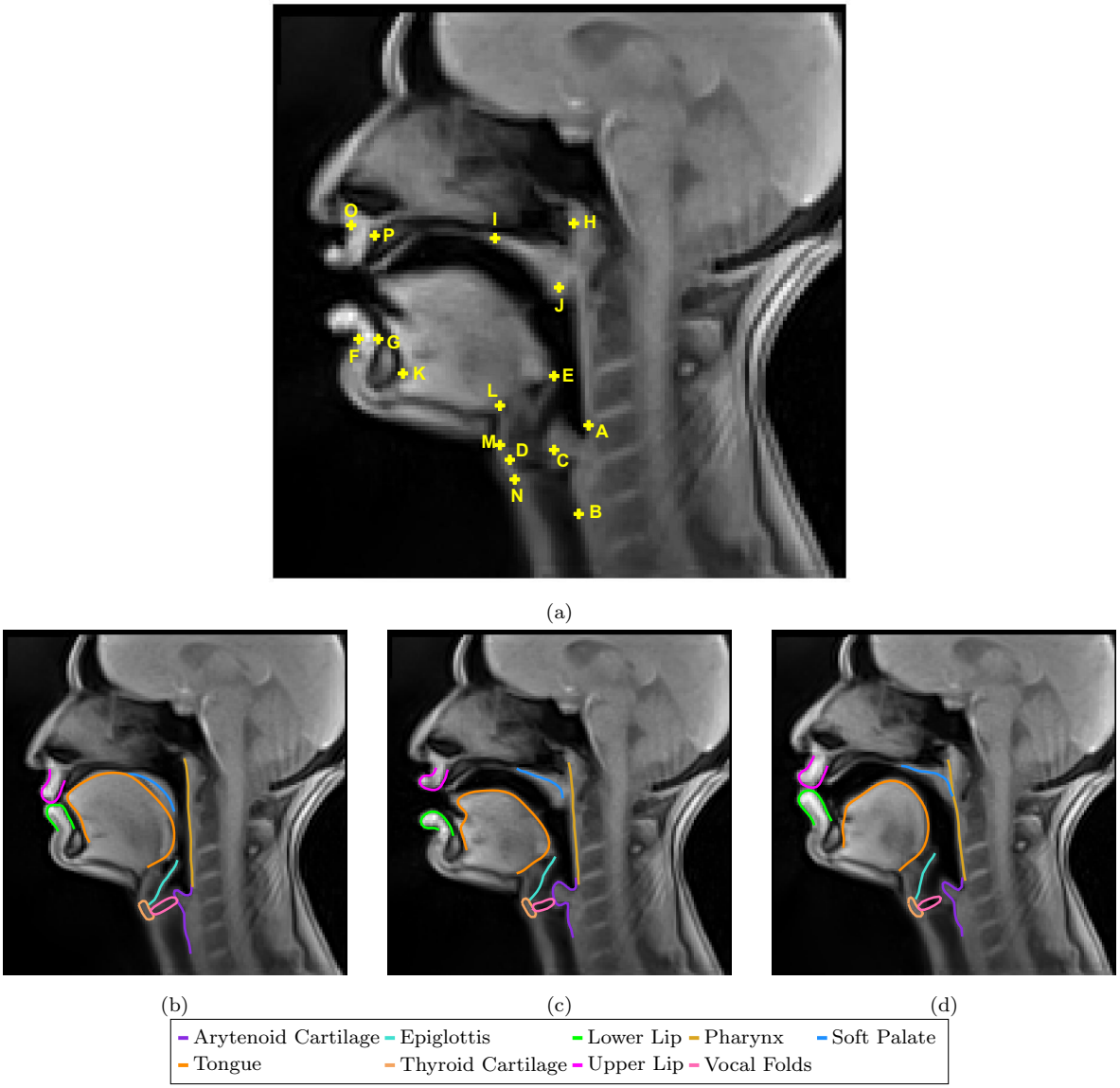


Figure 1: Landmarks in a mid-sagittal MRI sample and annotation samples exemplifying the procedure. (a) The mid-sagittal MRI sample shows the landmarks that were used to guide the annotation procedure described in subsection 2.2. (b-d) The annotation samples show how the articulators were annotated in the MRI sample.

2. A post-processing algorithm is applied to the network's outputs to construct the curve describing the articulator's shape. The nature of the algorithm and its hyperparameters depend on the articulator.

Section 3.1 describes the learning strategy used in the first phase, while Section 3.2 describes the algorithms applied to each articulator. The final contours are regularized using b-splines to match the target and predicted curve lengths[2].

3.1. Articulator Boundaries Segmentation

Unlike Isaieva et al. [36], who used the U-Net [37], we chose to work with the Mask R-CNN [42]. Mask R-CNN is a simple and flexible framework for object instance segmentation, and it is lightweight, easy to train, and can be applied to different tasks. These characteristics, as well as the availability of a pre-trained implementation[3] on standard deep learning libraries, make Mask R-CNN one of the preferred methods for medical image segmentation [43].

The Mask R-CNN architecture is advantageous for our problem because it performs three tasks simultaneously: object detection, classification, and segmentation. This approach allows localizing the region of interest before segmenting it, which avoids spurious predictions in image regions that do not correspond to the articulator.

The models were pre-trained on the `COCO train2017` dataset [44], which contains RGB images of common objects. However, the MRI frames are grayscale. Therefore, we used the temporal dimension to build a more contextualized input. The network's inputs were formed by putting frames $t-1$, t, and $t+1$ in the first, second, and third input channels, respectively.

[2]The b-spline regularization function can be found at https://gitlab.inria.fr/multispeech/vt/vt_tools/-/blob/main/vt_tools/bs_regularization.py

[3]https://pytorch.org/vision/main/models/generated/torchvision.models.detection.maskrcnn_resnet50_fpn.html

3.2. Post-processing Algorithms

The post-processing of the network's outputs depends on the articulator. For each articulator, a specific algorithm is chosen, including additional sub-steps and adjustments of several hyperparameters. This section explains the two approaches developed according to the articulator contour's closed/open nature.

For articulators that were annotated as closed contours, we utilize the largest contiguous ISO-valued contour (Section 3.2.1). For open contours, we utilize a graph-based algorithm (Section 3.2.2). Figure 2 presents one sample of the network's output for each articulator, illustrating the inputs of the post-processing algorithms.

3.2.1. Largest Contiguous ISO-valued Contour

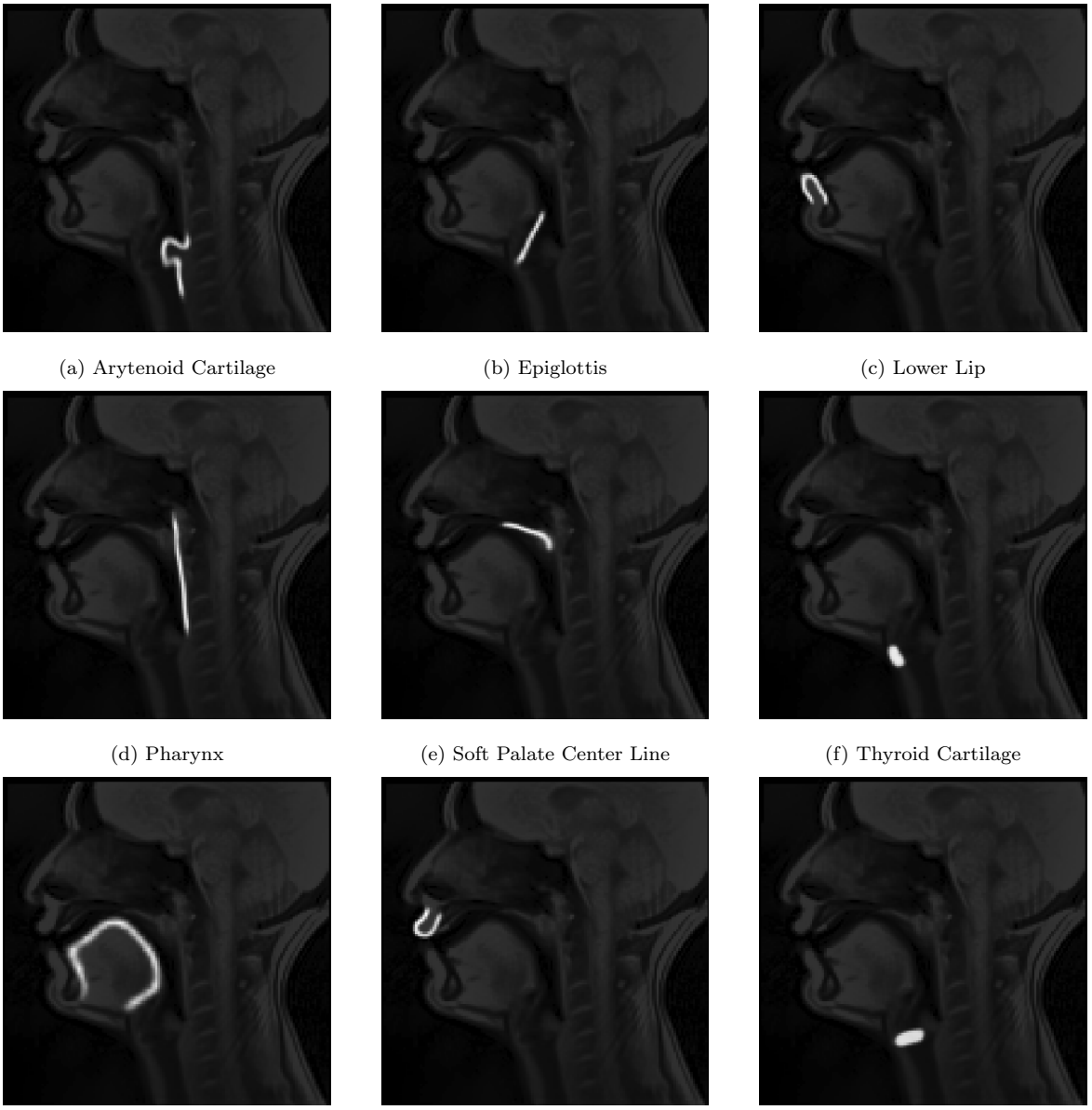
For articulators annotated as closed contours, the contour can be found by calculating the largest contiguous ISO-valued contour in the probability map. We used the `find_contours` function[4] from `scikit-image` [45], which uses the marching squares method to compute the ISO-valued contours of the input 2D array for a particular level value. In our case, the level value is 1, obtained after thresholding the probability map. Figure 3 presents each step of the algorithm on a custom synthetic image.

We use this method for the **thyroid cartilage** and **vocal folds**, with thresholds of 0.7 and 0.8, respectively. In rare cases, the network may output two separate blobs for one articulator, producing two non-contiguous contours. In these cases, we choose to keep the largest area as the true contour.

3.2.2. Graph-based Algorithm

The open contours can be found by expressing the non-zero pixels in the network's outputs as graph nodes and connecting the extremities using Dijkstra's shortest path algorithm [46]. We use this algorithm for the **arytenoid cartilage**, **epiglottis center line**, **lower and**

[4] https://scikit-image.org/docs/0.8.0/api/skimage.measure.find_contours.html

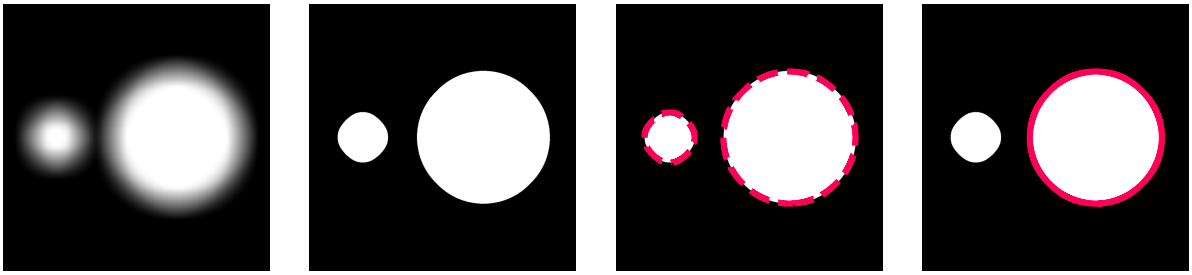


(a) Arytenoid Cartilage (b) Epiglottis (c) Lower Lip

(d) Pharynx (e) Soft Palate Center Line (f) Thyroid Cartilage

(g) Tongue (h) Upper Lip (i) Vocal Folds

Figure 2: Illustration of the segmentation mask for each articulator in one MRI sample. The segmentation masks are superimposed on the MRI sample with very low transparency to help the reader localize the articulator in the MRI. These segmentation masks are the inputs of the post-processing algorithms.



(a) Probability map input with multiple blobs

(b) Probability map after thresholding

(c) ISO-valued contours for both blobs

(d) Contour with the largest area

Figure 3: Steps of the largest contiguous ISO-valued contour algorithm for an illustrative artificial input.

upper lips, **pharynx**, **soft palate center line**, and **tongue**. The method requires a particular set of steps to be performed, which depends on the articulator.

The first step is thresholding to limit the number of nodes in the graph, which is done in two ways. In the first, the pixel value is given by

$$p_{\texttt{new}} = \begin{cases} 0, & \text{if } p_{\texttt{orig}} \leq T \\ 1, & \text{otherwise} \end{cases}$$

In the second, the pixel value is given by

$$p_{\texttt{new}} = \begin{cases} 0, & \text{if } p_{\texttt{orig}} \leq T \\ p_{\texttt{orig}}, & \text{otherwise} \end{cases}$$

where $p_{\texttt{orig}}$ is the original pixel value, $p_{\texttt{new}}$ is the updated pixel value, and T is the threshold value. The first is helpful when the relevant information is whether the pixel belongs or not to the articulator contour. Contrarily, the latter is helpful when a graded information is necessary.

The second step is skeletonization performed using the `skeletonize` function[5] from `scikit-image`, which favors inner pixels in the probability map. Then, the centers of the

[5] https://scikit-image.org/docs/stable/api/skimage.morphology.html#skimage.morphology.skeletonize

non-zero pixels in the image are converted to the nodes of a graph, and the edges between the nodes are created based on the Euclidean distance between them and the probability of each node. The weight is given by

$$w_{ij} = \alpha \cdot d(i,j) + \beta \cdot (1 - p_j)$$

where $d(i,j)$ is the Euclidean distance between node i and node j and p_j is the probability of node j. An edge is set between two nodes if the infinity norm between them is lower than two pixels.

The next step is determining the contour extremities using one of three methods: the greatest angular distance, vertical extremities, or horizontal extremities. The graph's center of mass (CM) is used as the reference for the greatest angular distance in all cases except for the arytenoid cartilage. For the arytenoid cartilage, only the CM's y-coordinate is used, and the x-coordinate is set to the right-most edge of the image. The two points with the greatest angular distance from the reference are selected as the extremities.

For the vertical extremities, the top-most and the bottom-most nodes in the graph are used, while for the horizontal extremities, the left-most and the right-most nodes in the graph are used. Finally, Diijkstra's algorithm is used to connect the two extremities, and the final contour is output. Table 3 summarizes the specific steps and parameters of the graph-based algorithm for each articulator. Figure 4 illustrates the graph-based algorithm's main steps for the tongue.

3.3. Experimental Design

We aimed to develop a speaker-independent method to accurately and individually track non-rigid vocal tract articulators in RT-MRI movies. We also wanted to investigate how speaker adaptation could improve the method's performance for a new subject.

We carried out two experiments. The first was a LOOCV protocol. We removed subject S7.1/S7.2 from the test phase in the LOOCV pipeline because they account for the most

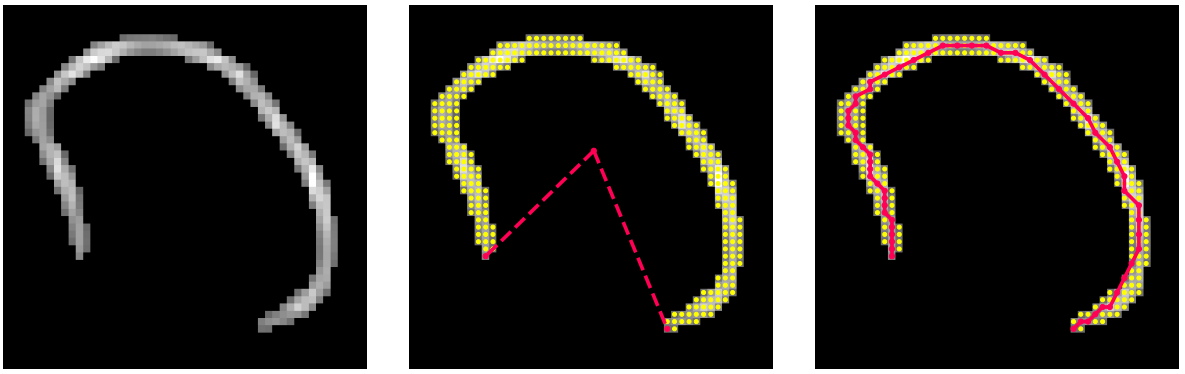
Table 3: Steps and parameters of the graph-based algorithm for all articulators.

Articulator	Threshold Value (Type)	Skeletonize	Extremities Choice	Ang. Distance Reference	α	β
Arytenoid Cartilage	0.2 (0/1)	Yes	Angular distance	CM's y-coordinate + Right-most x-coordinate	1	10
Epiglottis	0.3 (0/1)	Yes	Vertical extremities	–	1	10
Lower Lip	0.4 (0/1)	Yes	Angular distance	CM	1	10
Pharynx	0.3 (0/1)	Yes	Vertical extremities	–	1	10
Soft Palate Center Line	0.1 (0/1)	Yes	Horizontal extremities	–	1	10
Tongue	0.2 ($0/p_{\texttt{orig}}$)	No	Angular distance	CM	10^{-7}	1
Upper Lip	0.4 (0/1)	Yes	Angular distance	CM	1	10

images in the database, but still kept it for training. Leaving S7.1 and S7.2 out would have resulted in a significant reduction in the training set, making it difficult to determine the cause of the performance improvement. From the remaining six subjects, we isolated one at a time and trained a model with the remaining subjects. We then tested the model on the test set of the left-out subjects. We also tested all of the LOOCV models on S7.1 and S7.2 test sets.

In the second experiment, we fine-tuned each initially trained model with its respective left-out subject. We did this using 10, 40, and all the training samples. We then evaluated the improvement in performance on the test sets of the left-out subject and the test sets of subjects S7.1 and S7.2. Ideally, the adapted model would improve its performance for the target subject while keeping the performance of the previously seen subjects constant.

The models were trained using the Adam optimizer [47] with the cyclic learning rate



(a) Tongue's probability map after thresholding

(b) Graph nodes and contour's extremities given by the maximum angular distance

(c) Dijkstra's shortest path

Figure 4: Steps of the graph-based algorithm for an illustrative artificial input.

Table 4: Hyperparameters of the articulator segmentation model. This table summarizes the hyperparameters used to train and evaluate the segmentation network. The second part of the table refers to the speaker adaptation experiments.

Hyperparameter	Value
Batch size	8
Early-Stopping Patience	20
Weight Decay	10^{-3}
Sched. Max. Learning Rate	10^{-4}
Sched. Base Learning Rate	2×10^{-6}
Adapt. Num. Epochs.	20
Adapt. Learning Rate.	10^{-5}
Adapt. Sched. Red. Factor	10
Adapt. Sched. Patience	10

scheduler policy [48]. The training continued for the speaker adaptation experiments using the reduced learning rate on plateau scheduler policy. The hyperparameters of the training are given in Table 4. The machine learning code was developed using `PyTorch` [49]. The complete code for reproducing our results and using our software is available in our public repositories[6][7].

[6] https://github.com/vribeiro1/vocal-tract-seg

[7] https://gitlab.inria.fr/multispeech/vt/vt_tracker

3.4. Evaluation

The primary evaluation metric for articulators with open contours is the root mean square (RMS) value of the P2CP from Labrunie et al. [24]. The P2CP metric is calculated by finding the minimum distance between each point in the target contour and the predicted contour. The $\texttt{P2CP}_{\texttt{RMS}}$ is given by

$$\texttt{P2CP}(i,u,v) = \frac{1}{2}\Big(\min_{j\in\{1,2,\ldots,n\}} d(v_i,u_j) + \min_{j\in\{1,2,\ldots,n\}} d(u_i,v_j)\Big)$$

$$\texttt{P2CP}_{\texttt{RMS}}(u,v) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\texttt{P2CP}(i,u,v))^2}$$

where $u \in \mathbb{R}^{n\times 2}$ and $v \in \mathbb{R}^{n\times 2}$ are the target and predicted contours, respectively, and $d(u_i, v_j)$ is the Euclidean distance between points u_i and v_j.

This metric does not suit well articulators with closed contours. Therefore, the primary metric for the thyroid cartilage and the vocal folds is the Jaccard inndex; the P2CP is also computed for completeness. The Jaccard index, also known as Intersection-over-Union (IoU), is calculated by finding the intersection and union of the areas delimited by the target and the predicted contours. The Jaccard index ranges from 0 to 1, with 1 indicating a perfect match.

The inter-subject reproducibility of the results was assessed with the one-way ANOVA test for subjects S1-S6 and the unpaired t-test for subjects S7.1 and S7.2. $p < 0.05$ was considered significant.

4. Results

Generally, the proposed method demonstrated a good segmentation quality. The typical examples of the automatic segmentation compared to the ground truth contours and superimposed with the MRI images are shown in Figure 5 and Figure 6. They are representative of the overall method's performance, showing that it is adequate for different speakers and vocal tract positions. Figure 6 shows cases of swallowing (non-speech), which led to the

worst results. It should be noted that we excluded swallowing images from the test set, and these cases are only discussed for completeness. Table 5 shows the $\texttt{P2CP}_{\texttt{RMS}}$ values for each of illustrative case in millimeters.

Table 6 and Table 7 show the results in the LOOCV. Figure 7 and Figure 9a presents the results in the form of boxplots to make it easier to visualize and compare the results between subjects. The statistical test (one-way ANOVA) shows that the results were significantly different for all articulators except the soft palate center line. However, a visual analysis of Figure 7 and Figure 9a demonstrates that the low p-values are usually explained by a single outlier.

Table 8, Figure 8, and Figure 9b show the results when the models are evaluated on the S7.1 and S7.2 test sets. The statistical test demonstrates a significant difference between the two sets for all articulators except the lower lip and the soft palate center line. However, the differences between the means $\texttt{P2CP}_{\texttt{RMS}}$ distances tend to remain less than 0.5 mm for most articulators. The articulators that do not satisfy this condition are epiglottis, thyroid cartilage, and vocal folds.

The LOOCV results show that the overall performance has an error of less than 2.2 mm, slightly above one pixel (1.36 pixels). The segmentation of articulators annotated as closed contours (thyroid cartilage and vocal folds) provides a Jaccard index of about 60% in both cases. As pointed out, the tables and figures do not include swallowing cases.

Figure 10 and Figure 11 show speaker adaptation's impact on the models' performance, and the x-axis represents the size of the adaptation training set; the y-axis represents the $\texttt{P2CP}_{\texttt{RMS}}$/Jaccard index value on the test set. It can be seen that in case of initially poor inter-subject prediction, the $\texttt{P2CP}_{\texttt{RMS}}$ curve rapidly decreases after retraining with a small amount (usually ten) of additional images.

The supplementary material includes three videos that illustrate our method's result. Each video consists of the RT-MRI acquisition for one utterance with the original audio, extracted contours, phoneme at each frame, and annotation regarding the speaker, sequence,

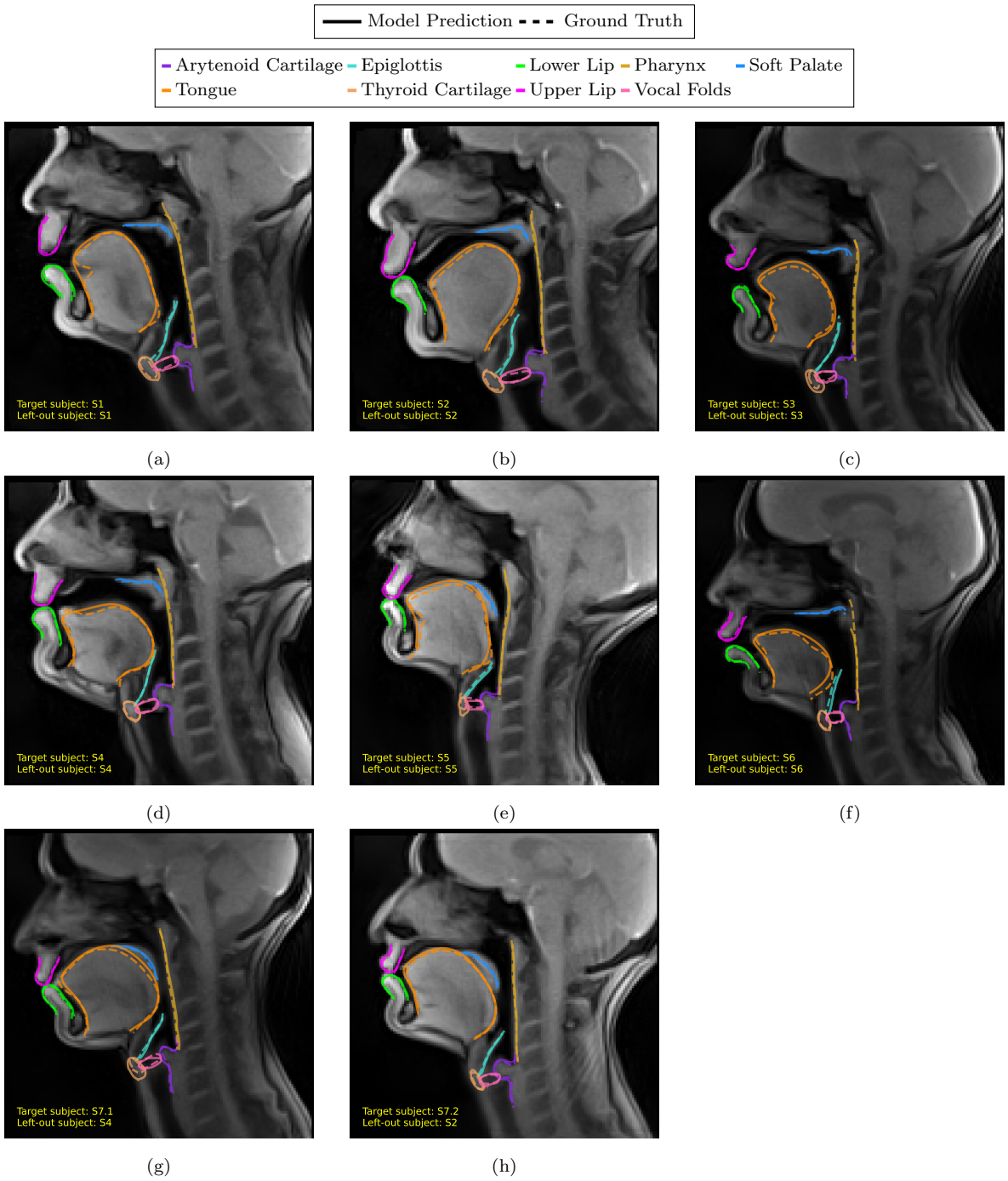
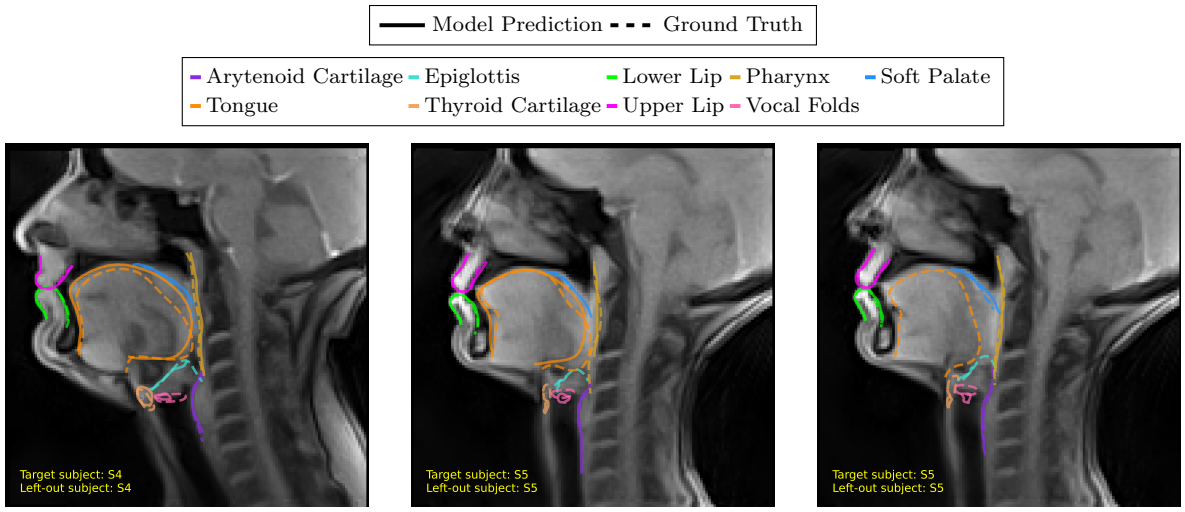


Figure 5: MRI samples of each subject superimposed with the predicted and ground truth contours after b-spline regularization. The text in the images indicates the ID of the subject in the image (target subject) and the ID of the left-out subject during the training of the model that produced that output. This figure shows how the predicted contours compare to the ground truth contours for each subject. The left-out subject is the subject that was not used to train the model.



(a) (b) (c)

Figure 6: MRI samples of swallowing superimposed with the predicted and ground truth contours after b-spline regularization. The text in the images indicates the ID of the subject in the image (target subject) and the ID of the left-out subject during the training of the model that produced that output. The figure shows how the predicted contours compare to the ground truth contours for each subject. The left-out subject is the subject that was not used to train the model. Note that for (c), the model completely missed the tongue.

and frame number.

Table 5: $\texttt{P2CP}_{\texttt{RMS}}$ error (in millimeters) and the Jaccard index (for closed contours) for the samples presented in Figure 5 and Figure 6.

Figure	Arytenoid Cartilage	Epiglottis	Lower Lip	Pharynx	Soft Palate Center line	Thyroid Cartilage		Tongue	Upper Lip	Vocal Folds	
	$\texttt{P2CP}_{\texttt{RMS}}$	$\texttt{P2CP}_{\texttt{RMS}}$	$\texttt{P2CP}_{\texttt{RMS}}$	$\texttt{P2CP}_{\texttt{RMS}}$	$\texttt{P2CP}_{\texttt{RMS}}$	$\texttt{P2CP}_{\texttt{RMS}}$	Jacc.Ind.	$\texttt{P2CP}_{\texttt{RMS}}$	$\texttt{P2CP}_{\texttt{RMS}}$	$\texttt{P2CP}_{\texttt{RMS}}$	Jacc.Ind.
Figure 6a	1.25	2.80	1.02	1.54	3.47	1.52	0.59	3.71	0.53	2.71	0.30
Figure 6b	5.23	5.14	0.95	4.82	1.55	5.28	0.06	3.75	0.93	2.32	0.36
Figure 6c	2.72	4.92	0.85	1.51	1.69	5.83	0.02		0.91	3.26	0.17
Figure 5a	0.96	1.41	1.39	1.42	0.88	1.44	0.73	2.62	0.98	1.04	0.69
Figure 5b	2.07	1.35	0.84	0.96	1.03	1.10	0.73	1.95	0.69	1.06	0.78
Figure 5c	1.29	1.07	0.67	1.29	1.07	2.21	0.52	2.26	0.96	1.05	0.75
Figure 5d	1.12	0.52	0.69	0.91	1.23	0.63	0.89	1.79	0.64	0.72	0.85
Figure 5e	1.06	1.72	0.88	0.87	1.21	1.86	0.48	2.89	1.31	1.25	0.53
Figure 5f	0.56	2.22	1.18	3.82	0.99	0.59	0.85	2.33	1.20	0.57	0.81
Figure 5g	1.01	1.15	0.83	1.03	1.30	2.38	0.52	2.01	0.64	1.17	0.73
Figure 5h	1.25	0.80	0.53	0.73	0.73	0.87	0.70	1.63	0.77	0.77	0.73

Table 6: $\texttt{P2CP}_{\texttt{RMS}}$ error (in millimeters) for each articulator for each left-out subject in the LOOCV setting. The symbol † indicates cases in which the model missed that articulator in one or more images. The column in bold shows the mean ± std in each row. The row in bold shows the mean ± std in each column. The p-values were calculated using a one-way ANOVA test using the subjects as the treatment variable.

Articulator	S1	S2	S3	S4	S5	S6	mean ± std	p-value
Arytenoid Cartilage	1.10 ± 0.41	2.08 ± 0.50	1.11 ± 0.17	1.17 ± 0.28	1.12 ± 0.30	0.92 ± 0.26	$\mathbf{1.25 \pm 0.42}$	1.80×10^{-19}
Epiglottis	1.15 ± 0.32	1.94 ± 0.83	0.90 ± 0.19	1.08 ± 0.33	1.56 ± 0.56	1.14 ± 0.62	$\mathbf{1.29 \pm 0.38}$	2.22×10^{-8}
Lower Lip	1.16 ± 0.26	0.73 ± 0.16	1.00 ± 0.24 †	0.86 ± 0.17	0.95 ± 0.22	1.03 ± 0.23	$\mathbf{0.96 \pm 0.15}$	2.01×10^{-7}
Pharynx	1.07 ± 0.28	0.98 ± 0.17	1.26 ± 1.35	1.07 ± 0.36	1.09 ± 0.17	2.28 ± 1.44	$\mathbf{1.29 \pm 0.49}$	9.34×10^{-6}
Soft Palate Center Line	1.01 ± 0.20	1.15 ± 0.30	1.14 ± 0.40	1.17 ± 0.47	0.98 ± 0.39	1.13 ± 0.45	$\mathbf{1.10 \pm 0.08}$	0.49
Thyroid Cartilage	1.70 ± 0.46	1.77 ± 0.46	1.75 ± 0.26	1.21 ± 0.42	3.78 ± 2.29	0.93 ± 0.27	$\mathbf{1.86 \pm 1.00}$	2.48×10^{-14}
Tongue	2.08 ± 0.42	1.91 ± 0.24	2.24 ± 0.42	1.93 ± 0.36	3.07 ± 0.67	1.87 ± 0.32	$\mathbf{2.18 \pm 0.46}$	8.62×10^{-16}
Upper Lip	0.84 ± 0.18	0.73 ± 0.12	0.81 ± 0.19	0.90 ± 0.39	1.20 ± 0.32	0.99 ± 0.25	$\mathbf{0.91 \pm 0.17}$	1.20×10^{-6}
Vocal Folds	1.35 ± 0.41	1.77 ± 1.21	1.20 ± 0.38	1.35 ± 0.45	2.84 ± 2.03	1.13 ± 0.65	$\mathbf{1.61 \pm 0.64}$	3.15×10^{-6}
mean ± std	$\mathbf{1.27 \pm 0.39}$	$\mathbf{1.45 \pm 0.55}$	$\mathbf{1.27 \pm 0.45}$	$\mathbf{1.19 \pm 0.31}$	$\mathbf{1.84 \pm 1.08}$	$\mathbf{1.27 \pm 0.48}$		0.2207

Table 7: Jaccard index for each articulator with closed contour for each left-out subject in the LOOCV setting. The column in bold shows the mean ± std in each row. The p-values were calculated using a one-way ANOVA test using the subjects as the treatment variable.

Articulator	S1	S2	S3	S4	S5	S6	mean ± std	p-value
Thyroid Cartilage	0.65 ± 0.10	0.61 ± 0.10	0.56 ± 0.07	0.70 ± 0.11	0.33 ± 0.19	0.74 ± 0.07	$\mathbf{0.60 \pm 0.14}$	2.61×10^{-20}
Vocal Folds	0.69 ± 0.09	0.59 ± 0.27	0.70 ± 0.09	0.67 ± 0.14	0.35 ± 0.25	0.67 ± 0.19	$\mathbf{0.61 \pm 0.13}$	5.65×10^{-8}

Table 8: The mean ± standard deviation of the $\texttt{P2CP}_{\texttt{RMS}}$ and Jaccard index (for closed contours only) when the models were tested with S7.1 and S7.2. The symbol † indicates cases in which the model missed that articulator in one or more images. The row in bold shows the mean ± std in each column. The p-values were calculated using the unpaired t-test using the test subjects as the treatment variable.

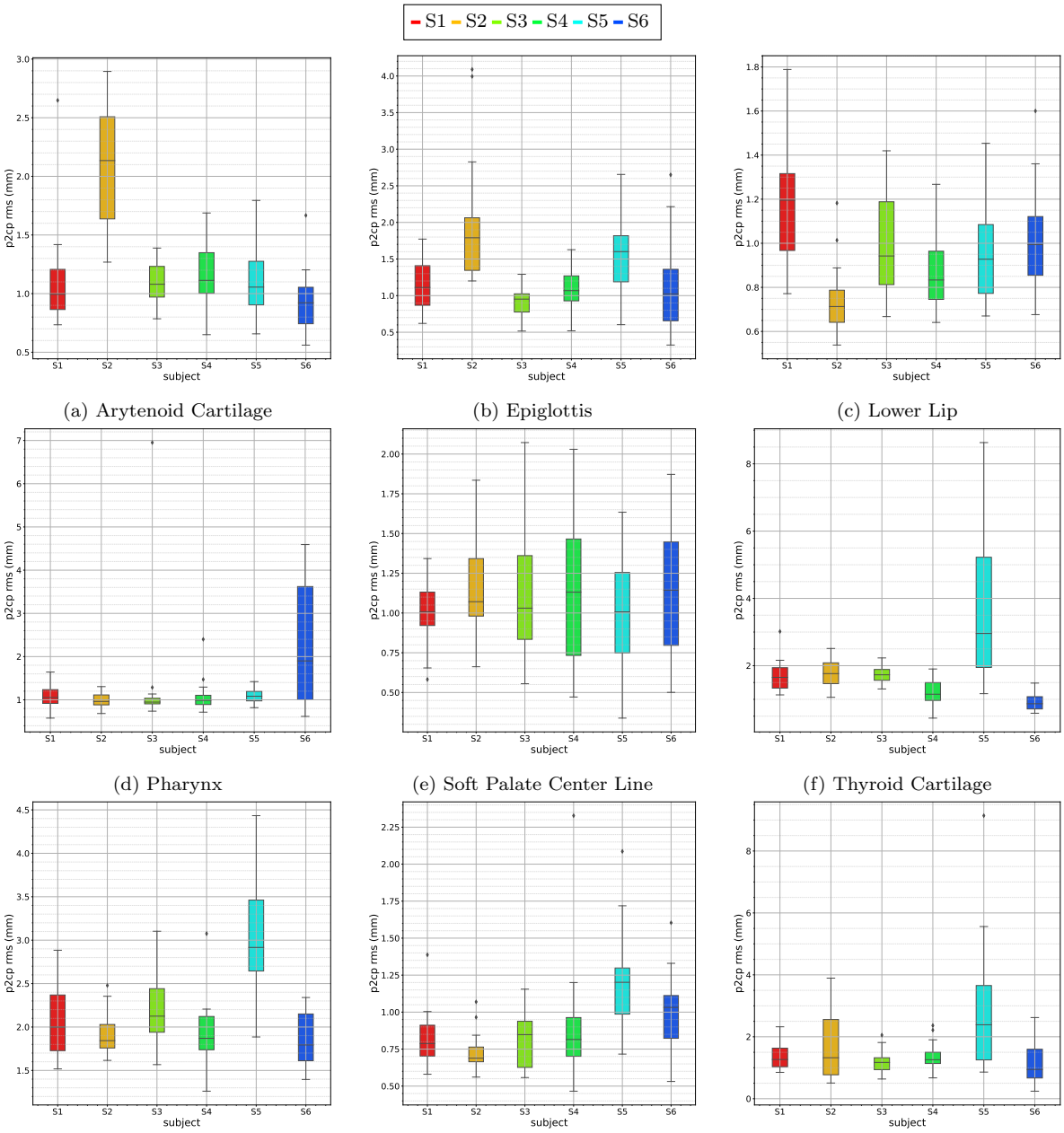
	$\texttt{P2CP}_{\texttt{RMS}}$ (mm)			Jaccard index		
Articulator	S7.1	S7.2	p-value	S7.1	S7.2	p-value
Arytenoid Cartilage	1.09 ± 0.03	1.20 ± 0.06	1.3×10^{-4}			
Epiglottis	1.53 ± 0.08	0.93 ± 0.08	10^{-64}			
Lower Lip	0.80 ± 0.03	0.79 ± 0.03 †	0.35			
Pharynx	0.84 ± 0.03	0.78 ± 0.02	10^{-3}			
Soft Palate	0.94 ± 0.02	0.96 ± 0.05	0.59			
Center line						
Thyroid Cartilage	2.09 ± 0.05	1.03 ± 0.03	10^{-158}	0.53 ± 0.01	0.69 ± 0.01	10^{-74}
Tongue	1.86 ± 0.06	1.39 ± 0.12	10^{-51}			
Upper Lip	0.86 ± 0.04	0.94 ± 0.02	1.8×10^{-3}			
Vocal Folds	1.64 ± 0.05	0.99 ± 0.03	10^{-65}	0.58 ± 0.01	0.74 ± 0.01	10^{-59}
mean ± std	$\mathbf{1.29 \pm 0.49}$	$\mathbf{1.00 \pm 0.19}$	0.1161			

5. Discussion

Our models can segment non-rigid vocal tract articulators with low error and outstanding generalization across subjects, as demonstrated by the LOOCV protocol. Although formal statistical analysis shows significant inter-subject variations of the mean annotation error, these differences are much less than the pixel size.

The model generally performs poorly for the sublingual cavity, as observed in Figure 5a and Figure 5e. It happens because the shortest path algorithm can sometimes miss accentuated curvatures. Nevertheless, the acoustic relevance of the sublingual cavity is minor compared to the tongue tip and tongue dorsum.

Another notable case is Figure 5e, where the tongue deviation is close to 3 mm. In this case, the divergence is due to a possible inconsistency in the annotators' decision. While both annotations could be considered as correct, the annotator of this image selected a more internal part of the tongue body, while the model delineated a more external contour.



(a) Arytenoid Cartilage (b) Epiglottis (c) Lower Lip

(d) Pharynx (e) Soft Palate Center Line (f) Thyroid Cartilage

(g) Tongue (h) Upper Lip (i) Vocal Folds

Figure 7: Distribution of the $\texttt{P2CP}_{\texttt{RMS}}$ error (in millimeters) for each articulator for each left-out subject in the LOOCV setting. The similar information is also shown in Table 6. Attention to the different y-scales when comparing the plots.

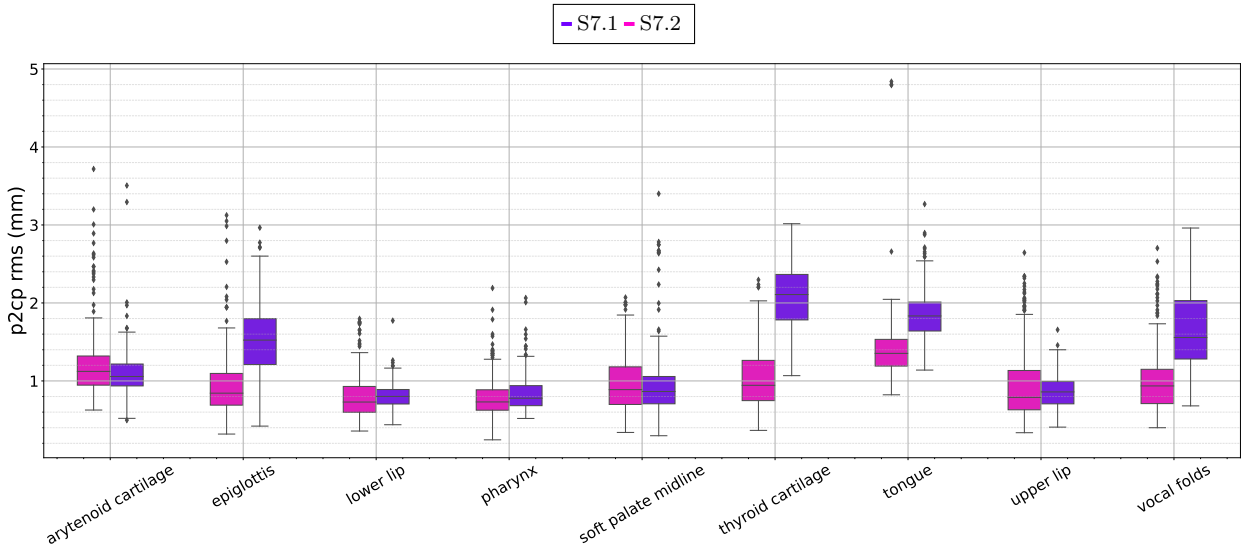
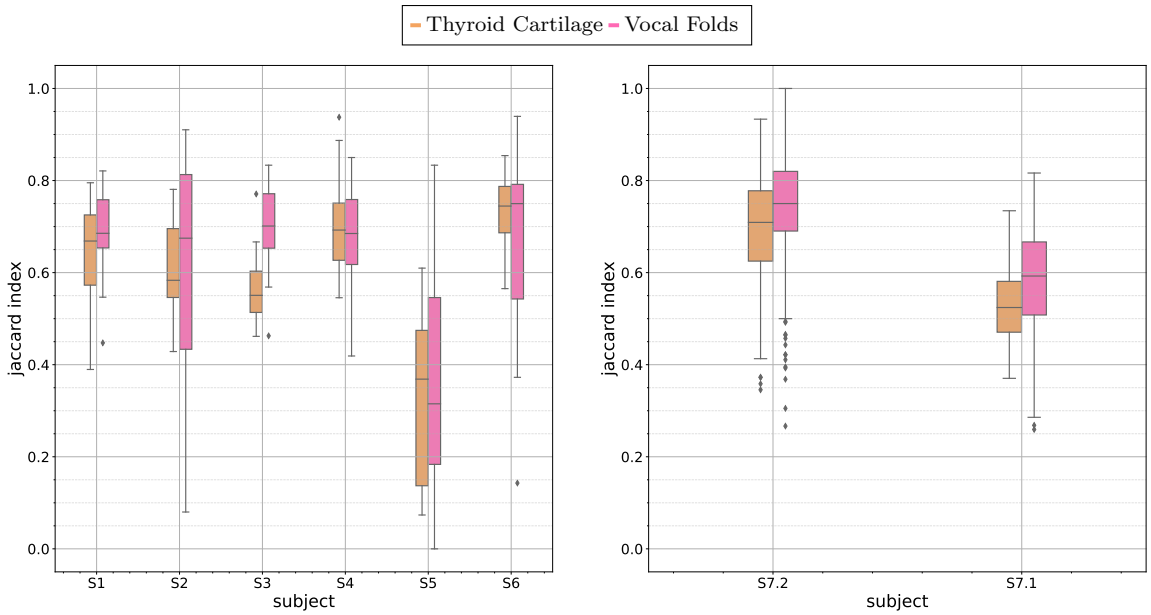


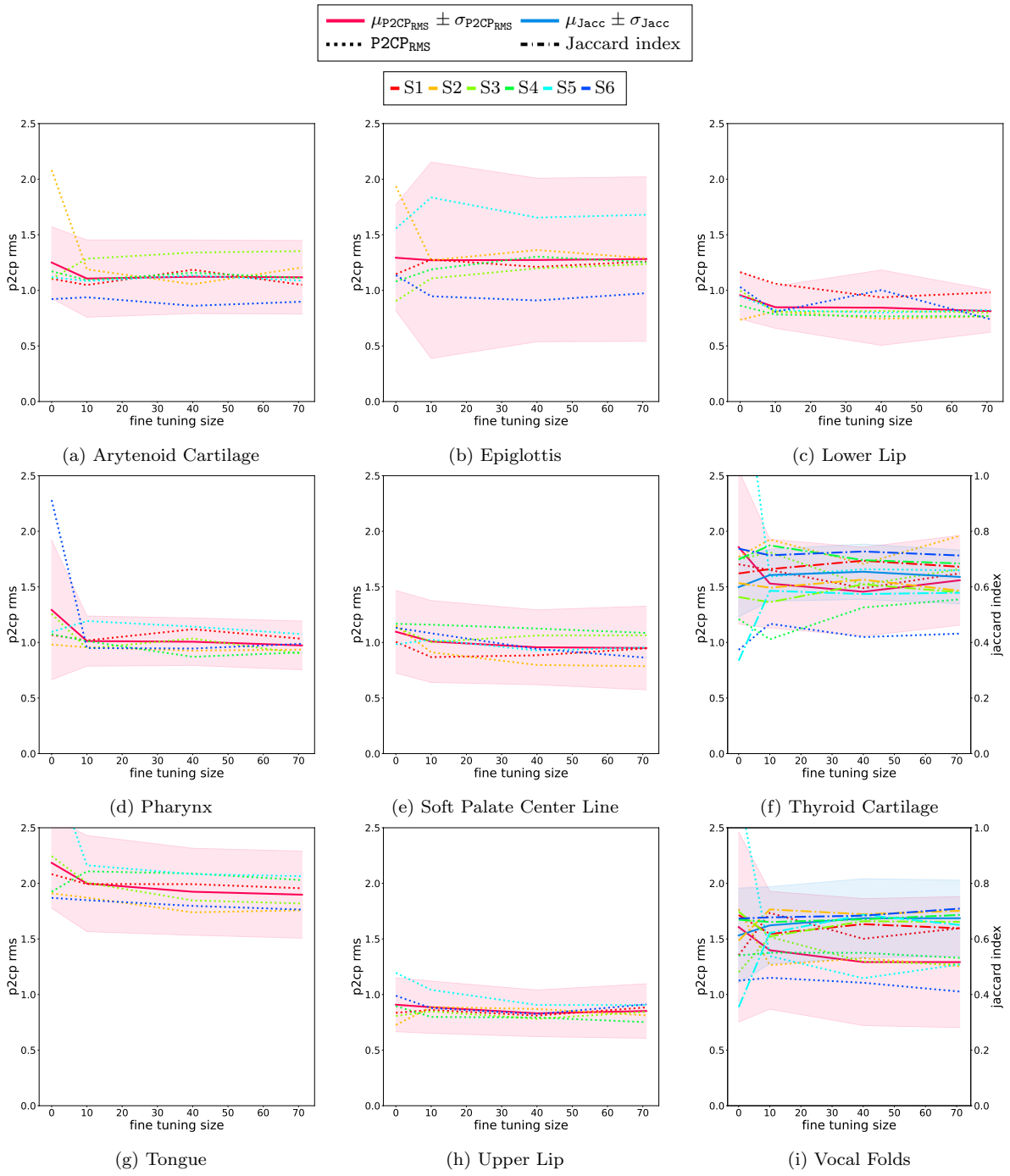
Figure 8: Distribution of the $\texttt{P2CP}_{\texttt{RMS}}$ error (in millimeters) for each articulator for S7.1 and S7.2. Similar information is also shown in Table 8. Attention to the y-scales when comparing with Figure 7.



(a) LOOCV

(b) S7.1 & S7.2

Figure 9: Distribution of the Jaccard index for the articulators with closed contours (a) for each left-out subject and (b) for S7.1 and S7.2. Similar information is also shown in Table 8.

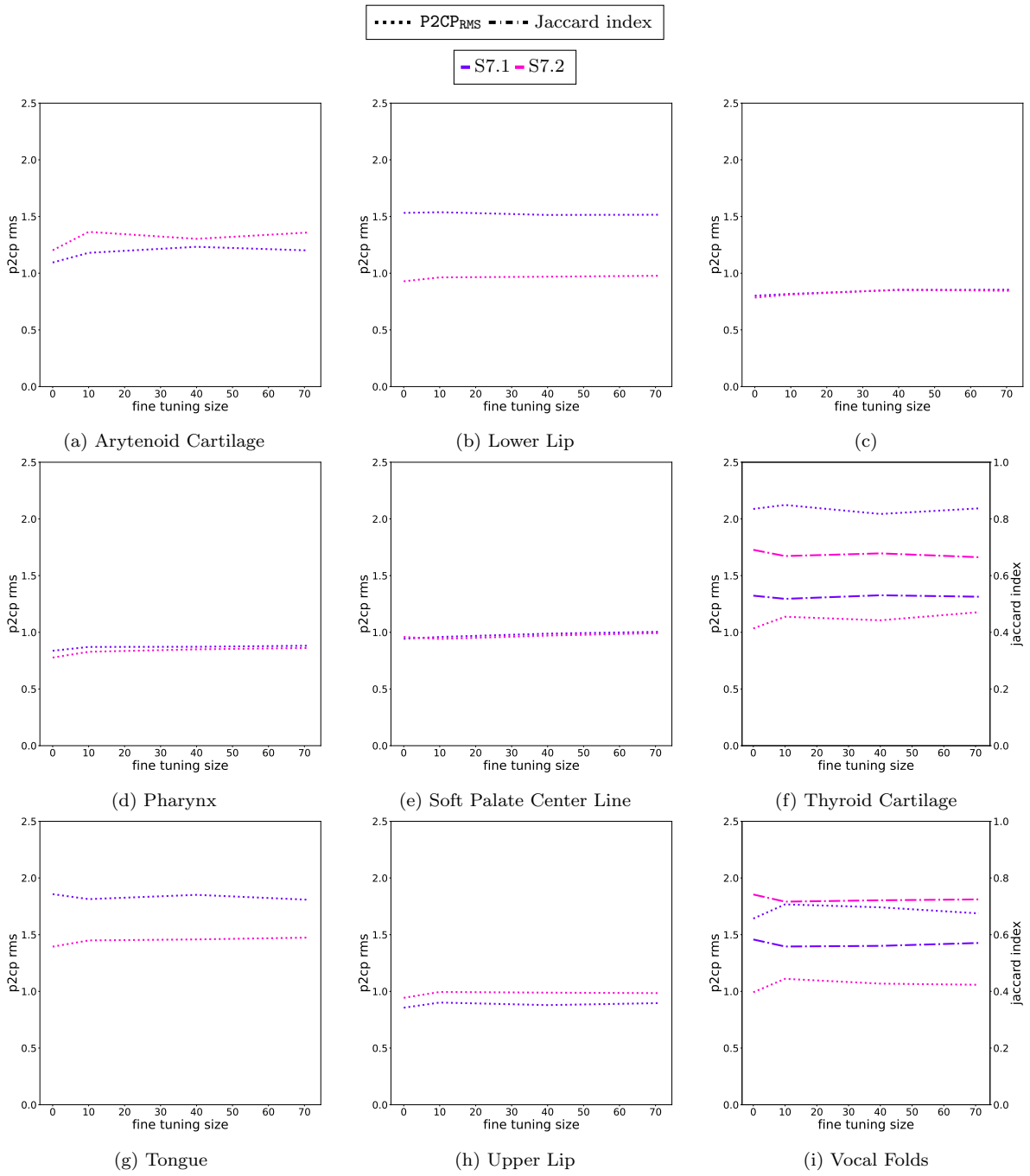


(a) Arytenoid Cartilage (b) Epiglottis (c) Lower Lip

(d) Pharynx (e) Soft Palate Center Line (f) Thyroid Cartilage

(g) Tongue (h) Upper Lip (i) Vocal Folds

Figure 10: P2CP_{RMS} and Jaccard index (only for closed contours) calculated on the test sets of the left-out subjects for each articulator when the models were adapted with varying numbers of training samples of the left-out subject. The colored lines represent the results for each test subject, the pink (and blue) solid lines represent the mean metric, and the shaded regions represent one standard deviation.



(a) Arytenoid Cartilage (b) Lower Lip (c)

(d) Pharynx (e) Soft Palate Center Line (f) Thyroid Cartilage

(g) Tongue (h) Upper Lip (i) Vocal Folds

Figure 11: $\texttt{P2CP}_{\texttt{RMS}}$ and Jaccard index (only for closed contours) calculated on the test sets of S7.1 and S7.2 for each articulator when the models were adapted with a varying number of training samples of the left-out subject. The colored lines represent the results for each test subject.

Probable explanation of this phenomenon is that the model privileged the tongue contour selected by other annotators (who worked on other images of the dataset). These discrepancies are usual in machine learning. The performance metrics are calculated in reference to a human annotator. However, even if different specialists provide their annotations for the same images, the annotations are likely to differ. This effect is known as inter-annotator agreement, a common phenomenon in image segmentation tasks [50]. A slight deviation from the ground truth is acceptable, but the hypothetical inter-annotator agreement should constrain it. Otherwise, the model would copy a specific specialist instead of learning the task. However, an extended analysis of the vocal tract segmentation inter-annotator agreement is beyond the scope of this research.

The predictions for the larynx articulators (arytenoid cartilage, epiglottis, thyroid cartilage, and vocal folds) are adequate, which is encouraging since this is a challenging region for human annotators. The most significant errors for the epiglottis occur at its extremities, as seen in Figure 5f. The prediction is correctly located for the thyroid cartilage and the vocal folds, and the errors are related to the size of the articulator; the algorithm usually yields a larger area. However, for articulatory speech research, the correct location of vocal folds is much more important than the precise contour since it is the source of the voice and directly impacts synthesis and control models.

For S5, the head is slightly rotated, as the subject seems to be leaning upwards, contrary to the others leaning forward. This case differs from a simple image rotation, which data augmentation could easily handle. The head rotation produces a slight deformation in the larynx, which is more pronounced in the thyroid cartilage and vocal folds region. For this reason, the model struggles to accurately predict their shapes when subject S5 is left out of the training set (Table 7). On the other hand, when the model is adapted to it, the performance improvement is very noticeable.

The case when the contour's extremities accounts for the largest errors also affects other articulators such as the tongue, pharynx, and soft palate. The contour's extremities also

accounts for the largest errors for other articulators such as the tongue, pharynx, and soft palate. A few cases for the tongue are visible in Figure 5c and Figure 5e. Nevertheless, the difference is less acoustically relevant for most of these articulators since they occur in a region that does not alter the final vocal tract air column.

Figure 6 shows a few cases of swallowing, which led to the worst results. Swallowing is an essential human process but is also one of the most difficult to annotate and predict. It is because the epiglottis lays over the arytenoid cartilage during swallowing, covering the glottis and preventing anything but air from entering the lungs. It creates constrictions between the articulators and the bolus, removing air-tissue boundaries. As a result, the articulators are almost indistinguishable, making the annotation difficult even for specialists.

Not surprisingly, the model provides the most unreliable results for swallowing cases. The tongue and the epiglottis errors are around 3 mm, and the Jaccard index for the vocal folds and thyroid cartilage is low. In some cases, the model even misses some articulators completely. It would have been possible to obtain better results for swallowing by significantly increasing the number of swallowing examples in the training set. However, this is not the focus of our work, which is concerned with articulatory speech research. We leave this possibility of improving the results for swallowing for future work.

We compared our work to that of Labrunie et al. [24] due to the similarities we found between them. Labrunie et al. [24] used RT-MRI data and considered most of the articulators as we did, except for the thyroid cartilage and vocal folds. However, the two studies had substantial differences in the annotation decisions. Labrunie et al. [24] did not include the sublingual cavity, starting the tongue annotation from the tip, and they chose to annotate the contour of the epiglottis and soft palate, not only the center lines. Most importantly, the two works used different test sets. Despite these differences, the similarities between the two works allow for some level of comparison. For a fair benchmark, it is desirable to have access to the same images and a standardized annotation procedure.

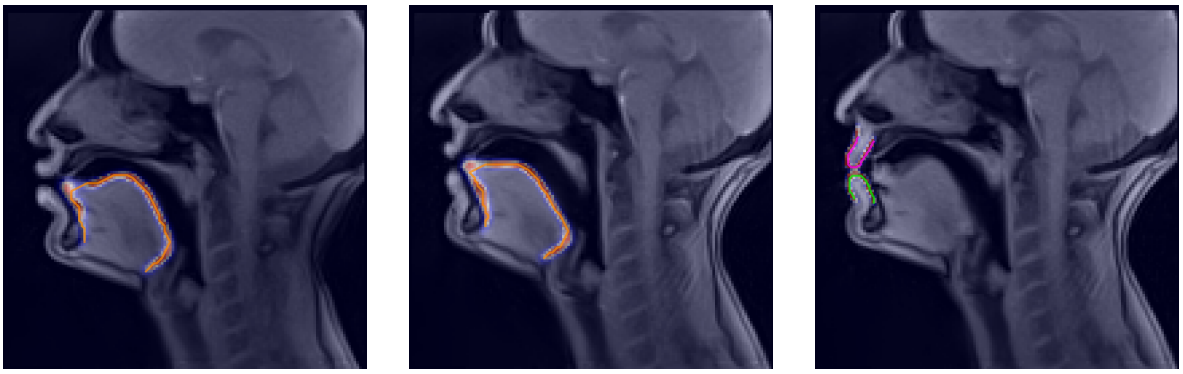
On average, our results are close to the Multiple Linear Regression (MLR) while under

performing compared to the best approach – the mASM. However, our method has the advantage of being speaker-independent by design, while the method proposed by Labrunie et al. [24] is subject-specific, limiting its usability for the community.

The second experiment evaluated the adaptation to an unseen speaker. The results from Figure 10 suggest that there is a significant improvement in the results with only ten additional training images when the model initially performed poorly, such as the vocal folds and thyroid cartilage for S5 and the pharynx for S6, even though on average the gain, if any, is minimal. The result indicates that adaptation is beneficial when the target subject has a more pronounced anatomical or postural differences from the training subjects. The adaptation gain is lower for cases where the target subject is standardized, such as the same head position.

During the speaker adaptation procedure, the model could specialize in the new subject and "forget" the previous ones. This phenomenon is known in the machine learning literature as catastrophic forgetting. It could be avoided using elastic weight consolidation [51]. However, the results of the second experiment shown in Figure 11 suggest that the model does not forget the previous subjects. It is likely because the speaker adaptation procedure was conservative, using a lower learning rate for a few epochs, preventing the model from diverging but also limiting the improvement.

It is essential to note that our method has limitations. The main one is not including rigid articulators, such as the jaw, the upper incisor, and the hard palate. These articulators are indispensable for speech production but are challenging to segment in MRI because bones are indistinguishable from the air in the image, so we can only observe the teeth root trace, which contains a small amount of water. Since these articulators are rigid, the problem is restricted to finding their location in the image. Therefore, an alternative for tracking these articulators would be sliding a pre-computed mask in the image and retrieving the image region with the highest correlation with the mask. However, it can be laborious by requiring several manual adjustments in the tracking.



(a) (b) (c)

Figure 12: MRI samples illustrating cases in which the model failed to predict the contact between articulators. The images include the original MRI sample, the predicted segmentation mask, and the contour without b-split regularization.

Furthermore, it is difficult to track contacts between articulators, especially for the tongue tip, as evidenced by the samples in Figure 12 and the videos in the supplementary material. Figure 12 shows MRI samples with a segmentation mask overlay and the shortest path contour without b-split regularization. Although the most extreme point in the tongue tip might have a higher probability, it corresponds to a much longer path, which the shortest path algorithm rejects; thus, the contact between the tongue tip and the alveolar region is missed.

An alternative is to give a higher weight to points closer to the alveolar region. Nevertheless, it requires locating the hard palate and upper incisor previously to the contour computation. Another alternative is training a neural network to post-process the segmentation mask, learning which is the relevant contour given the pixel probabilities.

Finally, another limitation is the RT-MRI frame rate. Our method was trained with a frame rate of 50 frames per second, meaning that each frame corresponds to 20 milliseconds of speech. However, some phonemes, such as /l/ have a shorter duration than the MRI frame, which means the image will be blurred, and the true articulator position will be uncertain. In these cases, the performance of the model will be lower.

6. Conclusions

In this study, we proposed and assessed a method for segmenting the vocal tract shape in real-time MRI using a deep learning approach. We also proposed a transfer learning method operating with a small dataset (in comparison to other deep learning applications). The method accurately estimated the shapes of nine non-rigid articulators that delimit the vocal tract tube and are essential for articulatory speech synthesis. We also showed that the model can generalize to new subjects. If the position of a new subject significantly differs from the previous set, the model may need to be adapted to the target subject. Nevertheless, we showed that only a small amount of data (about ten images) is required for a good adaptation.

Our recent research in articulatory synthesis of speech has shown that the method is helpful for this task [4, 5, 52]. We plan to address some of the method's limitations in future work, such as the difficulty of segmenting contact regions, especially the tongue tip. Also, our source code is available as a `Python` package that makes exploring RT-MRI data to investigate speech production easy. The package only covers the soft articulators, but we plan to extend it to rigid bodies to complete the vocal tract shape. Finally, `ASD1` images are already available [38]. We plan to publicly release the `ASD2` dataset together with the manual and automatic annotations for both datasets.

We hope this work will be helpful for other researchers interested in investigating speech production using RT-MRI data.

7. Acknowledgements

This research was supported by the French ANR project Full3DTalkingHead, CPER IT2MP, Région Lorraine and FEDER.

References

[1] Peter Ladefoged and Ian Maddieson. The sounds of the world's languages. *Language*, 74(2):374–376, 1998.

[2] Catherine P. Browman and Louis Goldstein. Articulatory phonology: An overview. *Phonetica*, 49(3-4):155–180, 1992.

[3] Pierre Badin, Gerard Bailly, Lionel Reveret, Monica Baciu, Christoph Segebarth, and Christophe Savariaux. Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics*, 30(3):533–553, 2002.

[4] Vinicius Ribeiro, Karyna Isaieva, Justine Leclere, Pierre-André Vuissoz, and Yves Laprie. Towards the Prediction of the Vocal Tract Shape from the Sequence of Phonemes to be Articulated. In *Proc. Interspeech 2021*, pages 3325–3329, 2021.

[5] Vinicius Ribeiro, Karyna Isaieva, Justine Leclere, Pierre-André Vuissoz, and Yves Laprie. Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated. *Speech Communication*, 2022. ISSN 0167-6393.

[6] Samuel Silva and António Teixeira. Critical Articulators Identification from RT-MRI of the Vocal Tract. In *Proc. Interspeech 2017*, pages 626–630, 2017.

[7] Shinji Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech production and speech modelling*, pages 131–149. Springer, 1990.

[8] Britta Grimme, Susanne Fuchs, Pascal Perrier, and Gregor Schöner. Limb versus speech motor control: A conceptual review. *Motor control*, 15(1):5–33, 2011.

[9] Kenneth L Moll. Cinefluorographic techniques in speech research. *Journal of Speech and Hearing Research*, 3(3):227–241, 1960.

[10] Gunnar Fant. Acoustic theory of speech production. Mouton, The Hague, 1960.

[11] Sarah N Dart. A bibliography of x-ray studies of speech. *UCLA Working Papers in Phonetics*, 66:1–97, 1987.

[12] Teja Rebernik, Jidde Jacobi, Roel Jonkers, Aude Noiray, and Martijn Wieling. A review of data collection practices using electromagnetic articulography. *Laboratory Phonology*, 12(1), 2021.

[13] Michael Proctor, Erik Bresch, Dani Byrd, Krishna Nayak, and Shrikanth Narayanan. Paralinguistic mechanisms of production in human "beatboxing": A real-time magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, 133(2): 1043–1054, 2013.

[14] Yinghua Zhu, Asterios Toutios, Shrikanth Narayanan, and Krishna Nayak. Faster 3D vocal tract real-time MRI using constrained reconstruction. In *Proc. Interspeech 2013*, pages 1292–1296, 2013.

[15] Ziwei Zhao, Yongwan Lim, Dani Byrd, Shrikanth Narayanan, and Krishna S Nayak. Improved 3D real-time MRI of speech production. *Magnetic Resonance in Medicine*, 85 (6):3182–3195, 2021.

[16] Sajan Goud Lingala, Brad P Sutton, Marc E Miquel, and Krishna S Nayak. Recommendations for real-time speech MRI. *Journal of Magnetic Resonance Imaging*, 43(1): 28–44, 2016.

[17] Donald G. Miller, Arend M. Sulter, Harm K. Schutte, and Rienhart F. Wolf. Comparison of vocal tract formants in singing andnonperiodic phonation. *Journal of Voice*, 11(1): 1–11, 1997.

[18] Peter W. Iltis, Jens Frahm, Dirk Voit, Arun A. Joseph, Erwin Schoonderwaldt, and Eckart Altenmüller. High-speed real-time magnetic resonance imaging of fast tongue

movements in elite horn players. *Quantitative imaging in medicine and surgery*, 5(3): 374, 2015.

[19] Zeynab Raeesy, Sylvia Rueda, Jayaram K. Udupa, and John Coleman. Automatic segmentation of vocal tract MR images. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 1328–1331. IEEE, 2013.

[20] Sylvia Rueda and Jayaram K. Udupa. Global-to-local, shape-based, real and virtual landmarks for shape modeling by recursive boundary subdivision. In *Medical Imaging 2011: Image Processing*, volume 7962, pages 1329–1341. SPIE, 2011.

[21] Jiamin Liu and Jayaram K. Udupa. Oriented active shape models. *IEEE Transactions on medical Imaging*, 28(4):571–584, 2008.

[22] Samuel Silva and António Teixeira. Unsupervised segmentation of the vocal tract from real-time MRI sequences. *Computer Speech & Language*, 33(1):25–46, 2015.

[23] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In *European conference on computer vision*, pages 484–498. Springer, 1998.

[24] Mathieu Labrunie, Pierre Badin, Dirk Voit, Arun A. Joseph, Jens Frahm, Laurent Lamalle, Coriandre Vilain, and Louis-Jean Boë. Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning. *Speech Communication*, 99:27–46, 2018.

[25] S Suganyadevi, V Seethalakshmi, and K Balasamy. A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval*, 11(1):19–38, 2022.

[26] Valliappan Ca, Renuka Mannem, and Prasanta Kumar Ghosh. Air-Tissue Boundary Segmentation in Real-Time Magnetic Resonance Imaging Video Using Semantic Seg-

mentation with Fully Convolutional Networks. In *Proc. Interspeech 2018*, pages 3132–3136, 2018.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[28] Ian Fasel and Jeff Berry. Deep belief networks for real-time extraction of tongue contours from ultrasound during speech. In *2010 20th International Conference on Pattern Recognition*, pages 1493–1496. IEEE, 2010.

[29] Aurore Jaumard-Hakoun, Kele Xu, Pierre Roussel-Ragot, Gérard Dreyfus, and Bruce Denby. Tongue contour extraction from ultrasound images based on deep neural network. *arXiv preprint arXiv:1605.05912*, 2016.

[30] Nan Zhang, Shifei Ding, Jian Zhang, and Yu Xue. An overview on restricted boltzmann machines. *Neurocomputing*, 275:1186–1199, 2018.

[31] Mohammad Eslami, Christiane Neuschaefer-Rube, and Antoine Serrurier. Automatic vocal tract segmentation based on conditional generative adversarial neural network. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pages 263–270, 2019.

[32] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[33] Sasan Asadiabadi and Engin Erzin. Vocal tract contour tracking in rtMRI using deep temporal regression network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3053–3064, 2020.

[34] S. Ashwin Hebbar, Rahul Sharma, Krishna Somandepalli, Asterios Toutios, and Shrikanth Narayanan. Vocal tract articulatory contour detection in real-time magnetic resonance images using spatio-temporal context. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7354–7358. IEEE, 2020.

[35] Matthieu Ruthven, Marc E Miquel, and Andrew P King. Deep-learning-based segmentation of the vocal tract and articulators in real-time magnetic resonance images of speech. *Computer Methods and Programs in Biomedicine*, 198:105814, 2021.

[36] Karyna Isaieva, Yvez Laprie, Nicolas Turpault, Alexis Houssard, Jacques Felblinger, and Pierre-André Vuissoz. Automatic tongue delineation from MRI images with a convolutional neural network approach. *Applied Artificial Intelligence*, 34(14):1115–1123, 2020.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[38] Karyna Isaieva, Yves Laprie, Justine Leclère, Ioannis K Douros, Jacques Felblinger, and Pierre-André Vuissoz. Multimodal dataset of real-time 2D and static 3D MRI of healthy french speakers. *Scientific Data*, 8(1):258, 2021.

[39] Martin Uecker, Shuo Zhang, Dirk Voit, Alexander Karaus, Klaus-Dietmar Merboldt, and Jens Frahm. Real-time MRI at a resolution of 20 ms. *NMR in Biomedicine*, 23(8): 986–994, 2010.

[40] Yves Laprie, Benjamin Elie, Anastasiia Tsukanova, and Pierre-André Vuissoz. Centerline articulatory models of the velum and epiglottis for articulatory synthesis of speech. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2110–2114. IEEE, 2018.

[41] Errol M. Bellon, E. Mark Haacke, Paul E. Coleman, Damon C. Sacco, David A. Steiger, and Raymond E. Gangarosa. MR artifacts: A review. *AJR. American journal of roentgenology*, 147:1271–81, 12 1986.

[42] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[43] Priyanka Malhotra, Sheifali Gupta, Deepika Koundal, Atef Zaguia, and Wegayehu Enbeyle. Deep neural networks for medical image segmentation. *Journal of Healthcare Engineering*, 2022, 2022.

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[45] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[46] Edsger W. Dijkstra et al. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[47] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[48] Leslie N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmai-

son, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[50] Vinicius Ribeiro, Sandra Avila, and Eduardo Valle. Handling inter-annotator agreement for automated skin lesion segmentation. *arXiv preprint arXiv:1906.02415*, 2019.

[51] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[52] Vinicius Ribeiro and Yves Laprie. Autoencoder-Based Tongue Shape Estimation During Continuous Speech. In *Proc. Interspeech 2022*, pages 86–90, 2022.