Tracking the tongue contours in rt-MRI films with an autoencoder DNN approach

Karyna Isaieva[a], Yves Laprie[b], Alexis Houssard[b], Jacques Felblinger[a,c] and Pierre-André Vuissoz[a]
karyna.isaieva@univ-lorraine.fr

[a]*INSERM, IADI, Université de Lorraine, Nancy, France;* [b]*CNRS, Inria, LORIA, Université de Lorraine, Nancy, France;* [c]*CIC-IT, INSERM, CHRU de Nancy, Nancy, France*

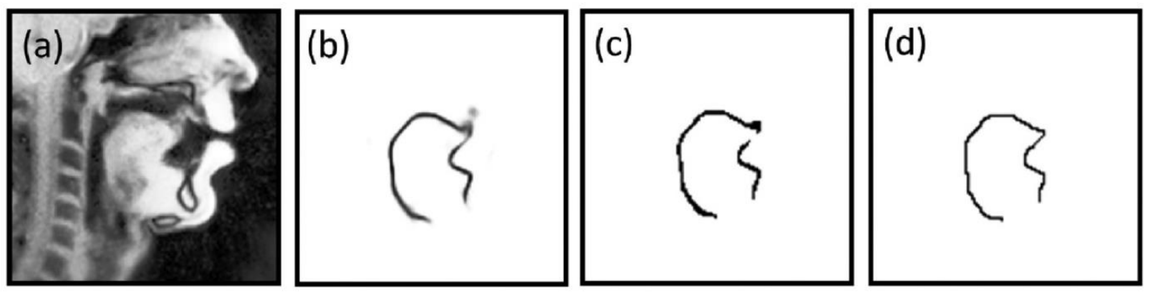
Introduction. Nowadays, magnetic resonance imaging (MRI) represents a non-invasive safe method of the speech imaging. Modern technologies allow imaging of articulators in mid-sagittal plane with temporal resolution of 20 ms [1]. However, tongue contour extraction from real-time magnetic resonance images is a nontrivial task due to the presence of artifacts manifesting in form of blurring or ghostly contours. Blurring mainly appears because of the relatively large slice thickness (generally 8 mm) resulting in a partial volume effect, i.e. the fact that one part of the slice volume corresponds to flesh and the other to air, especially when the tongue groove is marked. Moreover, displacement of the articulators during the acquisition of each image leads to motion artifacts and the undersampling which speeds up the acquisition causes some specific artifacts. Thus, some interpretation work is required. There are several works devoted to similar tasks. In [2] articulators' delineation was done with help of several supervised learning algorithms. In [3] air-tissue boundary segmenation was done with U-Net convolutional network. However, classic segmenation where the result of prediction is represented by a mask, required some post-processing to extract contours which are not mandatory closed. In our study, we concentrate on the extraction of the tongue contours from real-time MR images and the corresponding learning scheme and overall strategy to easily exploit results as curves and not masks. The question addressed in this work is whether learning can be designed to use 1-pixel wide contours as a segmentation mask and then lead to an acceptable delineation quality with very few spurious contour points. In our case, we chose U-Net auto-encoding CNN [4] which turned out to be very efficient for biomedical images.

Materials and methods. We used real-time images from the ArtSpeechMRIfr database [5]. The subjects are two native French speakers, both males of 35 and 32 years which will be denoted below as S1 and S2, respectively. The images were acquired with a Siemens Prisma-fit 3 T scanner (Siemens, Erlangen, Germany). We used the radial RF-spoiled FLASH sequence [1] with the slice thickness of 8 mm and image resolution of 136 × 136. The acquisition time varied from 34 sec to 90 sec, mostly about 60 sec. We followed the protocol described in [6]. Images were recorded at a frame rate of 55 frames per second and reconstructed with the algorithm presented in [1].

For our analyses, we consider a subset of 600 images of speaker S1. They include 400 consecutive images corresponding to one sentence and some manually selected non-similar images of the same speaker. Also, we took 100 random images of speaker S2. All images were delineated manually. The data of S1 were divided into training, validation, and testing sets (400, 100, and 100 images, respectively), and the images of S2 were used only for testing. We performed 6-fold cross-validation test.

We used the Keras framework for the U-Net implementation. The initial size of the images was 128 × 128. The sample weights were set at 0.8 and 0.2 for the tongue contour class and noncontour class, respectively. The batch size was 8 and the number of epochs was defined automatically by early stopping with a patience of 10.

We developed a post-processing algorithm to extract the 1-pixel width contours from the predicted probability maps. The two contour extremities are found as two adjacent contour points having the greatest angular distance with respect to the contour gravity center. Then a graph was constructed by connecting the points whose distance was less than the distance between the two extremities minus 1 pixel. Applying Dijkstra's shortest path search algorithm between the two extremities with the quadratic distance as a cost provides the tongue contour.



we used the Mean Sum of Distance metrics (MSD) to evaluate our results:

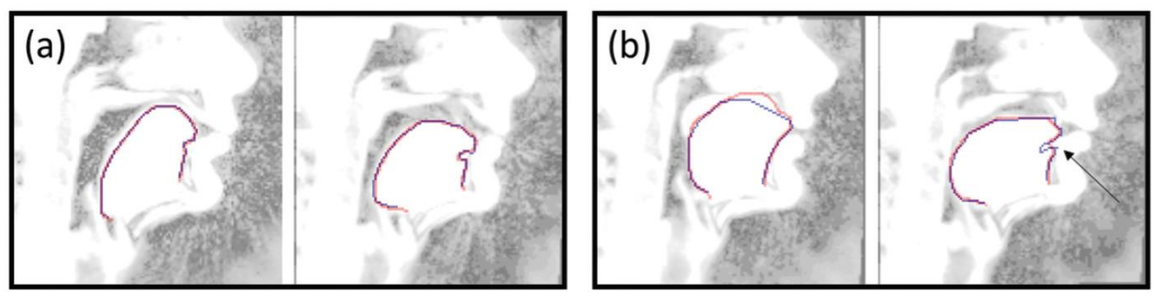
$$MSD(U,V) = \frac{1}{n1+n2}\left(\sum_{i=1}^{n1}\min_j|u_i - v_j| + \sum_{i=1}^{n2}\min_j|u_j - v_i|\right),$$

where $u_i \in U$ and $v_i \in V$ are ground truth and predicted curves, n_1 and n_2 is a total number of points in corresponding curves.

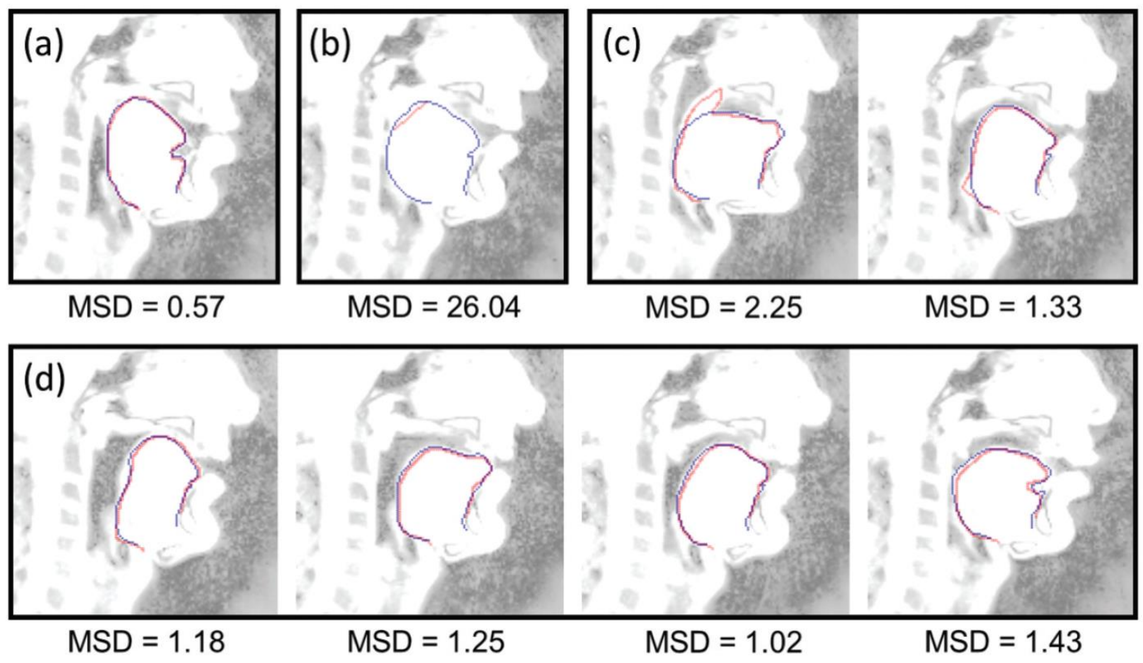
Results. The resulting mean squared distances and corresponding standard deviations between manually delineated and predicted contours for all iterations of sixfold cross-validation are presented in the table.

# iteration	Epoch number	MSD valid (mm)	MSD test (mm)	MSD S2 (mm)	Mean MSD (mm)
1	32	0.63 ± 0.21	0.62 ± 0.17	1.29 ± 0.36	0.92 ± 0.43
2	35	0.64 ± 0.24	0.62 ± 0.21	1.13 ± 0.32	0.88 ± 0.38
3	56	0.62 ± 0.20	0.60 ± 0.16	1.16 ± 0.39	0.88 ± 0.41
4	27	0.66 ± 0.24	0.66 ± 0.25	1.40 ± 2.49	1.04 ± 1.82
5	54	0.64 ± 0.21	0.64 ± 0.18	1.16 ± 0.29	0.91 ± 0.36
6	29	0.63 ± 0.17	0.64 ± 0.21	1.12 ± 0.32	0.88 ± 0.35

In general, intra-subject prediction proved to be very good since the error is less than 0.7mm. The best results are presented on the figure (a) and the problematic cases are shown on the figure (b).



Less accuracy was expected for the inter-subject prediction. The best result is shown on the figure (a), the worst is given on the figure (b), the problematic cases are illustrated on (c) and some typical results are presented on (d).



Conclusions. This work shows that a very good delineation accuracy, i.e. less than 0.7mm (for intra-speaker validation), can be achieved for the tongue contour with a rather limited training size if we consider that only 600 images. This approach thus shows promising results and slightly outperforms existing methods despite the presence of multiple image artifacts.

Also, we took care to control the learning process in a way to get 1-pixel wide contours and very few spurious points.

The tests carried out on the second speaker showed that the average accuracy is slightly less (1.21 mm) but still quite acceptable. Further research will focus the other speech articulators and algorithms intended to minimize the set of images that have to be hand labeled.

The work is published in Applied Artificial Intelligence journal [7].

The mean MSD value for intra-subject prediction was 0.63 mm which slightly outperform existing methods of the tongue delineation. In [2], the best result for the tongue delineation was MSD = 0.68 mm for prediction by modified active shape models.

Acknowledgments. We thank Arun A. Joseph, Dirk Voit, and Jens Frahm for their help with the data acquisition; Hamza Taybi for his help with the manual contour delineation. Research supported by the project ArtSpeech of ANR (Agence Nationale de la Recherche), France, CPER "IT2MP", "LCHN" and FEDER.

References

1. Uecker, M., S. Zhang, D. Voit, A. Karaus, K.-D. Merboldt, and J. Frahm. 2010. Real-time MRI at a resolution of 20 Ms. *NMR in Biomedicine* 23 (8):986–94. Wiley Online Library.
2. Labrunie, M., P. Badin, D. Voit, A. A. Joseph, J. Frahm, L. Lamalle, C. Vilain, and L.-J. Boe. 2018. Automatic segmentation of speech articulators from real-time Midsagittal MRI based on supervised learning. *Speech Communication* 99 Elsevier:27–46.
3. Somandepalli, K., A. Toutios, and S. S. Narayanan. 2017. Semantic edge detection for tracking vocal tract air-tissue boundaries in real-time magnetic resonance images. *Interspeech*, Stockholm, Sweden, 631–35.
4. Ronneberger, O., P. Fischer, and T. Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–41. Cham: Springer International Publishing.
5. Douros, I., J. Felblinger, J. Frahm, K. Isaieva, A. Joseph, Y. Laprie, F. Odille, A. Tsukanova, D. Voit, and P.-A. Vuissoz. 2019. A multimodal real-time MRI articulatory corpus of french for speech research. In *Proceedings of Interspeech 2019,* 1556-1560. Graz, Austria.
6. Niebergall, A., S. Zhang, E. Kunay, G. Keydana, M. Job, M. Uecker, and J. Frahm. 2013. Realtime MRI of speaking at a resolution of 33 Ms: Undersampled radial FLASH with nonlinear inverse reconstruction. *Magnetic Resonance in Medicine* 69 (2):477–85. Wiley Online Library.
7. Isaieva, K., Laprie, Y., Turpault, N., Houssard, A., Felblinger, J., & Vuissoz, P. A. (2020). Automatic Tongue Delineation from MRI Images with a Convolutional Neural Network Approach. *Applied Artificial Intelligence*, 1-9.