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

Karyna Isaeva[(1)], Yves Laprie[(2)], Alexis Houssard[(2)], Jacques Felblinger[(1)], Pierre-André Vuissoz[(1)]

[(1)]IADI, France, karyna.isaieva@univ-lorraine.fr

[(2)]Loria, France, yves.laprie@loria.fr

1 Introduction

Speech production is an eminently dynamic process and the study of articulatory gestures is therefore a central research topic. For this reason many dynamic data acquisition devices have been developed, including electromagnetic articulography and ultrasound. Only X-rays and magnetic resonance provides a global view of the vocal tract. Unlike X-rays, magnetic resonance imaging does not present any health hazard for the subjects and is therefore an essential tool. More recently, dynamic MRI has appeared and offers a high acquisition rate.

Raw images cannot be exploited easily and for this reason it is necessary to extract the contours of the articulators from images. A great deal of efforts [5, 2] has been put into designing automatic tracking algorithms for X-ray image processing, but the poor quality of those images has never led to acceptable results, and the tracking has often been done by hand. This acceptable solution for X-ray images which are in small numbers is not at all acceptable for MRI images which are available in very large numbers.

Unlike X-ray images, there is no superimposition of organs in MRI images, which are real slices of a small thickness. In our case, we use 8 mm slices acquired at 50 Hz on a Siemens Prisma machine with the radial FLASH MRI sequence [6] from ArtSpeechMRIfr database [1]. The contouring is therefore easier, but there are blurring or ghosting effects because of displacement of the articulators during acquisition of each image. This means that a simple contour extraction based on a gradient, for instance, is insufficient since some interpretative work is required, especially when the tongue tip is rapidly approaching the teeth to articulate a dental sound. A learning method is therefore a natural choice and we have used an auto-encoding approach to address this problem. In our case we chose U-Net DNN [4] which turned out to very efficient for biomedical images.

2 Implementation

We chose 609 non-similar images for the training set and 100 random images for the testing set. For the tongue contour detection we chose a variation of U-Net architecture proposed in [8]. Each image from the training and the test sets was annotated by a human, and mask was chosen to be a curve of a single pixel width. Loss function was the weighted binary cross-entropy [7]. The model was trained with Adam optimizer with learning rate of 0.0005 and batch size of 8. We created check-points each 250 training iterations and we found that 6250 iterations give the best delineation. All hyper parameters were defined from the training set itself using Dice coefficients as the metrics.

The output of tracking is a binary image, i.e. an unorganized cloud of points, which has to be post-processed to get true contours. To connect some gaps and exclude some overlapping parts in the resulting contours we thus developed an algorithm based on Dijktsra's shortest path search on a weighted graph. For evaluation of the results we used the Mean Sum of Distance metrics (MSD) [3].

3 Results and discussion

In general, tongue contours produced by the model have a high accuracy. Our preliminary calculations shown that MSD for intra-subject precision was less than 1 mm. Examples of predicted contours are shown on Fig. 1 (see also https://artspeech.loria.fr/resources/).

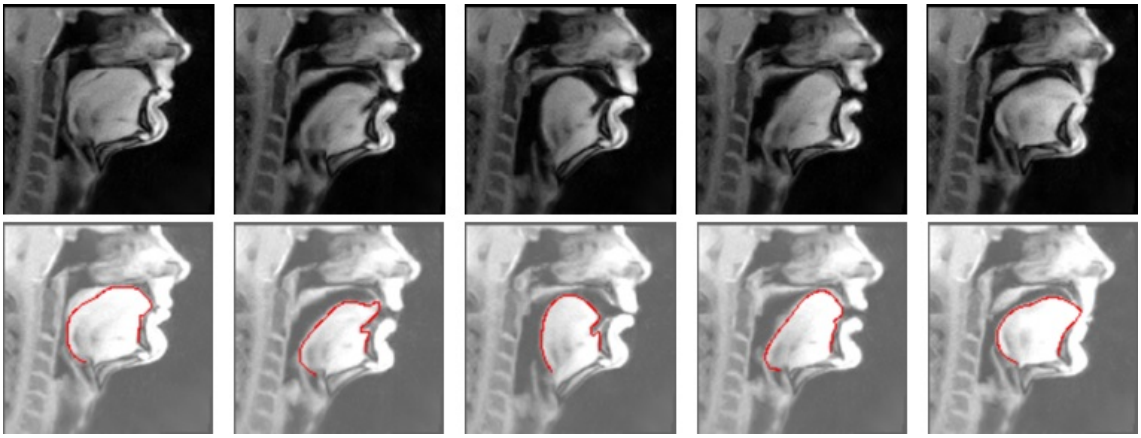

Figure 1. Results of automatic tongue delineation. Top row represents original images, bottom row shows the same images with the automatically detected tongue contour

Despite the model appeared to deal well with the MRI artifacts (ghost edges close to the middle part of the tongue surface, motion artifacts manifesting in form of blurring, curly edges), in some cases predicted edges were somewhat distorted by presence of artifacts. Also, in few cases the model failed to find the edge between the tongue and other articulators when too tight contact had place. Inter-subject prediction was evaluated visually. Its results are acceptable but require further training with additional images especially when there is a contact between the tongue and the pharyngeal wall, the palate or teeth.

References

[1] Ioannis Douros et al. "A Multimodal Real-Time MRI Articulatory Corpus of French for Speech Research". In: *Interspeech*. 2019.

[2] J. Fontecave Jallon and F. Berthommier. "A Semi-Automatic Method for Extracting Vocal-Tract Movements from X-Ray Films". In: *Speech Communication* 51.2 (2009), pp. 97–115.

[3] Min Li, Chandra Kambhamettu, and Maureen Stone. "Automatic contour tracking in ultrasound images". In: *Clinical linguistics & phonetics* 19.6-7 (2005), pp. 545–554.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab et al. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4.

[5] Georg Thimm. "Tracking Articulators in X-ray Movies of the Vocal Tract". In: *Computer Analysis of Images and Patterns*. Ed. by Franc Solina and Alešs Leonardis. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 126–133. ISBN: 978-3-540-48375-5.

[6] Martin Uecker et al. "Real-time MRI at a resolution of 20 ms". In: *NMR in Biomedicine* 23.8 (2010), pp. 986–994.

[7] Saining Xie and Zhuowen Tu. "Holistically-nested edge detection". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1395–1403.

[8] Jian Zhu, Will Styler, and Ian Calloway. "A CNN-based tool for automatic tongue contour tracking in ultrasound images". In: *arXiv preprint arXiv:1907.10210* (2019).