AUTOMATIC VOCAL TRACT SEGMENTATION BASED ON CONDITIONAL GENERATIVE ADVERSARIAL NEURAL NETWORK

Mohammad Eslami[1], *Christiane Neuschaefer-Rube*[1], *Antoine Serrurier*[1]

[1] *Clinic for Phoniatrics, Pedaudiology & Communication Disorders, University Hospital and Medical Faculty of the RWTH Aachen University, Germany*
meslami@ukaachen.de

Abstract: Speech production is characterized by high articulatory variability both in the space and the time domains. MRI in the last decades and real-time MRI more recently have proved to be particularly adapted to study the speech articulations. The data generated require however a high amount of processing but the quantity generated exclude manual processing and call for automatic segmentation methods. Nowadays, deep learning shows very promising results in many aspects of image processing problems including segmentation. In this paper, the segmentation of the jaw, the tongue and the vocal tract are explored based on a modified version of the pix2pix algorithm, taking advantage of the conditional generative adversarial networks. The experimental results are evaluated via a leave-one-out cross-validation scheme on midsagittal static MRI images of 10 subjects sustaining 62 different articulations. Both qualitative and quantitative assessments of the proposed method show promising and reliable performance and open the way for possible future works in speech articulatory modelling.

1 Introduction

Speech production is characterized by high articulatory variability both in the space and the time domains. Modelling the articulators of the vocal tract require therefore high amount of data and heavy processing. Magnetic Resonance Imaging (MRI) has proved in the last decades to be particularly adapted to articulatory modelling as an harmless and non-invasive imaging techniques providing images of good quality and resolution. A recent development of this modality, real-time MRI, allows the acquisition at a very high frame rate, making this technique particularly adapted for articulatory modelling [1, 2]. The amount of images generated excludes however manual segmentation of the speech articulators and calls for automatic or semi-automatic methods [2]. These methods have to deal with the high intra- and inter-speaker articulatory variability observed in practise.

Over the years, many studies have proposed different automatic and semi-automatic methods to segment the vocal tract and its surrounding articulators. Labrunie *et al.* [2] recorded a large database of real-time MRI midsagittal images for a French speaker. Their method involves training various contour models based on multiple linear regression and shape particle filtering. The technique proposed by Sampaio *et al.* [3] is based on level-set-methods. The recent rise of neural networks and deep learning techniques benefited also to the automatic segmentation of the vocal tract and speech articulators. The technique proposed by Valliappan *et al.* [4] makes use of semantic segmentation based on a deep learning architecture called fully convolutional networks with additional post-processing to enhance the results. Lastly, a convolutional neural network with an encoder-decoder architecture combined with post-processing is used to jointly detect the vocal tract air-tissue boundaries and label them in [5].

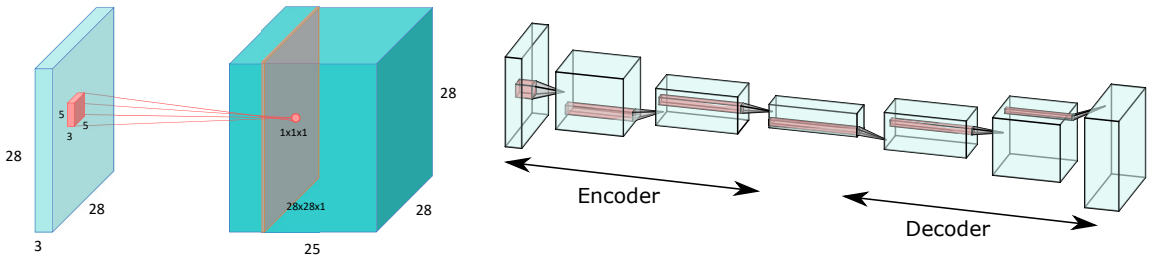


Figure 1 – Schematic representations of a CNN layer (left) and an auto-encoding architecture based on CNN (right).

So far, until now, most of the existing methods does not achieve a fully automatic segmentation of the vocal tract and its surrounding articulators which have no or very limited manual initialization and bring a high accuracy. The ambition of this paper is to rely on end-to-end training and recent deep learning segmentation techniques to move towards this objective. More specifically, deep learning and convolutional neural network methods used in adversarial schemes show very promising results [6]. In our case, we aim at modifying and training a conditional Generative Adversarial Neural Network (cGAN) to segment the jaw, the tongue and the vocal tract on MRI images as preliminary analyses.

The rest of this paper is as follow. Section 2 describes the background for segmentation via deep learning and adversarial schemes. The method and the data used for validation are presented in section 3. Results and discussion are presented in section 4 and concluding words are in section 5.

2 Segmentation based on neural networks

Neural networks are effective to process one dimensional data but show limitations for higher dimensions. Convolutional Neural Networks (CNNs) have been proposed to overcome this issue. Similar to neural networks, CNNs perform for each neuron the dot product of the output of the previous layer with the current neuron weights, add the bias, and apply a linear or non-linear activation function to the result. However, unlike a regular neural network, the neurons of a CNN are arranged in 3 dimensions, representing the height, width and depth of an image. Every layer of a CNN transforms the 3D input volume to a 3D output volume via filtering with a 3D volume filter (also called kernel) and applying an activation function. The depth of the output volume corresponds to the number of different filters in the layer and the filter or kernel values are the unknown weights which should be trained by the system. Figure 1a shows a convolutional layer with 25 filters/kernels of size 5x5x3 and used on an input of size 28x28x3.

Deep learning techniques and specially CNNs proved to be very efficient for high-level computer vision problems such as object detection and classification. It has been made possible by their end-to-end training fashion which learns optimized features instead of using hand-crafted features [7]. Such networks have therefore been extended to solve other problems such as semantic segmentation, which is a pixel-wise labelling problem. For that purpose, the CNNs are usually embedded in an auto-encoding architecture. In this architecture, the input image is 'encoded' step-by-step to extract the representational features and these features are 'decoded' step-by-step to create the output image containing the pixel-wise labels. A schematic representation is proposed in Figure 1b.

GANs and cGANs have been under attention in recent years because of their promising performance in image generation, segmentation and translation [6]. As an illustration, an approach for segmentation and conversion of clothes in images [8], a real time approach for segmentation

of insulators at a pixel level [9] or an approach for road detection and segmentation [10] have been proposed based on GANs. GANs are characterized by an adversarial scheme [6] where two 'adversarial' networks are trained in competition: one *generator* aiming at generating an image fooling a discriminator and one *discriminator* aiming at discriminating a real from a fake image.

More specifically, the generator of a GAN learns a mapping between a random noise vector z and an output image y, *i.e.* $G : z \Rightarrow y$. In conditional mode (cGAN), the generator is additionally conditioned on ground truth labels or images, *i.e.* it is constrained to generate images related to the conditions, for instance to generate the segmentation ground truth. In other words, the generator learns a mapping between an input image x and a random noise vector z to an output image y on the other hand, *i.e.* $G : \{x,z\} \Rightarrow y$. The generator G is trained to generate outputs that cannot be distinguished from "real" images while the discriminator D is trained to detect the generator's "fakes".

The objective of a cGAN, *i.e.* in a general optimization definition, can be expressed in equation (1), where $\mathbb{E}$ is the *Expectation* over population. G tries to minimize this objective against D which tries to maximize it, *i.e.* a *minimax* game as $\hat{G} = arg\ min_G\ max_D\ \mathscr{L}_{cGAN}$.

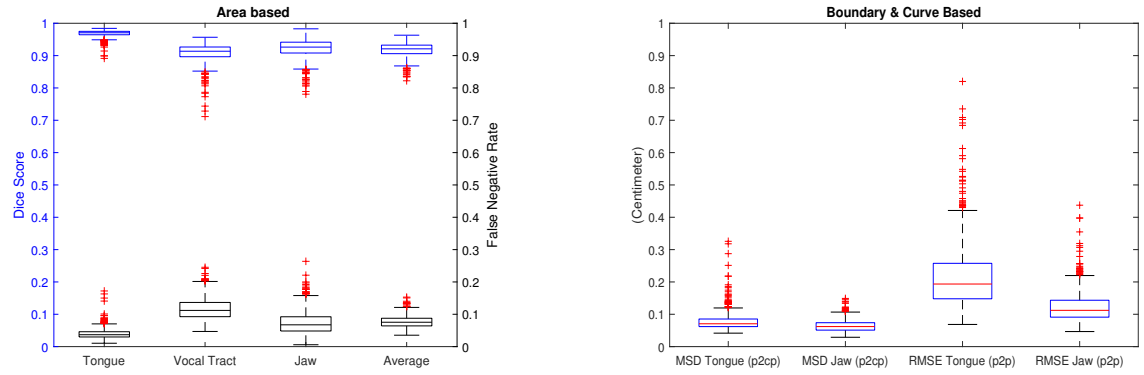
$$\mathscr{L}_{cGAN}(G,D) = \mathbb{E}\left[logD(x,y)\right] + \mathbb{E}\left[log(1 - D(x,G(x,z)))\right] \tag{1}$$

3 Material and Method

3.1 Method

In this work, we have selected an instance of cGAN named *pix2pix* [11] to use as basis of our method. This instance has been selected for its proven results regarding segmentation (*e.g.* [8]), for the public availability of its code [12], and for its adaptability for future works due to its cGAN architecture, specially for speech analysis and generation. The generator of pix2pix, which learns to transform input images into desired target images, is made of an auto-encoding CNN architecture [11]. The discriminator, which learns to determine whether an input image is real or a fake created by the generator, is made of a CNN called *PatchGAN*, where a patch is subset part of an image, *i.e.* discriminator attempts to determine whether each patch in an image is real or fake. In our study, the input images of the generator are midsaggital MRI images of the vocal tract and the output images the corresponding images where the jaw, the tongue and the vocal tract are coded with respectively blue, red, green pixels and the remaining pixels being coded in black. An example is visible on Figure 3.

In pix2pix, the objective function is $\hat{G} = arg\ min_G\ max_D\ [\mathscr{L}_{cGAN} + \lambda \mathscr{L}_{L1}(G)]$ where $\mathscr{L}_{L1} = \mathbb{E}\left[\| y - G(x,z) \|_1\right]$ is the L1 distance between the target and output images decreasing the blurring effect. The default loss function provided with the pix2pix code is very general and designed to evaluate the cost for image-to-image translation. It has been modified in the present study to deal with the more specific problem of segmentation by adding the loss from the Soft Jaccard score between target and output images. The soft Jaccard score is similar to the hard Jaccard score, which compares the similarity of two batch of thresholded data, but unlike it deals with non-thresholded intensities, which makes it derivative, a desired attribute for the loss function of a neural network. The objective function used in our study is shown in (2) where $\mathscr{L}_J(G)$ is the soft Jaccard score and defined in (3). Y_is and x_is are the pixel intensities and i is



(a) **(b)**

Figure 2 – Box plots of overall segmentation results for the Dice score and FNR (a) and the RMSE (b).

the pixel counter.

$$\hat{G} = arg\ min_G\ max_D\ [\mathscr{L}_{cGAN} + \lambda \mathscr{L}_{L1}(G) + \beta \mathscr{L}_J(G)] \tag{2}$$

$$\mathscr{L}_J(G) = 1 - soft\ \ jaccard\ score\ (y, G(x,z)) = 1 - \frac{2\ \Sigma_i(y_i\ x_i) + \varepsilon}{\Sigma_i(y_i^2\ x_i^2) + \varepsilon} \tag{3}$$

3.2 Data and Evaluation

To test the method, static midsagittal MRI were considered for 10 French speakers (5 males, 5 females) sustaining 62 articulations and the articulators surrounding the vocal tract have been manually segmented [13, 14]. The vocal tract contour have been obtained as the resultant contour of the articulator contours. Note that one speaker has been discarded in comparison to [13] and [14] due to the significantly lower quality of the images. We assume that the current method will not be used in the future on images of such quality. The corpus, designed to be representative of the French phonemic repertoire, consisted of the 10 French oral vowels [i e ɛ a y ø œ u o ɔ], 2 of the 4 nasal vowels [ã ɔ̃], and each of the 10 consonants [p t k f s ʃ m n ʁ l] in the 5 symmetric vowel contexts [i e ɛ a u].

The method is evaluated following a leave-one-out cross validation (LOOCV) procedure on the articulations of each speaker individually as preliminary analyses. In other words, the network is trained for each speaker on 61 articulations and evaluated on the remaining 62^{nd} articulation. The training was done via a system equipped by NVIDIA gpu Tesla P100 with 16 GB memory. The network has 57,190,084 parameters to be trained and it takes almost 30 minutes for 600 epochs with a batch size of 60 sample. λ and β in equation 2 are set to 10 experimentally. The size of input/output images and kernel(or filter) are 256x256x3 and 4x4. The auto-encoding network has the following architecture: Encoder of generator: 256x256x3 -> 128x128x64 -> 64x64x128 -> 32x32x256 -> 16x16x512 -> 8x8x512 -> 4x4x512 -> 2x2x512 -> 1x1x512; Decoder of generator: 1x1x512 -> 2x2x512 -> 4x4x512 -> 8x8x512 -> 16x16x512 -> 32x32x256 -> 64x64x128 -> 128x128x64 -> 256x256x3; Discriminator: 256x256x3 => 256x256x6 -> 128x128x64 -> 64x64x128 -> 32x32x256 -> 31x31x512 -> 30x30x1 (the => symbol codes the concatenation of the output and target images).

The segmented regions of the jaw, of the tongue and of the vocal tract generated by the model, *i.e.* taken from the output images, are compared with the same regions generated from the manual segmentation by means of the Dice score and the False Negative Rate (FNR), following standard evaluation in image processing. These metrics evaluate the accuracy in terms of surface while our interest lies in the contours. Lastly, they consider regions that do not have impact on the shape of the vocal tract, like the root of the tongue. To overcome these limitations,

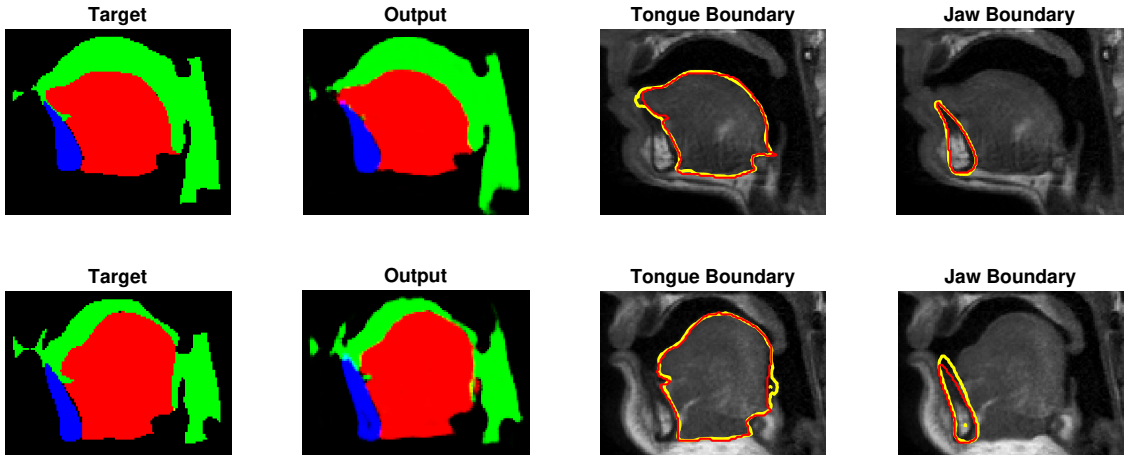


Figure 3 – Best and worst segmentation results based on the Dice score.
Top) Best result: speaker: 3, articulation: [m^{e}], average Dice: 0.96, average FNR: 0.04.
Bottom) Worst result: speaker: 8, articulation: [m^{u}], average Dice: 0.83, average FNR: 0.12.

for each of the three considered regions, the contours corresponding to the boundary with the vocal tract have been selected. This approach has been applied to the images generated by the model and by the manual segmentations in order to generate homogeneous sets of data. The two sets of contours have been compared by means of Root Mean Squared Error (RMSE) and Mean sum of distances (MSD) [2]. The distances in a point-to-point (p2p) fashion are considered for RMSE whereas in a point-to-closest-point (p2cp) fashion for MSD [2]. Note that p2p distances are inherently larger than p2cp distances.

4 Results and Discussions

4.1 Results

Figure 2a shows the statistical box plots for the overall Dice scores and FNR of the tongue, the jaw, the vocal tract and their average. The average and standard deviation of the Dice score and FNR are 0.92 ± 0.02 and 0.08 ± 0.02 respectively, which shows promising performance. The results corresponding to the best and worst Dice scores are shown in Figure 3 for illustration.

Regarding MSD, the overall mean and standard deviation for the tongue and the jaw are 0.08 ± 0.03 (cm) and 0.06 ± 0.02 (cm) respectively, whereas 0.22 ± 0.10 (cm) and 0.12 ± 0.05 (cm) for RMSE. The corresponding statistical box plots are visible in figure 2b. These results confirm the promising performance noticed for the Dice score and the FNR. The achieved MSD result for tongue is comparable with proposed method in [2] which is reported as 0.73 (mm) (The result of jaw is not addressed). In addition, we believe that exploiting the proposed method in this paper is more easily and need very limited manual configurations. The results corresponding to the best and worst RMSE are shown in Figure 4 for illustration.

In order to visualize the results per articulation the box plots of the Dice scores and MSDs for the 62 different articulations are shown in Figure 5. All the Dice scores and MSDs are in the range $[0.9\ 0.95]$ and $[0.05\ 0.1](cm)$ respectively. Although the results appear rather similar across articulations, the detailed results corresponding the 3 best and worst results are listed in Table 1. It appears that the articulations [f^{u}], [l^{u}] and [n^{a}] are more challenging to segment.

The same approach is repeated for the speakers and the box plots are shown in Figure 6. Only little variation is observed between speakers, all of them showing equal performance within the range $[0.9\ 9.95]$ and $[0.05\ 0.1](cm)$ for Dice and MSD respectively. One speaker only shows results out of this range, (MSD: $0.12(cm)$); it corresponds also to images with

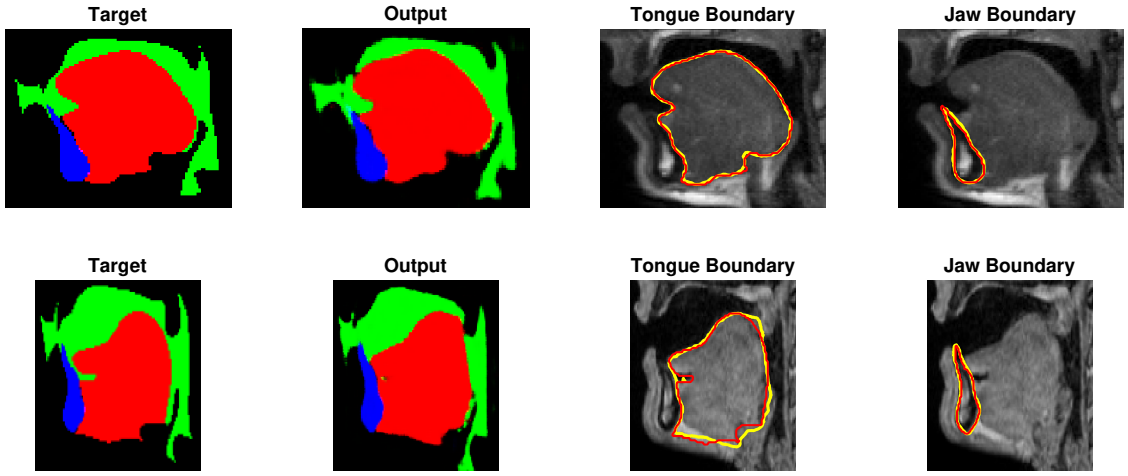


Figure 4 – Best and worst segmentation results based on the RMSE of the tongue.
Top) Best result: speaker: 2, articulation: [ʃ$^{\varepsilon}$], RMSE: 0.07 (cm), MSD: 0.04 (cm).
Bottom) Worst result: speaker: 4, articulation: [ʁ$^{\varepsilon}$], RMSE: 0.82 (cm), MSD: 0.14 (cm).

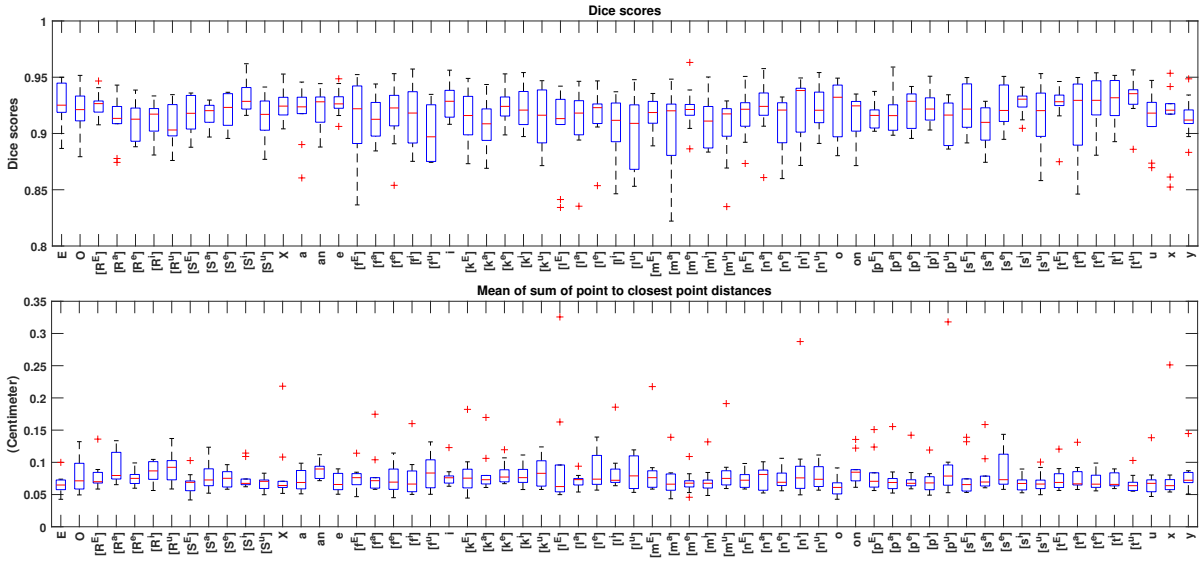


Figure 5 – Overall Dice scores (top) and MSD in cm (bottom) of the tongue the 62 articulations.
P.S. In this Figure [E O R S an on x X] stand for [ɛ ɔ ʁ ʃ ã ɔ̃ ø œ] respectively.

Table 1 – Detailed results for the three best and worst results measured on the tongue for the articulations.

	Measurement	**Top 1**	**Top 2**	**Top 3**
Best average of	distances by MSD *(cm)*	[o]: 0.06 ± 0.01	[ɛ]: 0.07 ± 0.02	[s^{i}]: 0.07 ± 0.01
Best average of	distances by RMSE *(cm)*	[t$^{\varepsilon}$]: 0.16 ± 0.04	[ʃ^{u}]: 0.16 ± 0.07	[f$^{\varepsilon}$]: 0.16 + 0.04
Best average of	Dice scores	[ʃ^{i}]: 0.93 ± 0.01	[tu]: 0.93 ± 0.02	[ti]: 0.93 ± 0.02
Best average of	False Negative Rates	[ʃ^{i}]: 0.07 ± 0.01	[te]: 0.07 ± 0.02	[ti]: 0.07 ± 0.02
Worst average of	distances by MSD *(cm)*	[p^{u}]: 0.10 ± 0.08	[l$^{\varepsilon}$]: 0.10 ± 0.09	[n^{i}]: 0.10 ± 0.07
Worst average of	distances by RMSE *(cm)*	[ɔ̃]: 0.31 ± 0.14	[ã]: 0.29 ± 0.19	[n^{a}]: 0.29 ± 0.14
Worst average of	Dice scores	[f^{u}]: 0.90 ± 0.02	[l^{u}]: 0.90 ± 0.03	[m^{u}]: 0.90 ± 0.03
Worst average of	False Negative rates	[f^{u}]: 0.09 ± 0.02	[l$^{\varepsilon}$]: 0.09 ± 0.03	[l^{u}]: 0.09 ± 0.03

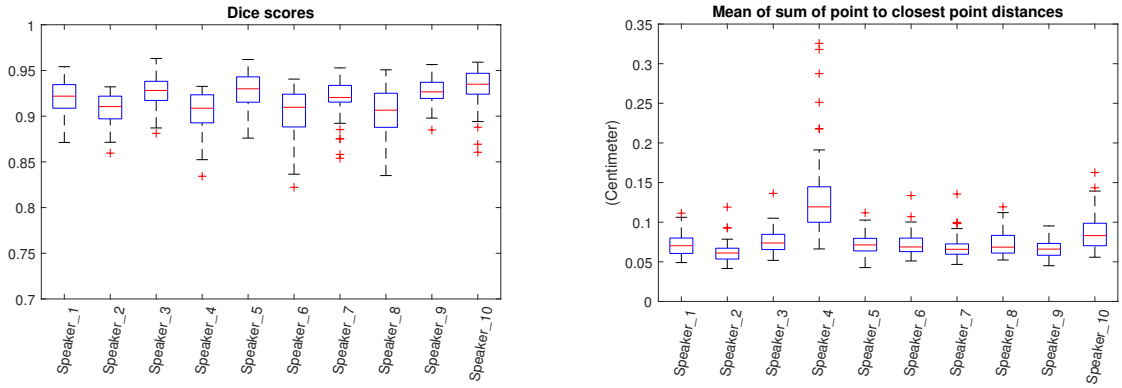


Figure 6 – Overall Dice scores (left) and MSD (right) of the tongue for the 10 speakers.

lower quality and higher noise (see Figure 4b).

4.2 Discussion and Future Works

The segmentations obtained with the method proposed in this study show in general good accuracy and the preliminary qualitative and quantitative assessments appear positive. The LOOCV procedure ensures that the network is able to segment articulations that have not been used for training and is therefore generalisable to any articulation.

Further works are needed to optimize the networks and improve the segmentation accuracy results. In particular, post-processing methods are necessary to transform the segmentation results into exploitable contours for the purpose of articulatory modelling. This involve the development of methods to extract automatically anatomical landmarks. Further, the method needs to be more robust in order to deal with very different speakers, articulations or lower image quality (such as on Figure 4b)

However, the preliminary analysis carried out in this study shows very promising results and further development plans include analyses on other articulators, a cross-speaker segmentation method and tests on real-time MRI data. In a longer term, the method proposed in the present study opens the way for promising perspectives in articulatory speech modelling and therapy.

5 Conclusion

Segmenting the vocal tract and its surrounding articulators is a crucial step for analyzing the speech production mechanisms and their disorders. In this work, a segmentation method based on modifying the pix2pix method, which is a conditional Generative Adversarial Network, has been proposed. The method shows very promising results for segmenting the jaw, the tongue and the vocal tract which is comparable with the state of the art. Also, the proposed method would be trained in an end-to-end fashion and requires very limited of manual procedure. These preliminary results open the way to a vast amount of future work in speech production modelling based on deep learning.

Acknowledgement

The authors are very grateful to P. Badin for providing the data, L. Lamalle recording them, and J.-A. Valdés Vargas and G. Ananthakrishnan for performing the majority of the initial tracings. This research project is supported by the START-Program of the Faculty of Medicine, RWTH Aachen. Also, this work has been partially funded by the French ANR (grant ANR-08-EMER-001-02 “ARTIS”).

References

[1] KIM, Y.-C. and Y.-C. KIM: *Fast upper airway magnetic resonance imaging for assessment of speech production and sleep apnea. Precision and Future Medicine*, 2(4), pp. 131–148, 2018.

[2] LABRUNIE, M., P. BADIN, D. VOIT, A. A. JOSEPH, J. FRAHM, L. LAMALLE, C. VILAIN, and L.-J. BOË: *Automatic segmentation of speech articulators from real-time midsagittal mri based on supervised learning*. *Speech Communication*, 99, pp. 27–46, 2018.

[3] SAMPAIO, R. D. A. and M. P. JACKOWSKI: *Vocal tract morphology using real-time magnetic resonance imaging*. In *Graphics, Patterns and Images (SIBGRAPI), 2017 30th SIBGRAPI Conference on*, pp. 359–366. IEEE, 2017.

[4] VALLIAPPAN, C., R. MANNEM, and P. K. GHOSH: *Air-tissue boundary segmentation in real-time magnetic resonance imaging video using semantic segmentation with fully convolutional networks*. In *Proc Interspeech*, vol. 2018, pp. 3132–6. 2018.

[5] SOMANDEPALLI, K., A. TOUTIOS, and S. S. NARAYANAN: *Semantic edge detection for tracking vocal tract air-tissue boundaries in real-time magnetic resonance images*. *Proc. Interspeech 2017*, pp. 631–635, 2017.

[6] YI, X., E. WALIA, and P. BABYN: *Generative adversarial network in medical imaging: A review*. *arXiv preprint arXiv:1809.07294*, 2018.

[7] GARCIA-GARCIA, A., S. ORTS-ESCOLANO, S. OPREA, V. VILLENA-MARTINEZ, and J. GARCIA-RODRIGUEZ: *A review on deep learning techniques applied to semantic segmentation*. *arXiv preprint arXiv:1704.06857*, 2017.

[8] ZHANG, H., Y. SUN, L. LIU, X. WANG, L. LI, and W. LIU: *Clothingout: a category-supervised gan model for clothing segmentation and retrieval*. *Neural Computing and Applications*, pp. 1–12, 2018.

[9] CHANG, W., G. YANG, J. YU, and Z. LIANG: *Real-time segmentation of various insulators using generative adversarial networks*. *IET Computer Vision*, 2018.

[10] HAN, X., J. LU, C. ZHAO, S. YOU, and H. LI: *Semi-supervised and weakly-supervised road detection based on generative adversarial networks*. *IEEE Signal Processing Letters*, 2018.

[11] ISOLA, P., J.-Y. ZHU, T. ZHOU, and A. A. EFROS: *Image-to-image translation with conditional adversarial networks*. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976. IEEE, 2017.

[12] *Pix2pix code via tensorflow*. `https://phillipi.github.io/pix2pix/`, ???? Accessed: 2019-01-20.

[13] VALDÉS VARGAS, J. A.: *Adaptation of orofacial clones to the morphology and control strategies of target speakers for speech articulation*. Ph.D. thesis, Université de Grenoble, 2013.

[14] SERRURIER, A., P. BADIN, L.-J. BOË, L. LAMALLE, and C. NEUSCHAEFER-RUBE: *Inter-speaker variability: speaker normalisation and quantitative estimation of articulatory invariants in speech production for french*. In *18th Annual Conference of the International Speech Communication Association (Interspeech 2017)*, pp. 2272–2276. 2017.