POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Vocal tract segmentation of dynamic speech MRI images based on deep learning for neurodegenerative disease application

Laurea Magistrale in Biomedical Engineering - Ingegneria Biomedica

Author: Angelica Bonà, Matteo Cavicchioli

Advisor: Professor Pietro Cerveri

Co-advisor: MSc Matteo Rossi, PhD Maria Luisa Mandelli

Academic year: 2020-2021

1. Introduction

1.1. Clinical context

Speech articulation is the most complex motor activity humans perform. Difficulties with articulation are called motor speech impairments and may be the first symptoms present in several neurodegenerative diseases such as non-fluent/agrammatic variant primary progressive aphasia (nfvPPA), manifesting as progressive apraxia of speech (AOS) and/or dysarthria [1]. Currently, diagnosis relies on the perceptual and subjective judgment of clinicians. Linking the vocal tract's shape alterations, extracted automatically, with clinical and acoustic speech evaluations has the potential of better defining the anatomical changes of specific articulation deficits and might provide an effective tool for diagnosis and monitoring of neurodegenerative diseases. MRI sequences of the vocal tract have several advantages over other existing instrumental approaches that either have limited spatial coverage of the vocal tract (ultrasound, electropalatography), are invasive (cine x-ray and optical coherence tomography), or alter articulatory kinematics (electromagnetic articulography). In particular dynamic speech MRI (dsMRI), an innovative MR technology, offers a unique opportunity for fast, direct, non-invasive, real-time visualization of the changes in the vocal tract during speech.

1.2. MRI analysis methods

Because of number of articulators involved in the speech production and the complexity of their anatomy, there are still no gold standard datasets available and this makes difficult to evaluate the performance of proposed automated analysis. The existing techniques used to analyze dsMRI images to investigate speech properties can be summarized in four classes [2]:

1. *Basic decomposition or matrix factorization techniques*: these methods obtain spatio-temporal basis functions of the articulators' movement associated to linguistic gestures;
2. *Region of interest (ROI)-based*: they are based on the manual demarcation of the regions of interest of which variation can provide useful information regarding linguistic or clinical questions;
3. *Grid-based*: they are based on a reference coordinate system that is superposed a sagittal view of vocal tract to facilitate the calculation of the vocal tract area functions

by the identification of points of intersection between soft tissue and gridlines;

4. *Contour-based*: they are based on the extraction of all the tissue boundaries belonging to structures recruited during speech production.

The last class of methods is the one that was used in this work to develop an automated image-segmentation tool, exploiting a deep learning approach (advanced UNet), to extract the contouring of the main articulators that are critically involved in speech production. The final aim of this tool is to extract quantitative metrics that can provide useful information to detect motor speech impairments and follow the progression of disease over time.

2. Materials and Methods

2.1. Data Collection

A multidisciplinary team from Department of Neurology and Department of Radiology of University California San Francisco (UCSF) enrolled 10 young and 20 older healthy controls as well as 15 nfvPPA patients from active projects at the UCSF Memory and Aging Center (MAC) and the Language Neurobiology Laboratory (ALBA). Speech stimuli were provided to the subjects to guarantee a wide range of permissible articulatory movements in Standard American English. After training, all participants underwent MRI on a 3T Siemens Prisma scanner where they repeated the speech stimuli during the MRI acquisition. A series of mid-sagittal slices for dynamic speech MRI were acquired during the speech stimuli and grouped into dynamic speech videos. The study was approved by the UCSF Committee on Human Research and all subjects provided written informed consent.

2.2. Dataset preprocessing

Because of the restrictions due to the COVID-19 pandemic, only 4 young control subjects and 1 nfvPPA patient were able to complete the study protocol. In Table 1 the speech videos provided by UCSF and used to extract the images and build the dataset are listed.

Table 1: Each video includes the Subject ID, the stimulus provided and the belonging to patient or control group.

Subject ID	Stimulus	Control/Patient
1	SEGREGATION	Control
1	MICROSCOPIC	Control
1	TOPCOP	Control
2	MICROSCOPIC	Control
2	SEGREGATION	Control
2	TOPCOP	Control
3	PATAKA	Control
3	MICROSCOPIC	Control
3	WELCOME	Control
4	PA	Control
4	KA	Control
4	COUNT	Control
5	SEGREGATION	Patient
5	MICROSCOPIC	Patient
5	TOPCOP	Patient

A graphical user interface (GUI) was developed in Python 3.7.9 to extract the desired number of frames (equally spaced) from videos, create and organize the dataset (made by 970 images). A manual annotation of the anatomical contouring of the main articulators was provided under the supervision of an expert radiologist using 3D Slicer software to obtain the ground truth segmentations. The manual segmentation took approximately 15 minutes for each image. Considering the onset of fatigue of the operators after about 20/25 images, it took approximately 250 hours to complete the segmentation process. Since pixels of dsMRI images assumed values between 0 and 256, they needed to be normalized between 0 and 1 to obtain better performances. A min-max normalization was applied using 0 as minimum value and the 90th percentile (about 130) as maximum value. Non zero pixels belong to the so called *foreground*, whereas zero pixels belong to the so called *background*. Figure 1 shows an example of a dsMRI frame and the corresponding manual segmentation where seven regions of interest were identified: Upper Lip (UL): green, Hard Palate (HP): yellow, Soft Palate (SP): soft brown, Tongue and Epiglottis (TO): light blue, Lower Lip and Jaw: red, Head (HE): orange and Background (BK): black or dark grey.

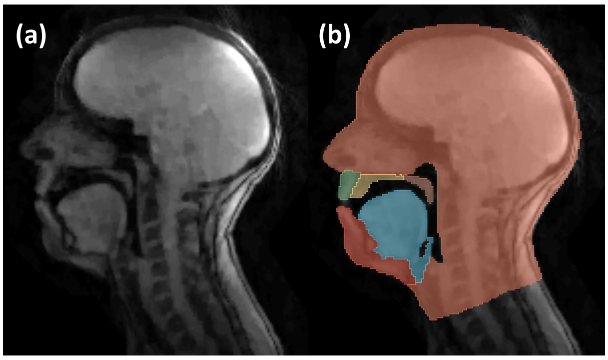

Figure 1: a) Example of dsMRI frame of dimensions 256 × 256, b) Corresponding ground truth.

2.3. UNet

A groundbreaking architecture called UNet appears to be promising with respect to segmentation problems. It can efficiently process large images by combining information from both local and global features. UNet include two operational paths connected each other through a bottleneck. The first is the contraction path where the network learns by coding (convolution) and compressing (maxpooling) the initial image into a series of feature maps. The second is the expansion path where the information obtained is brought back to the initial resolution (deconvolution) and decoded in the segmentation mask. To enhance network performances, skip connections are used between the same levels of the two paths. Due to them the expansion path can better recover spatial information by merging features skipped from the various resolution levels on the contracting path. Based on these considerations, different UNet architectures were developed and trained with different loss functions. Their output is a 7-layers Softmax that provide the seven classes probability maps and their accuracy was evaluated by some properly chosen metrics. UNet architectures developed were five, but only the most effective ones will be explained below.

2.3.1 QT-UNet

QT-UNet architecture has an encoding unit structured as a dense block, where each convolutional layer receives all the previous outputs as inputs. This means that bottleneck receives portions of all the previous layers as input, enhancing information flow. Furthermore, this architecture replaces deconvolution as mean of up-sampling with an unpooling layer based on nearest-neighbor interpolation.

2.3.2 IM-UNet

IM-UNet uses residual blocks as encoding unit formed by two separate convolutional branches, with a branch having kernel size 1 × 1, joined by a pixelwise sum. This allows the network to propagate multi-scale information to the bottleneck, which is instead made up of four dilated convolutions that further increase the receptive field, without reducing resolution. Each encoding unit is equipped with a dropout layer to reduce overfitting. A new version of this architecture was also tested, obtained with the addition of an Attention Block. This was introduced in correspondence of the skip connections to diminish the number of redundant features that are brought from the down-sampling path to the up-sampling path [3].

2.4. Loss Functions

Loss function quantifies the discrepancy between ground truth and prediction, updating network parameters (weights and bias) every training epoch, through back-propagation. The loss functions used to train the networks are depicted in Figure 2. They are subdivided in three classes according to their operating method.

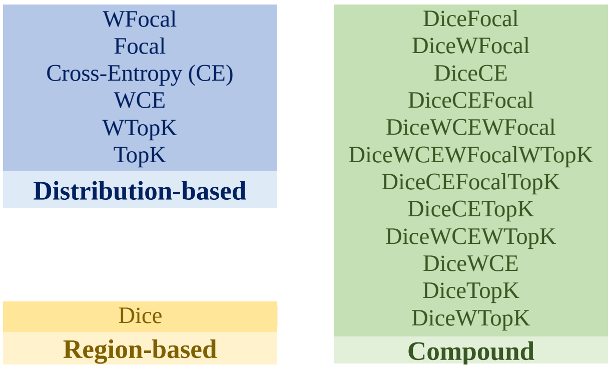

Figure 2: Graphical representation of the losses used, subdivided in their classes.

1. The Distribution-based class includes all the losses that aim to minimize the difference between two distributions. They were tested also in their weighted form for class imbalance problem, with weights given by:

$$w_c = \frac{npix_c}{Npix}, \tag{1}$$

where $npix_c$ represents the number of pixels belonging to the foreground of class c, while $Npix$ is the sum of the total pixels belonging to the different classes foregrounds;

2. The Region-based class includes all the losses that measure the difference between ground truth and prediction, trying to maximize the overlap between the two;
3. Compound losses are conceived as the sum of multiple losses, with the aim of improving performance by combining the different strengths. In literature there are compound losses composed of a maximum of two elements, while in this study this concept was extended to three and four elements.

2.5. Metrics

Networks segmentation accuracy was assessed using three complementary metrics. They were the most uncorrelated ones among those that focus on the properties that are most inherent to the considered task (small segment, complex boundary, low densities, importance of contour). First, Dice metric (DICE) was used to quantify the overlap between the ground truth segmentations and the predicted segmentations. Second, the Hausdorff Distance (HD) was used to quantify the precision of the prediction's spatial position and boundaries compared to the ground truth. Third, the Global Consistency Error (GCE) was able to focus particularly on the overlap of small portions of the images and take into account also the amount of true negatives. To evaluate the overall goodness of networks an overall metric (OM) was introduced:

$$OM = (1 - DICE) + GCE + HD \qquad (2)$$

2.6. Networks Training and Evaluation

A portion of the dataset, composed by 820 images of control subjects, was split into training (80%), validation (10%) and test (10%) sets. Networks taken into account were given by the combination of loss functions and architectures mentioned above. They were all trained with a batch size of 8, 70 epochs and Adam optimizer with learning rate of 0.001. The best networks in term of the overall metric were saved and applied on the test set. The project was developed in *Google Colab* environment using *TensorFlow v2.8.0*. The initial trainings were conducted exploiting *Google Colab* GPU NVIDIA Tesla T4 with RAM of 25 Gigabyte and each model took about 1 hour and 20 minutes to be trained. The successive trainings exploited a cluster of NVIDIA Tesla A100 with RAM of 40 Gigabyte each and each model took about 40 minutes to be trained. Particularly, the mean value of inference time for a single image was about 0.092s.

2.6.1 Statistical analysis

A statistical analysis was made to rank the 95 networks and asses their statistical differences. First the overall metric distribution on the test set was extracted from each network. These distributions were used to perform Kruskal Wallis test which produced a p-value equal to 0, meaning that at least two networks were significantly different each other. Then Tukey Kramer test was applied to see which networks were or not significantly different from one another. All the analysis was performed in MATLAB R2021b environment.

2.6.2 Cross Validation

Cross validation was applied on the best networks selected after the statistical analysis. The subject-one-out cross validation was implemented using the 4 control subjects to guarantee that each subject could appear in both the training and the test set, enhancing the variability of data.

2.7. Post processing

Since the seven classes were predicted separately by networks and their predictions were fuzzy $[0, 1]$, they were converted into crisp segmentations $\{0, 1\}$ and reassembled to obtain the overall predicted segmentation. The conversion carried out by introducing a threshold of 0.5 on the pixels produced holes in the most uncertain areas of Softmax probability map. So the entire image was scanned and each pixel was assigned to the class to which it has the highest probability of belonging (Argmax). This way all holes were filled and all pixels were assigned to a class.

2.8. Vocal Tract Segmentation Tool

In order to make this work accessible to clinicians and allow them to benefit from the segmentations, a user friendly application was developed using Python 3.7.9 and KV language. Vocal tract segmentation tool (VTS-tool) uses the best networks to perform rapid segmentation of the MRI recording and it allows to interactively visualize the images with their predicted segmentations superposed. It also gives the possibility to compute some clinical metrics both interactively (distances) and automatically (dynamic computation of articulators areas). This way clinicians can evaluate the trend of the change of the articulators areas over time.

3. Results

The three best networks, with their Dice medians, obtained from the statistical analysis were:

- IM-UNet with Attention Block (IMUNetAtt) trained with Dice-CrossEntropy-Focal-TopK loss (0.9375);
- IM-UNet trained with Dice-CrossEntropy-Focal loss (0.9285);
- QT-UNet trained with Dice-CrossEntropy-Focal loss (0.9256)

After the post processing they also obtained a median Hausdorff Distance of 0.32, a median Global Consistency Error of 0.0011 and a median overall metric of 0.38 on the control subjects. Since their best possible value is 0 and the worst possible value is 1, these results are considered satisfactory. Figure 3 provides a graphical example of the best networks capacity to correctly predict a control subject segmentation.

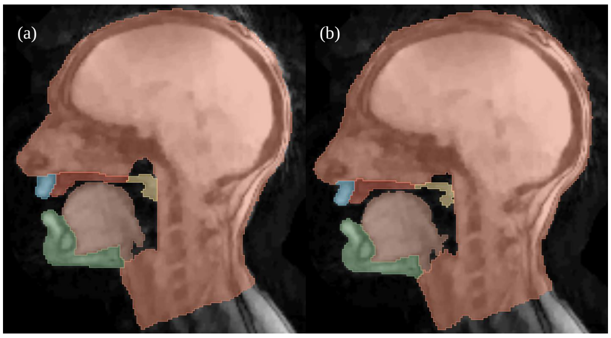

Figure 3: Example of a) ground truth segmentation of a control subject and b) the correspondent predicted and post processed segmentation.

In Figure 4 was depicted an example of the manual delineation of the articulators as well as the predicted areas with the corresponding under-segmented and over-segmented areas. Specifically, under-segmented areas are those ones that should have been included in a region but were not; over-segmented areas are those ones that should have been excluded from a region but were not.

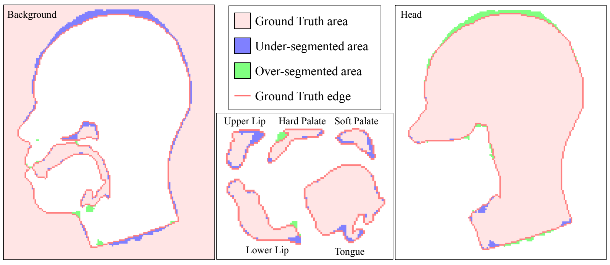

Figure 4: Example of a map showing, for a control subject, the ground truth area of regions with the superposition of areas that were under-segmented and over-segmented.

Cross validation produced the following results: a median Dice of 0.91, a median Hausdorff Distance of 0.28, a median Global Consistency Error of 0.001 and a median overall metric of 0.37 on the control subjects. These results prove that these networks don't suffer from overfitting problem and they are able to correctly segment images when trained and tested with different sets.

Taking into account that patient articulators are quite different from the ones of control subjects, the same networks applied on his/her 150 images (out of 970) produced worse results. Specifically, they produced a median Dice of 0.82, a median Hausdorff Distance of 0.38, a median Global Consistency Error of 0.0026 and a median overall metric of 0.55. Figure 5 shows the change of the articulators areas in the patient over the repetition of a task obtained from the VTS-tool. This is one of the possible clinical metric that can be used to discriminate between a physiological production of speech and the presence of some motor speech impairments. In particular, as mentioned in 1.1, apraxia of speech (AOS) is characterized by inconsistent speech patterns, while dysarthria is connoted by consistent patterns. This way, by looking at areas trend, is possible, for example, to discriminate between these two pathological conditions.

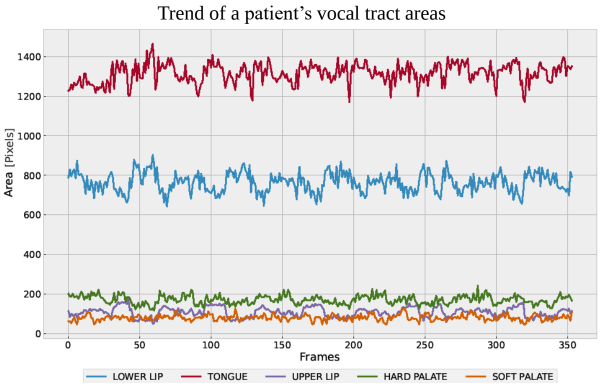

Figure 5: Trend of patient articulators areas during the repetition of the task *Segregation*.

4. Discussion

Networks that perform the best in the segmentation task were those ones built with compound losses with three and four elements. This means that combining *Dice loss* and *Cross-Entropy loss* succeeds in taking into account the dissimilarities between the two distributions and also the overlap degree. Then, adding *Focal loss* allows to penalize the well-classified samples to focus on the worst ones. This loss amplifies precision to the detriment of recall, meaning that it privileges a sub-segmentation rather than oversegmentation. This tendency is confirmed by the map shown in Figure 4, where a strong predominance of under-segmented areas can be seen. Eventually, adding *TopK loss* allows to focus on the most difficult pixels.

The reason why IMUNet, IMUNetAtt and QTUNet prevailed over the other two architectures may be linked to the propagation of the initial information through all the layers before the bottleneck. The IMUNet and IMUNetAtt, in their encoding path, propagate the initial information layer by layer, adding the linear projection of the input with a deeper convolution. QTUNet, instead, propagates the input to the bottleneck through the connections of the dense block, increasing the information flow received. This concept, as well as improving the quality of the network outputs, also improves its performances, as back-propagation is facilitated.

Dice values are all equal or greater than 0.92, meaning that the amount of overlap between the predicted classes and their ground truth is satisfactory. Hausdorff Distance values are quite close to 0, it means that the prediction’s spatial position and boundaries are close to the ones of ground truth. Global Consistency Error values are very close to zero, meaning that also the overlap of the smallest portions of the regions is guaranteed and the amount of the true negatives is high as well as the amount of the true positives. Eventually, the Overall metric values are quite close to 0 and it means that the overall performance of the networks can be considered satisfactory.

5. Conclusions

The automatic segmentation of vocal tract in its main articulators was successfully performed by the best networks obtained. They gained satisfactory metrics results on control subjects and good results on patient. These networks achieve quite good generalizability and don’t suffer from overfitting problem. The VTS-tool developed allows clinicians to save time, because it is not necessary to perform the manual segmentation of each dsMRI image, which is a very time-consuming activity. It also allows to obtain quantifiable and objective clinical information that can help clinicians making an early diagnosis and a better monitoring of speech diseases.

6. Acknowledgements

This work was carried out at Department of Electronics, Information and Bioengineering of Politecnico di Milano, in collaboration with Department of Neurology and Department of Radiology of University California San Francisco.

References

[1] J. M. Ogar, N. F. Dronkers, *et al.*, “Progressive Nonfluent Aphasia and Its Characteristic Motor Speech Deficits,” *Alzheimer Disease & Associated Disorders*, vol. 21, no. 4, S23–S30, Oct. 2007.

[2] V. Ramanarayanan, S. Tilsen, *et al.*, “Analysis of speech production real-time MRI,” *Computer Speech & Language*, vol. 52, pp. 1–22, Nov. 2018.

[3] O. Oktay, J. Schlemper, *et al.*, *Attention U-Net: Learning Where to Look for the Pancreas*, Apr. 2018.