A Vocal Tract Segmentation and Analysis over a European Portuguese MRI Database[1]

Inês Carbone, Paula Martins*, António Teixeira, Augusto Silva

*Escola Superior de Saúde da Universidade de Aveiro

***Abstract* – Knowledge of the speech production mechanism is essential for the development of speech production models and theories. Magnetic Resonance Imaging delivers high quality imaging of soft tissue. To our knowledge, there are no complete systematic Magnetic Resonance Imaging studies of European Portuguese production. With this work, we intend to create a database of Magnetic Resonance Imaging images of European Portuguese and to implement segmentation techniques well known on the image processing field, but rarely used on the context of vocal tract segmentation. We present 2D vocal tract contours as well as their evaluation using a similarity metric, the Pratt Index. This metric allows not only to validate the segmentation technique, but also to infer on the differences among the sounds in study.**

I. INTRODUCTION

Our knowledge about speech production and perception is still incomplete. More information is definitely needed. Recently, better ways of measuring vocal tract configurations have become an increased research interest. One of the most promising imaging techniques is Magnetic Resonance Imaging (MRI), which has been used for the study of several languages: American English (e.g. [1]), French (e.g. [2], [3]), Swedish (e.g. [4]), Japanese (e.g. [5]), German (e.g. [6]), Tamil (e.g. [1]), etc. For European Portuguese, there is only one restrict study, [7], but 3D and Real Time acquisitions were not reported.

The viability of a useful MRI database is determined by the existence of a reliable, fast and with low human interaction segmentation method. This is particularly relevant when using Real Time MRI, where the number of images to process increases exponentially. In this work we present Region Growing based contours of the vocal tract and quality tests made between the contours. This is innovative in the sense that the only comparison article between contours using several segmentation techniques that we found was made by Soquet and coworkers in 1998, [2]. In the cited study, authors compare manual segmentation (used, for e.g. in [3]), semi-automatic thresholding techniques (used, for e.g. in [7]), and a semi-automatic snake technique (used, for e.g. in [8]). They concluded that these techniques lead to no significant differences between contours. However, all of these techniques need manual initialization, manual control of the development of the contours and/or manual correction of the final contour.

The importance of the development of a quick, easy to use and accurate segmentation method for the segmentation of MRI images in this field is related to the development of the MRI acquisition method. Not only the quality of images and volumes generated tend to improve, but also the time spent to obtain the images tends to decrease. These improvements open the research problems to acquisitions in Real Time [1], very important in the speech study because it has been noted [4] that acquisitions of artificially sustained sounds may lead to misinterpretations of the real speech production, a dynamic process. Concerning these problems, we intend to improve the segmentation methods used, having developed and proved the efficiency of the Region Growing method in these images. The main advantage of this method is that the only human interaction is represented by the choice of a point inside the region of interest (the “seed” point) and a threshold.

The study of the robustness of the segmentation method is also very important. We need to make sure that the contours generated are truthful enough to represent the vocal tract configuration of the sound being produced. The contours cannot contain errors that may lead to a misinterpretation and/or confusion of the sound with another one. This was evaluated with a metric called the Pratt Index.

The article is structured as follows: Section II gives some information on the image acquisition process, the corpus and the informant; Section III presents the Region Growing segmentation and the metric used to compare the contours; Section IV shows the results. We finish, in Section V, with some observations on the results and a few future work topics.

II. METHOD

For this study, we used a small part of a recently acquired European Portuguese MRI database. The sounds studied are described in Table I. These sounds were sustained and the informant was instructed to produce them in the context of a reference word. The informant was a 25 years old male with vocal and singing training having as native language European Portuguese.

The images were acquired using a 1.5 Tesla (Magneton Simphony, Maestro Class, Siemens, Erlanger, Germany) scanner equipped with Quantum gradients (maximum amplitude = 30 mT/m; rise time = 240 μs; slew rate = 125 T/m/s; FOV = 50 cm). Neck and brain phased array coils were used.

We used one of the four parts of the corpus: acquisition 2D

[1] This work was supported by FCT

Sound	Reference Word	Translation	Phonetic Transcription
a	Pato	Duck	[patu]
i	Pipo	Barrel	[pipu]
u	Buda	Buddha	[bud6]
6	Cada	Each	[k6d6]
@	Devi	Form of the verb to owe (money)	[d@vi]
E	Leva	Take	[lEv6]
e	Peca	hard to translate, meaning Atrophic	[pek6]
O	Pote	Pot	[pOt@]
o	Tôpo	Top	[topu]
6~	Canto	Corner	[k6~tu]
i~	Minto	I lie	[mi~tu]
e~	Pente	Komb	[pe~t@]
o~	Ponte	Bridge	[po~t@]
u~	Punto	Trade mark (Fiat)	[pu~tu]
m	Cama	Bed	[k6m6]
l	Laço	Bow	[lasu]
s	Sala	Room	[sal6]

TABLE I

List of the sounds studied. We use SAMPA (Speech Assessment Methods Phonetic Alphabet).

in the sagital plane. This corpus was acquired in the sagital plane with a TSE T1 weighted sequence (Slice thickness = 5 mm, FOV = 200 mm, Matrix = 192x256, ETL = 15). Each of these images took 5.6 seconds to be acquired.

III. Vocal tract segmentation

A. Segmentation Method

The segmentations were made with the Region Growing method, [9]. We start by manually placing a seed inside the vocal tract and it expands until it reaches the vocal tract wall. This expansion is based on grey level comparison between the mean grey level value of all the pixels already marked as inside the vocal tract and the neighborhood pixels of the contour of the region already delimited. The stop criterion is based on a maximum difference threshold between the pixel being tested and the mean value of all the pixels assumed to belong to the region of interest. This method was implemented in Matlab and its pseudo-code is represented below.

```
ROI= Seed
Mean=Image(ROI;
While ROI is not empty do:
 TestPixel=ROI(1)
 ROI=ROI- TestPixel
 For each one of the 8 Neighbors do:
  Error=Mean-Image(Neighbor$_i$)
  If Error > Threshold
   Contour= [Contour; Neighbor$_i$]
  Else
   ROI = [ROI; Neighbor$_i$]
   Mean=Image(ROI)
  End
 End
End
```

B. Evaluation Criterion

The analyses of the influence of the seed in the final contour as well as the comparisons between contours were made with the Pratt Index [10]: $PRATT = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\alpha d_i^2}$, where N is the number of corresponding points between contours; d_i^2 is the distance between two corresponding points; α is related with the contour size and was chosen, based on one of the authors previous work on other types of images [10], to be $1/9$. This Index has its range in the interval [0, 1] where 1 means that the two contours are equal.

IV. 2D Results

We generated 100 contours (each set takes about 35 minutes with the current implementation) with a randomly seed inside the vocal tract, for each image. The reference contour was chosen to be the mean of these contours. Fig. 1 shows some examples. We opted not to have a contour drawn by an expert because of several reasons: lack of time, lack of specialists and, the most important one, it has been proved, see [10], [11], that the contours manually drawn suffer of a large inter and intra variability.

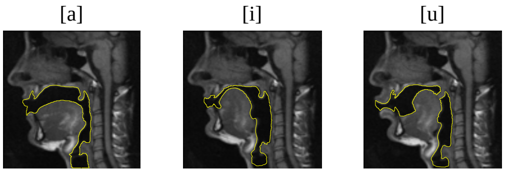

Fig. 1 - MRI images of the point vowels [a, i, u] and the mean contours (in yellow).

In some cases, we had to eliminate some of the 100 contours generated. These contours were significantly different from the rest of the contours generated and were considered as outliers. We like to stress that the quantity of contours considered as outliers did not exceed in any case 18 contours (18%).

Each contour was compared with the mean contour to assess reliability of results. Fig. 2 shows that the Region Growing Segmentation method is robust to changes in the seed (low intra-variability). The Pratt Indices are close to 1, having as a minimum the value 0.84.

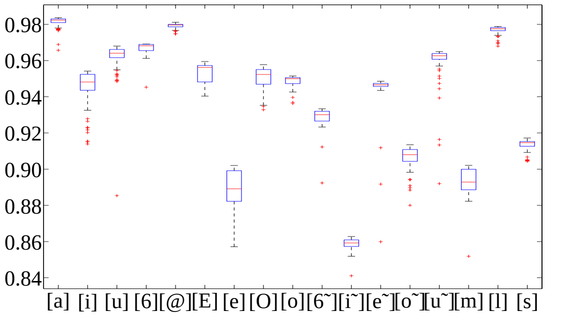

Fig. 2 - Boxplots of the Pratt Index differences using different seeds. Results for several different European Portuguese sounds (or phones) are presented.

The three worst cases were further studied looking at the dispersion of the contour points to detect where the variation is higher. The results are represented as ellipses, where the major axis is the variation on the contour in the horizontal and the minor axis is the variation on the contour in the vertical, Fig. 3.

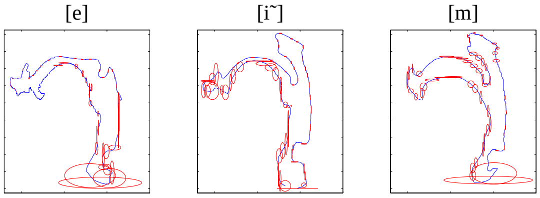

Fig. 3 - Study of the problematic zones for the sounds [e], [i˜], and [m]. Vocal tract contour in blue and dispersion ellipses in red.

One of the major difficulties that we found in this study is the influence of the teeth on the contour. As the teeth do not appear in the MRI images, they are confused with the vocal tract. This is visible in Fig. 3, where the zone corresponding to the teeth is one of the zones with the biggest variance. Another zone with high variance is the one in the bottom part of the images. This is not alarming since, in the articulatory studies, we are essentially interested in the vocal tract (between the lips and the glottis).

Also interesting, for speech production studies, is to compare the previous results with Pratt Index differences between the mean contours of the different sound tract shapes segmented. Fig. 4 presents these results.

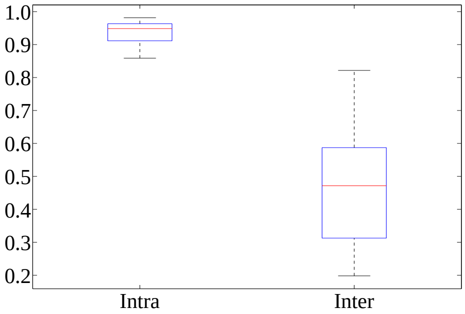

Fig. 4 - Comparison of the Pratt Index of the robustness of the contour with changes in the seed (Intra) with the Pratt Index of the mean contours (Inter).

The 95% confidence intervals are: $CI_p[0.92 \leq Intra \leq 0.96] = 95\%$ for the intra-variability, and $CI_p[0.44 \leq Inter \leq 0.49] = 95\%$ for the inter-variability, resulting in a statistically significant difference between variability due to segmentation method and real differences due to different sounds.

The comparison in Fig. 4 and the confidence intervals makes us conclude that the contours generated by the Region Growing method can be used to segment these kinds of images with no risk of confusion among the contours generated for different sounds.

As noted in [12], co-articulation is "a crucial issue in both speech synthesis and speech recognition to deal with ... when performing human speech behaviors". Based on this observation, we tested if our process is able to distinguish between the fricatives in three different contexts: VCV, where V represents one of the vowels belonging to [a], [i], [u]. An example of three contours for the fricative [f] is presented in Fig. 5.

Differences in segmentation results (observe the Fig. 5) can be attributed to the influence of the co-articulation and not to errors in the contours.

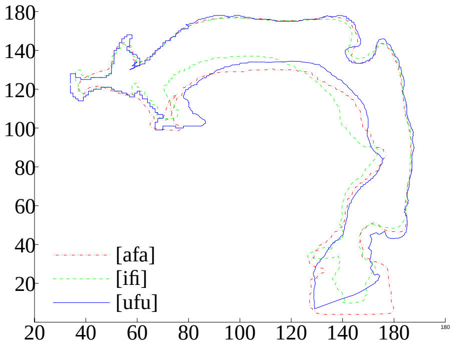

Fig. 5 - Context influence on Fricative [f].

V. CONCLUSION

We were able to implement a fast (less than 40 seconds per contour) segmentation method, with low human interaction. This method proved to have a high resistance to different seeds, and to be able to produce accurate contours for a wide set of Portuguese sounds.

Results, not completely presented in the paper due to space constraints, point to a possible use of the Pratt Index in the comparison of vocal tract configurations. Regarding our MRI database as a possible future standard, this Index could also be used as a measure of abnormal articulations.

Future Work: We are planning to implement other segmentation techniques and performing comparative evaluation. We hope to be able to improve even further the reliability of the contours, decrease segmentation time, and eliminate the user inputs.

We anticipate that segmentations will improve if some noise reduction as a pre-processing to the images is performed. This can be particularly relevant for lower quality Real Time images, which we want to process as automatically as possible.

It is mandatory to extend this line of work to 3D data in our database. The database acquired has some vocal tract volumes, and an initial sample segmentation is represented in Fig. 6. This ITK Snap [13] experiment took about one minute, suffering from a noticeable problem in the teeth zone.

At the moment we are studying the co-registration of the teeth (we made a 3D acquisition of the teeth for the same subject of this study) both on 2D images and 3D volumes.

VI. ACKNOWLEDGEMENTS

This work is part of project HERON (Ref.: POSI/PLP/57680/2004), funded by FCT (Portuguese Research Agency). We thank to the director of the Radiology Department of Coimbra University Hospitals and technical staff for providing the equipment and personal assistance and the informants for help and patience during the Image Acquisition session.

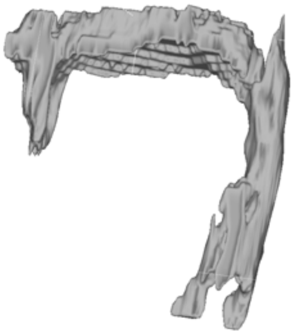

Fig. 6 - ITK Snap 3D segmentation for the vowel [a].

REFERENCES

[1] S. Narayanan, K. Nayak, S. Lee, and D. Byrd, “An approach to real-time magnetic resonance imaging for speech production”, *J Acoust Soc Am.*, vol. 115, no. 4, pp. 1771–1776, 2004.

[2] A. Soquet, V. Lecuit, T. Metens, B. Nazarian, and D. Demolin, “Segmentation of the airway from the surrounding tissues on magnetic resonance images: A comparative study”, *Proc. 5th ICSLP*, 1998.

[3] A. Serrurier and P. Badin, “Towards a 3D articulatory model of the velum based on MRI and CT images”, *ZAS Papers in Linguistics*, vol. 40, pp. 195–211, 2005.

[4] O. Engwall, “Are static MRI measurements representative of dynamic speech? Results from a comparative study using MRI, EPG and EMA”, *Proc. 6th ICSLP*, vol. I, pp. 17–20, 2000.

[5] H. Takemoto, T. Kitamura, H. Nishimoto, and K. Honda, “A method of tooth superimposition of MRI data for accurate measurement of vocal tract shape and dimensions”, *Acoustical Science and Technology*, vol. 25, no. 6, pp. 468–474, 2004.

[6] A. J. Beer, P. Hellerhoff, A. Zimmermann, K. Mady, R. Sader, E. J. Rummeny, and C. Hannig, “Dynamic near-real-time magnetic resonance imaging for analyzing the velopharyngeal closure in comparison with videofluoroscopy”, *J Magn Reson Imaging*, vol. 20, no. 5, pp. 791–797, 2004.

[7] S. Rua and D. Freitas, “Morphological dynamic study of human vocal tract”, in *CompIMAGE*, Coimbra, Portugal, 2006.

[8] E. Bresch, J. Adams, A. Pouzet, S. Lee, D. Byrd, and S. Narayanan, “Semi-automatic processing of real-time MR image sequences for speech production studies”, in *7th Int. Seminar Speech Prod.*, Ubatuba, Brazil, 2006.

[9] R. Adams and L. Bischof, “Seeded region growing”, *IEEE Trans Pattern Anal Mach Intell*, vol. 16, no. 6, pp. 641 – 647, 1994.

[10] B. Santos, C. Ferreira, J. Silva, A. Silva, and L. Teixeira, “Quantitative evaluation of a pulmonary contour segmentation algorithm in X-ray computed tomography images”, *Academic Radiology*, vol. 11, no. 8, pp. 868–878, 2004.

[11] I. Middleton and R. I. Damper, “Segmentation of magnetic resonance images using a combination of neural networks and active contour models”, *Medical Engineering & Physics*, vol. 26, no. 1, pp. 71–86, 2004.

[12] J. Dang, M. Honda, and K. Honda, “Investigation of coarticulation in continue speech in japanese”, *Acoustical Science and Technology*, vol. 25, no. 5, pp. 318–329, 2004.

[13] P. Yushkevich, J. Piven, H. Heather, R. Smith, S. Ho, J. Gee, and G. Gerig, “User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability”, *Neuroimage*, 2006, To appear.