Segmentation and 3D reconstruction of the vocal tract from MR images – a comparative study

S. R. Ventura
School of Allied Health Science – Porto Polytechnic Institute, V.N. Gaia, Portugal

D. R. Freitas
Faculty of Engineering, University of Porto, Porto, Portugal

I. M. Ramos
Radiology Service, São João Hospital and Faculty of Medicine, University of Porto, Porto, Portugal

João Manuel R. S. Tavares
Faculty of Engineering, University of Porto / Institute of Mechanical Engineering and Industrial Management Porto, Portugal

ABSTRACT: Speech production is an important human function involving a set of organs with specific morphological and dynamic aspects. The inter-speaker variability, the coarticulation or the nasality are some interesting aspects to improve a realistic 3D modeling of the vocal tract. For this, the understanding of the mechanism of speech production is crucial, as the current image data is not sufficient to reproduce truthfully the speaker's anatomy and articulation. Hence, the goal of 3D modeling is to generate the complete geometrical and dynamical information concerning the vocal tract from medical images, such as from magnetic resonance imaging (MRI). This work aims to describe and compare two different segmentation techniques to attain the 3D shape of the vocal tract during speech production from MR images: the former based on manual tracing of the vocal tract contours and the latter based on image thresholding. Thus, the segmented cross-sectional areas were measured, and 3D models were built from the sagittal data by blending the contours obtained from the two segmentation techniques. The mean error of the measures computed were low for both segmentation techniques, which let us conclude that the techniques are useful to evaluate the vocal tract geometry accurately. Additionally, the 3D models built using both segmentation techniques were also very similar and truthful. However, when the coronal data was used, various difficulties occurred.

1 INTRODUCTION

Magnetic resonance imaging (MRI) is a promising technique in all clinical research fields; the multiplanar imaging acquisition, high soft-tissue resolution and safety of MRI are some of the most important advantages for its use (Shadle et al. 1999; Narayanan et al. 2004). Vocal tract morphology is one of the main aspects to be considered during speech articulation that confers potential inter-speaker differences (Fuchs, Winkler, and Perrier 2008).

Until now, the knowledge concerning the morphology and articulation measurements of the vocal tract, based on MRI, is not sufficient to reproduce accurately the speaker's anatomy (Birkholz and Kroger 2006). However, this knowledge is demanded by different areas, such as bioengineering, medicine or speech therapy (Ventura, Freitas, and Tavares 2009).

Three-dimensional (3D) imaging based on magnetic resonance is essential to acquire the full geometry of the vocal tract, and thus to provide better knowledge concerning vocal tract shape and to obtain more data its realistic 3D modeling (Bresch et al. 2008; Apostol et al. 1999; Badin and Serrurier 2006; Kim, Narayanan, and Nayak 2009).

In order to study the vocal tract from MRI data, the acquired images must be processed considering the following main steps: image segmentation and 3D shape reconstruction. The former is the most important and difficult; mainly, because there are common problems related with the determination of the air-tissue boundaries, as previously reported by (Demolin, Metens, and Soquet 1996) and (Soquet et al. 1998), and is decisive to obtain suitable 3D models in the second step. For example, to achieve the completion of these tasks, (Serrurier and Badin 2005; Badin et al. 1998) extracted the vocal tract shape manually and the 3D mesh reconstruction was realized by intersecting these contours along a semi-polar grid. However, manual edition, besides of being very arduous and time consuming, is extremely user-dependent (increasing the uncertainty of results). On the other hand, (Behrends and Wismuller 2001) introduced a simple algorithm based on 3D region growing; and (Narayanan et al. 2004) focuses on methods to automatically segment and track the real-time MRI data using Kalman snakes and optical flow.

Despite the level of automation in performing the image processing tasks, the analysis of vocal tract from images remains complex, given the various problems that persist. Some of these problems are related with the MR image acquisition technique and with the morphology of the vocal tract (i.e. the non-identification of teeth and the similarity of signal intensity between the vocal tract and of some of the surrounding structures).

According to (Soquet et al. 1998), that presents a comparative study to assess the accuracy of three different segmentation techniques, the image thresholding method revealed an inferior dispersion, but the overall results presented small average error and large error distribution.

In this work, we describe and compare two different segmentation techniques to extract the contours to be used in 3D reconstruction of the vocal tract during speech production from MR images: one based on the manual tracing of the segmentation contours, and another based on image threshold.

The remaining of this paper is organized as follows. In the next section, the description of the MRI protocol, the speech corpus and the image analysis and assessment are described. Then, in section three, the area measurements obtained by using the two segmentation techniques under comparison are presented and discussed, and the 3D models built for the vocal tract are shown and examined. Finally, the conclusions are pointed out in the last section.

2 METHODS

According to the usual regulated safety procedures for MRI, the subject was previously informed about the undergoing imaging exam and subsequently instructed about the procedures to be adopted, whereupon the subject signed a consent form.

2.1 *MRI protocol and speech corpus*

This study was performed using a 1.5T Siemens Magneton Symphony system and a head array coil, involving one healthy young male, with speech therapy skills. Using a fast image acquisition mode, we collected an overall of seven slices in two different image orientations, considering the following acquisition parameters: field of view 150 mm, image matrix 128x128 and image resolution 0.853 px/mm.

The subject sustained the articulation during 9 s for the acquisition of three sagittal slices (slice thickness of 5 mm) and 9.9 s for the four coronal slices (slice thickness of 6 mm and with 10 mm of gap between slices) for each speech sound. The acquisition time was defined adopting a compromise between image resolution and the duration of the sustained articulation allowed by the subject.

The speech corpus consisted in four European Portuguese sounds: the vowels [i] and [u] and the lateral consonants.

Due to the MR acoustic noise produced during the acquisition process, the speech recording had not enough quality, and therefore, it was discarded.

2.2 *Image analysis and assessment*

The analysis of the 2D images was performed using *ImageJ - Image Processing and Analysis in Java* (from NIH, USA - http://rsb.info.nih.gov/ij/). Subsequently, the 3D models of the vocal tract were built using *Blender* (from Blender Foundation, Amsterdam, the Netherlands - http://www.blender.org/cms/Home.2.0.html).
The vocal tract region was extracted from each image slice using two segmentation techniques:
(1) Manual tracing of the segmentation contours based on Bézier curves;
(2) Manual imaging threshold.

The contour segmentation process resulted in a total of 28 2D contours, i.e. seven contours for each sound. Then, the resultant contours from all cross-sectional images were measured, Figure 1.

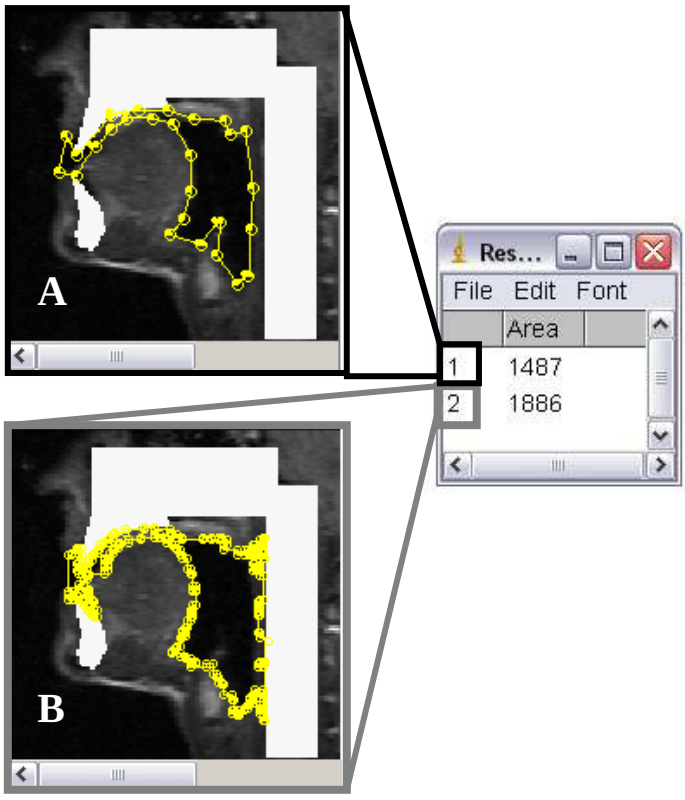

Figure 1. Measurement of the cross-sectional areas of the vocal tract obtained using manual tracing of the segmentation contours (A) and manual imaging threshold (B).

The contours obtained by the two segmentation techniques were then used to build the 3D skin mod-

els, after importing the contours in *.shapes* format, into the *Blender* software.

3 RESULTS AND DISCUSSION

3.1 *Cross-sectional area measurements*

The measurements on the cross-sectional areas, attained using manual tracing of the segmentation contours and manual imaging threshold in each image slice (Si), are indicated in Tables 1-3.

The manual tracing segmentation was taken as reference, mainly because the user can take into account its knowledge about the vocal tract anatomy in an easier way.

Table 1. Measurements (mm^2) on the cross-sectional areas obtained by manual tracing of the segmentation contours.

Speech Corpus	Sagittal			Coronal			
	S1	S2	S3	S1	S2	S3	S4
[i]	1487	1540	1568	185	53	43	856
[u]	1686	1826	1911	379	973	134	43
[L]	1358	1650	1662	368	678	186	83
[lh]	1603	1812	1811	337	368	100	1294

Table 2. Measurements (mm^2) on the cross-sectional areas obtained by manual imaging threshold.

Speech Corpus	Sagittal			Coronal			
	S1	S2	S3	S1	S2	S3	S4
[i]	1886	2022	1649	370	84	99	733
[u]	1480	1748	1906	404	969	114	57
[L]	1202	1612	1651	520	738	121	161
[lh]	1343	1577	1694	424	236	196	1395

Tables 1 and 2 indicate the areas measured using the segmentation techniques based on manual tracing and imaging threshold, respectively. The sagittal slice 2 was situated at the midsagittal anatomic plane (a plane passing vertically through the midline, dividing the body into left and right parts). The slice 1 and 3 were situated at right and left from slice 2, respectively.

The coronal (or frontal) plane was a vertical section that divides the body into anterior and posterior; sections 1, 2, 3 and 4 were situated in the oral cavity, from front to back, showing areas for the lips (S1), tongue apex (S2), tongue body (S3) and oropharynx (S4). The unrounded vowel [i] area at the lips plane (S1) is inferior to the area of rounded vowel [u] for both segmentation techniques. Similarly, these two vowels have different areas in slice 4: front vowel [i] area is superior to the back vowel [u] area.

In lateral consonants, an occlusion was observed somewhere along the tongue, while air was escaping at one or both sides of the tongue. As it can be realized from Tables 1 and 2, the coronal sections 2 and 3 of the palatal lateral approximant [lh] are inferior than slice areas 1 and 4 for both segmentation techniques. The consonant [L] is pronounced by the tongue's approximation to the velar region, as a result, the sections 3 and 4 areas are inferior when compared with sections 1 and 2 areas.

The average areas and standard deviation of the measurement errors between the two segmentation techniques under comparison are depicted in Table 3.

Table 3. Mean cross-sectional areas and relative errors of the measurements (mm^2).

Speech Corpus	Average sagittal area		Average coronal area	
	Relative Error	SD	Relative Error	SD
[i]	0.2094	147.63	0.1310	82.429
[u]	0.0533	101.74	0.0098	1.3066
[L]	0.0439	76.582	0.1711	34.969
[lh]	0.1171	58.342	0.0724	37.050

SD – Standard deviation.

For both segmentation techniques, the results indicate in Table 3 allow to conclude that the mean error is relatively small and the distribution of the errors is highly dispersed.

3.2 *3D vocal tract models*

From the contours extracted it is possible to reconstruct the 3D vocal tract shape. Figure 2 depicts the 3D vocal tract models built from the cross-sectional sagittal data segmented by the two segmentation techniques.

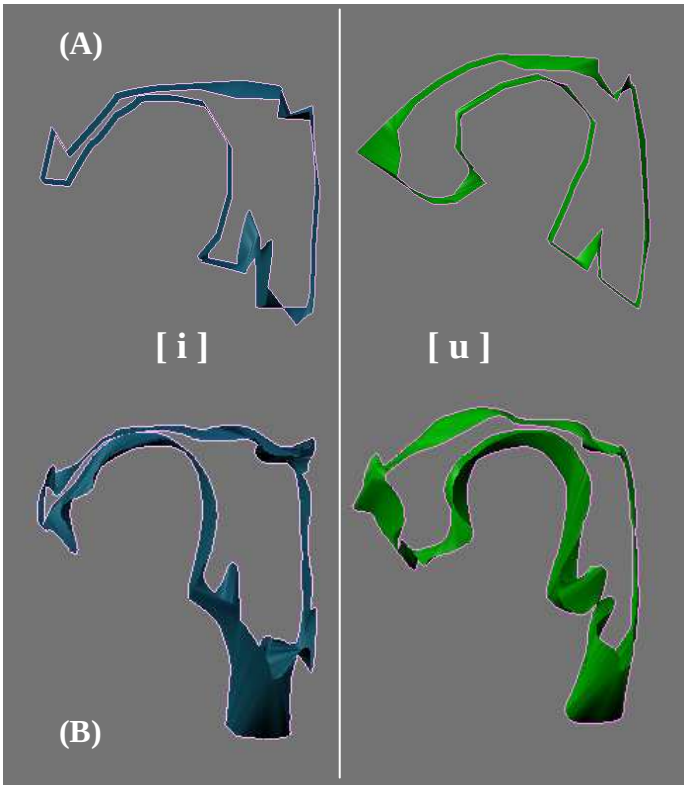

Figure 2. 3D vocal tract models built for vowels [i] and [u] from manual tracing of the segmentation contours (A) and manual imaging threshold (B).

The 3D vocal tract meshes created using the two segmentation techniques are very similar; i.e. the articulatory organs shape and position for these two vowels are identical. However, the non-rigid geometry of the vocal tract model generated using the segmentation technique based on manual imaging threshold provides a more likely anatomic shape.

The coronal data is particularly important to realize the lateral dimension of the oral cavity and the tongue's position. The 3D vocal tract models generated by the interpolation of cross-sectional contours of lateral consonants are shown in Figures 3 and 4.

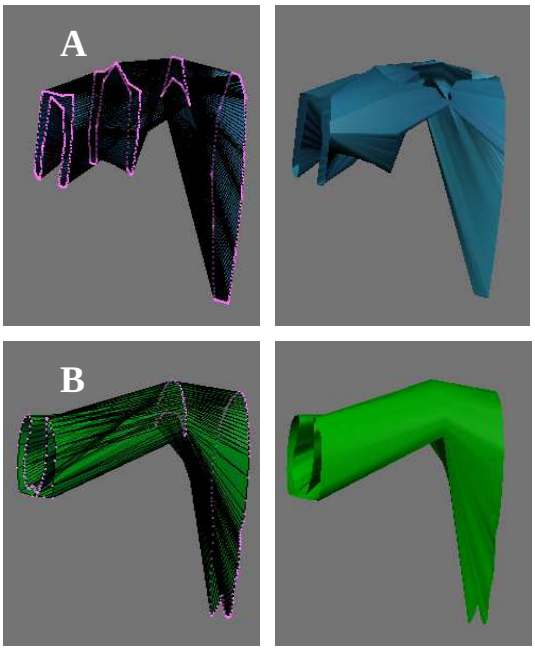

Figure 3. 3D vocal tract models built for consonant [lh] from manual tracing of the segmentation contours (A) and manual imaging threshold (B).

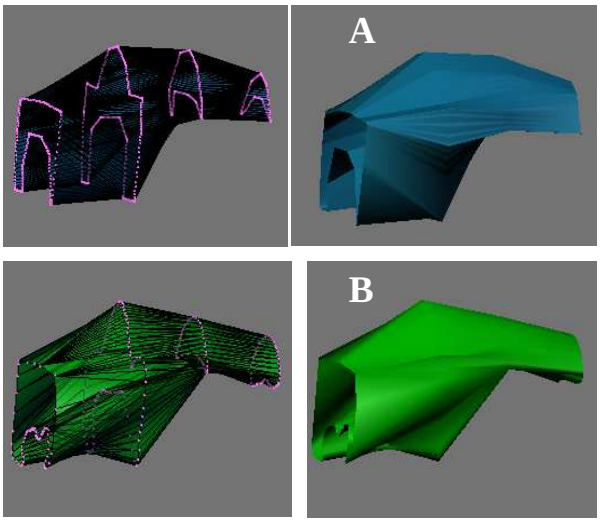

Figure 4. 3D vocal tract models built for consonant [L] from manual tracing of the segmentation contours (A) and manual imaging threshold (B).

Contrasting with the 3D models built from sagittal data, the vocal tract shapes built applying the two segmentation techniques (Figures 3 and 4) on the coronal data are reasonably different. Despite the more likely anatomic shape provided by 3D models built from the segmentations done by manual imaging threshold, the lateral dimensions of the vocal tract are overestimated, probably because the non-identification of the teeth and the similarity of signal intensity between the vocal tract and of some of the surrounding structures (e.g. nasal cavity). Consequently, the 3D model of the vocal tract built from coronal data must be carefully analyzed in order to avoid inaccurate measures.

4 CONCLUSIONS

In this work, we compared the measurements attained on cross-sectional areas obtained by two different image segmentation techniques, one based on manual contours tracing and another one based on imaging thresholding, in MR images acquired during speech production. Additionally, 3D models were built by blending the segmented contours.

For both segmentation techniques under comparison, the measurement of the segmented cross-sectional regions allowed to attain accurate data concerning the vocal tract geometry during speech production. In fact, low mean errors and errors distributed in a very dispersed manner were verified.

The success of the 3D modelling is conditioned by the segmentation technique used. By one hand, the technique based on the manual tracing of the segmentation contours gives results more accurate, while in the other hand, is more dependent on the user's skill and is more time consuming. The inferior dependency in the user's skill and the higher level of automation are advantages of the segmentation technique based on imaging threshold, but the vocal tract geometry reconstructed from this technique must be carefully analyzed as it can be inaccurate, particularly when applied on coronal data.

MRI data interpretation is difficult and some problems remain to be solved in order to obtain more realistic and accurate 3D models, as the correct identification of the teeth and of the vocal tract limits. On the other hand, the low number of MR slices that are always acquired has a negative impact in the resultant 3D models.

5 ACKNOWLEDGMENTS

The images considered in this work were acquired at the Radiology Department of the S. João Hospital, in Portugal, and we would like to express to our gratitude to the technical staff.

The first author would like to thank the support and contribution of the PhD grant from IPP – Insti-

tuto Politécnico do Porto and ESTSP – Escola Superior de Tecnologia da Saúde, in Portugal.

6 REFERENCES

Apostol, Lian, Pascal Perrier, Monica Raybaudi, and Christoph Segebarth. 1999. "3D Geometry of the Vocal Tract and Interspeaker Variability. "pp. 443-446 in *14th International Congress of Phonetic Sciences (ICPhS 99)*. San Francisco, USA.

Badin, P, G Bailly, M Raybaudi, and C Segebarth. 1998. "A Three-dimensional Linear Articulatry Model Based on MRI Data. "pp. 417-420 in *5th International Conference on Spoken Language Processing*, Eds R.H. Mannell & J. Robert-Ribes. Sydney, Australia.

Badin, Pierre, and Antoine Serrurier. 2006. "Three-dimensional modeling of speech organs: Articulatory data and models." *Transactions on Technical Committee of Psychological and Physiological Acoustics* 36:421-426.

Behrends, Johannes, and Axel Wismuller. 2001. "A Segmentation and Analysis Method for MRI Data of the Human Vocal Tract." *Forschungsberichte des Instituts fur Phonetik und Sprachliche Kommunikation der Universitat Phonetik und at Munchen (FIPKM)* 37:179-189.

Birkholz, Peter, and Bernd J Kroger. 2006. "Vocal Tract Model Adaptation Using Magnetic Resonance Imaging. "p. 493–500 in *7th International Seminar on Speech Production*. Ubatuba, Sao Paolo, Brazil.

Bresch, Erik, Yoon-chul Kim, Krishna Nayak, Dani Byrd, and Shrikanth Narayanan. 2008. "Seeing Speech: Capturing Vocal Tract Shaping and Real-time Magnetic Resonance Imaging." *IEEE Signal Processing Magazine* 123-129.

Demolin, D, T Metens, and A Soquet. 1996. "Three-dimensional measurement of the vocal tract by MRI. "pp. 272-275 in *4th Internat. Conf. on Spoken Language Processing (ICSLP 96)*, vol. 1p. Philadelphia, USA.

Fuchs, Susanne, Ralf Winkler, and Pascal Perrier. 2008. "Do Speakers' Vocal Tract Geometries Shape their Articulatory Vowel Space?. "pp. 333-336 in *8th International Seminar on Speech Production (ISSP)*. Strasbourg, France http://halshs.archives-ouvertes.fr/hal-00370715/.

Kim, Yoon-chul, Shrikanth S Narayanan, and Krishna S Nayak. 2009. "Accelerated Three-Dimensional Upper Airway MRI Using Compressed Sensing." *Magnetic Resonance in Medicine* 61:1434-1440.

Narayanan, S, K Nayak, S Lee, A Sethy, and D Byrd. 2004. "An Approach to Real-time Magnetic Resonance Imaging for Speech Production." *Journal Acoustical Society of America* 115:1771-76.

Serrurier, A, and P Badin. 2005. "Towards a 3D articulatory model of the velum based on MRI and CT images." *ZAS Papers Linguistics* 40:195-211.

Shadle, C.H., M. Mohammad, J.N. Carter, and P.J.B. Jackson. 1999. "Multi-planar Dynamic Magnetic Resonance Imaging: New Tools for Speech Research. "pp. 623-626 in *International Congress of Phonetics Sciences (ICPhS99)*. San Francisco.

Soquet, A., V. Lecuit, T. Metens, B. Nazarian, and D. Demolin. 1998. "Segmentation of the Airway from the Surrounding Tissues on Magnetic Resonance Images: A Comparative Study. "pp. 3083-3086 in *5th International Conference on Spoken Language Processing (ICSLP 98)*. Sydney, Australia.

Ventura, Sandra Rua, Diamantino Freitas, and João Manuel R S Tavares. 2009. "Application of MRI and biomedical engineering in speech production study Application of MRI and biomedical engineering in speech production study." *Computer Methods in Biomechanics and Biomedical Engineering* 12:671-681.