Three-dimensional modeling of tongue during speech using MRI data, Sandra Rua Ventura, Diamantino R. Freitas, João Manuel R. S. Tavares, CMBBE 2008 - 8th International Symposium on Computer Methods in Biomechanics and Biomedical Engineering, 6 pag., Porto, Portugal, 27th February-1st March 2008

THREE-DIMENSIONAL MODELING OF TONGUE DURING SPEECH USING MRI DATA

Sandra Rua Ventura[1], Diamantino Rui Freitas[2] and João Manuel R. S. Tavares[3]

1. ABSTRACT

The tongue is the most important and dynamic articulator for speech formation, because of its anatomic aspects (particularly, the large volume of this muscular organ comparatively to the surrounding organs of the vocal tract) and also due to the wide range of movements and flexibility that are involved. In speech communication research, a variety of techniques have been used for measuring the three-dimensional vocal tract shapes. More recently, magnetic resonance imaging (MRI) becomes common; mainly, because this technique allows the collection of a set of static and dynamic images that can represent the entire vocal tract along any orientation. Over the years, different anatomical organs of the vocal tract have been modelled; namely, 2D and 3D tongue models, using parametric or statistical modelling procedures. Our aims are to present and describe some 3D reconstructed models from MRI data, for one subject uttering sustained articulations of some typical Portuguese sounds. Thus, we present a 3D database of the tongue obtained by stack combinations with the subject articulating Portuguese vowels. This 3D knowledge of the speech organs could be very important; especially, for clinical purposes (for example, for the assessment of articulatory impairments followed by tongue surgery in speech rehabilitation), and also for a better understanding of acoustic theory in speech formation.

2. INTRODUCTION

The human speech production is a complex and individual mechanism due to the different anatomical structures involved and organs movements implicated. The vocal tract has some important features that are constrained to medical image techniques observation, namely the non-linear shape (similar to a tube L-shaped) and the significant length. Furthermore, the vocal tract organs or articulators change their positions during speech causing shape variation in this tube and subsequently in the air flow. The tongue is the most important and dynamic articulator for speech formation mostly because the large volume of this muscular organ and the wide range of movements and flexibility. For this reason, it is difficult to measure the movements of human tongue and to determine the surface deformation.

Many approaches have been used for measuring speech and vocal tract shapes, appearing with very useful results the magnetic resonance imaging. This image technique allows morphologic measurements in static [1-4] and also in dynamic studies [5-9], and can represent the whole vocal tract along any orientation with good soft

[1]Professor, Department of Radiology, School of Allied Health Science - IPP, Praça Coronel Pacheco 15, 4050-453 Porto, Portugal

[2] Professor, DEEC, FEUP - Faculty of Engineering of University of Porto, Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal

[3] Professor, DEMEGI, FEUP

tissues resolution.
A number of articulatory models of the tongue have been proposed with two main distinctions: (1) physiological approaches aiming the understanding and modelling the muscular structure and functions of the tongue; and (2) geometrical or statistical approaches based in assumptions on the geometry of articulators and movements (without a direct coupling between measurements and the model).
For example, Stone et al. (2000) used tagged cine-MRI to examine the internal deformation of the tongue during speech. These authors observed the principal strains detailed regions of compression and extension in the tongue and showed the linkage between internal tissue strain and surface deformation [10]. Dang and Honda (2000) reconstructed the tongue geometry from MR images and characterized their dynamic using x-ray microbeam data [11]. Several different measurement sources have been used to created 3D models of tongue, lips and face, based on MRI and video images [12], and subsequent based in the combination of MRI, electromagnetic articulography and electropalatography, for shape and parameters determination through statistical analysis [13]. Other studies have proposed an alternative method for tongue modelling using finite element modelling [14,15]. More recently, Doel et al. (2006) proposed geometrical models of the vocal tract using MRI [16]. The authors presented an approach that includes dynamical interaction between the tongue (modelled by fast 3D finite element modelling) and the air flow in the vocal tract (airway).
Our purpose is to present and describe some 3D reconstructed models from MRI data of some European Portuguese sounds, by means of stacks combination. This paper is organized in five sections. The methods section describes the equipment, corpus and subjects, and the procedures used for the speech study, namely for morphologic imaging of the vocal tract. The results are presented in next section, through the exhibition of some 2D contours and some three-dimensional models of the tongue. Finally, we present the conclusions and future work.

3. METHODS

The image data was collected from one male training subject, in supine position without speech disorders and using an MRI Siemens Magneton Symphony 1.5T system. The corpus consisted in nine sounds of European Portuguese (EP): five oral vowels, two nasals and lateral consonants. Because of the MRI environment, the acoustic recording of the produced speech was not possible during the images acquisitions.

3.1 MRI protocol for image acquisition

For each 3D model, a set of seven MR WT1-images using TSE sequences was achieved in sagittal and coronal orientations. In this study, the subject sustained the articulation during 9 seconds for the acquisition of three sagital slices and 9.9 seconds for the four coronal slices. The time acquisition is a compromise between image resolution and the time needed for sustained articulation allowed by the subject.
For the slice position we use as reference points the lips center. This procedure was realized to ensure the same location points among sounds and for stacks registration.
Sagittal stacks consists in a set of three contiguous slices with 5 mm thickness, one positioned at the midsagittal plan and the others at each side (right and left), as demonstrated in Figure 1. This slice orientation gives information about the length and shape of the vocal tract and namely tongue’s position.

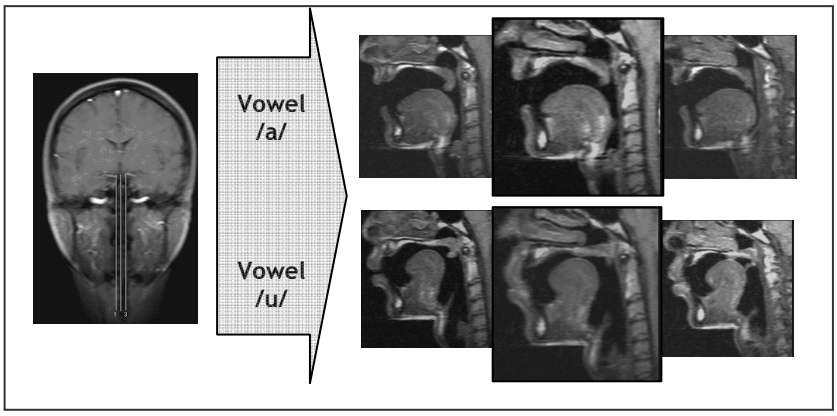

Fig. 1 Sagittal slices programmed in a reference image (left) and two examples (right) of the set of sagittal images for two vowels of EP.

The coronal stack consists in a set of four slices with 6 mm thickness and 16 mm spaced. The first slice started at lips center level and the others slices spaced back (Figure 2). This slice orientation gives lateral information of the vocal tract and namely lips shape and tongue's.

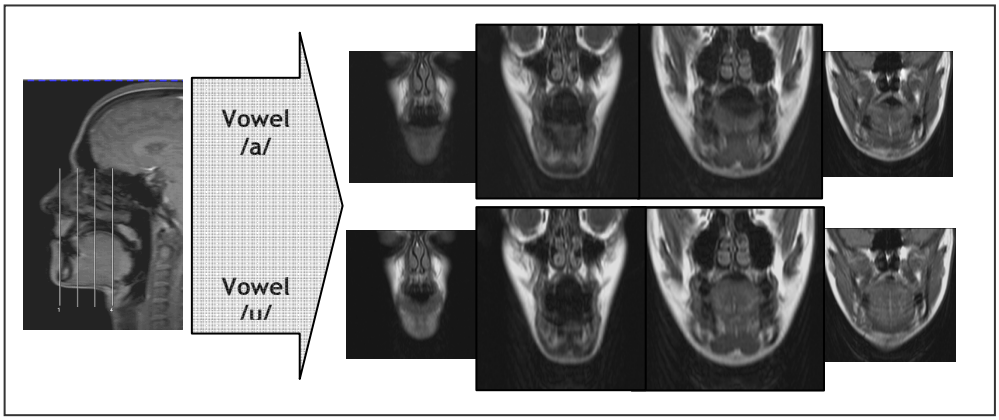

Fig. 2 Coronal slices programmed in a reference image (left) and two examples (right) of the set of coronal images for two vowels of EP.

3.2 Image segmentation and 3D reconstruction

Image analysis and 3D model construction was accomplished in two stages, namely: (1) image segmentation using the S*egmenting Assistant*, a 3D editing plugin of *Image J* the image processing software developed by the National Institute of Health (http://rsb.info.nih.gov/ij/) and subsequent 3D reconstruction, and (2) graphic representation and combination of orthogonal stacks using the *Blender* software for 3D graphics creation, version 2.41 (http://www.blender.org/).
The tongue contour was manually extracted from each image using interpolation with Bézier contours. The whole tongue was segmented as one unit and without including the epiglottis, as depicted in Figure 3.

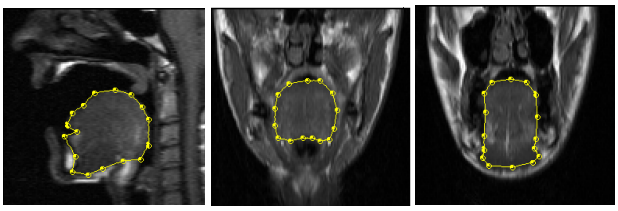

Fig. 3 Tongue contours extraction realized in sagittal and coronal images.

The contour extraction process resulted in a total of 63 planar contours (maximally 7 contours for each sound).
Outlines were subsequently used to generate a skin (a 3D object), after importing the contours in *.shapes* format, into the *Blender* software.
For each articulatory position, the next phase was the combination of sagittal and coronal outlines (2D curves). To make this possible, it is required that the outlines be well aligned – this process is usually called image registration. In computational vision, the term image registration means the process of transforming the different sets of images into one coordinate system, what is necessary in order to be able to compare or integrate the data obtained from different measurements.

4. RESULTS

The following images (Figure 4) represent different perspectives of the 3D model obtained for the vowel /a/. The blue skin represents the union of the three outlines extracted from the sagittal stack. The red skin represents the union of the three outlines extracted from the coronal stack. It must be observed that these reconstructions are not yet closed as should be for tongue reconstruction. This closing of the skin by unification of the different stacks is the next step in the processing.

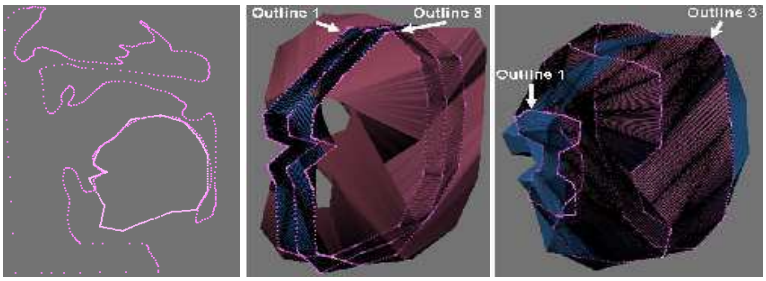

Fig. 4 Comparison between the 2D contour extracted from the midsagittal MR image (left) and two views of the 3D model generated for the vowel /a/ of EP.

Figure 5 depicts the extracted midsagittal contours and the resulting 3D models obtained during EP vowels articulation. Comparing this data we can see that the tongue shape and dimensions in oral cavity are different among sounds: the tongue moves from front to center and from center to back for the vowels [i, e, a] and [a, o, u] respectively.

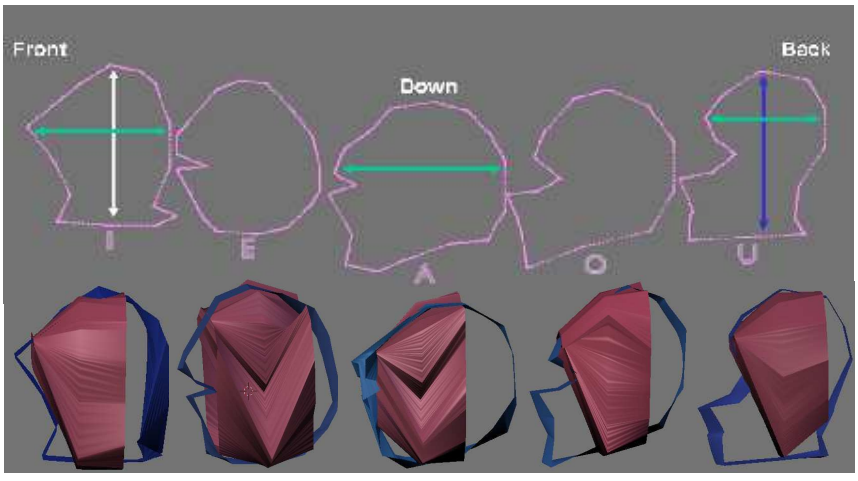

Fig. 5 Extracted midsagittal contours and 3D models generated for the five vowels of EP.

The following models in Figure 6 demonstrate the 3D models generated for the nasals and laterals consonants of EP. As observed, the coronal data is especially useful in the characterization of laterals consonants, where the tongue is in almost total contact with the palate at central position, and the air escapes by the tongue's side.

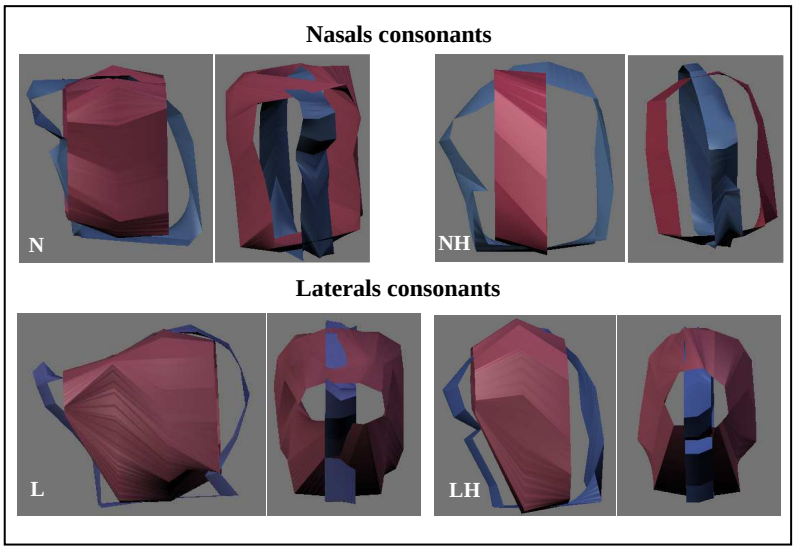

Fig. 6 Three-dimensional models of EP consonants.

5. CONCLUSIONS AND FUTURE WORK

This paper presented some 3D tongue shape extracted from MR images by means of combination of stacks: representative slices in sagittal and coronal orientations. It should be noted that these 3D tongue models are not complete yet and much work remains to be done. However, comparing with previous works, this study use a new approach for tongue modelling from MRI, with short time acquisitions minimizing subject effort during the sustentation of the sounds. Furthermore, until now no other study of characterization of the tongue gestures in EP based on MRI was found in literature. This 3D knowledge of the speech organs can be very important for a better understanding of the acoustic speech production and also, for clinical purpose (assessment of articulatory impairments followed by tongue surgery in speech rehabilitation).

The future work is to improve the MRI protocol (increasing number of slices), the image analysis (reducing the time need and user intervention) and the 3D models (namely closing the 3D models between stacks). This data is intended as a contribution needed for the construction of 3D articulatory models for speech synthesis.

6. REFERENCES

1. Apostol, L., Perrier, P., Raybaudi, M., Segebarth, C., 3D Geometry of the Vocal Tract and Inter-speaker Variability, Proceed. 14th International Congress of Phonetic Sciences (ICPhS), San Francisco, USA, 1999, 443-446.
2. Badin, P., Serrurier, A., Three-dimensional Modeling of Speech Organs: Articulatory Data and Models, IEICE Technical Committee on Speech, Kanazawa, Japan, 2006, 29-34.
3. Kröger, B.J., Winkler, R., Mooshammer, C., Pompino-Marschall, B, Estimation of Vocal

Tract Area Function from Magnetic Resonance Imaging: Preliminary Result, Proceed. 5th Seminar On Speech Production: Models And Data, München, Germany, 2000, 333-336.
4. Serrurier, A. & Badin, P., A Three-dimensional Linear Articulatory Model of Velum based on MRI data, Interspeech 2005: Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 2005, 2161-2164.
5. Abbott, M.B., Dardzinski, B.J., Donnelly, L.F., Using volume segmentation of cine MR data to evaluate dynamic motion of the airway in pediatric patients, AJR American Journal of Roentgenology, 2003, 181 (3), 857-859.
6. Avila-García, M.S., Carter, J.N., Damper, R.I., Extracting Tongue Shape Dynamics from Magnetic Resonance Image Sequences, Transactions on Engineering, Computing and Technology V2, 2004, 288-291.
7. Mády, K., Sader, R., Zimmermann, A., Hoole, P., Beer, A., Zeilhofer, H., Hannig, C., Assessment of Consonant Articulation in Glossectomee Speech by Dynamic MRI, Proceed. of 7th International Conference on Spoken Language Processing (ICSLP), Denver, USA, 2002.
8. Narayanan, S., Nayak, K., Lee, S., Sethy, A., Byrd, D., An Approach to Real-time Magnetic Resonance Imaging for Speech Production, Journal Acoustical Society of America, 2004, 115(4), 1771-1776.
9. Takemoto, H., Honda, K., Measurement of Temporal Changes in Vocal Tract Area Function during a continuous vowel sequence using a 3D Cine-MRI Technique, Proceed. 6th Int. Seminar on Speech Production, Sydney, Australia, 2003, 284-289.
10. Stone**,** M., Dick, D., Davis, E., Douglas, A., and Ozturk, C., Modelling the Internal Tongue Using Principal Strains, Proceedings of the Fifth Speech Production Seminar, Kloster-Seeon, Germany, 2000, 133-136.
11. Dang, J., Honda, K., Improvement of a Physiological Articulatory Model for Synthesis of Vowel Sequences Proceedings, Sixth International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China, 2000, vol.1, 457-460.
12. Badin, P., Borel, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., Towards an audiovisual virtual talking head: 3D articulatory modeling of tongue, lips and face based on MRI and video images, Proceed. 5th Speech Production Seminar, München, Germany, 2000, 261-264.
13. Engwall, O., Combining MRI, EMA & EPG measurements in a three-dimensional tongue model, Speech Communication, 41, 2003, 303-329.
14. Gérard, JM., Ohayon, J., Luboz, V., Perrier, P., Payan, Y., Indentation for Estimating the Human Tongue Soft. Tissues Constitutive Law: Application to a 3D. Biomechanical Model, Proceedings of Medical Simulation: International Symposium, ISMS 2004, Cambridge, MA, USA, 2004, 77-83.
15. Perrier, P., Payan, Y., Zandipour, M., Perkell, J., Influences of tongue biomechanicson speech movements during the production of velar stop consonants: A modeling study, JASA 114(3), 2003, 77-83.
16. Doel, K., Vogt, F., English, E., Fels, S., Towards Articulatory Speech Synthesis with a Dynamic 3D Finite Element Tongue Model, Proceedings of the 7th International Seminar on Speech Production, Brazil (ISSP 06), 2006, 59-66.

7. ACKNOWLEDGMENT

Images were acquired at the Radiology Department of Hospital S. João, Porto, with the collaboration of Isabel Ramos (Professor of Faculdade de Medicina da Universidade do Porto and Department Director) and the technical staff that is gratefully acknowledged.