MAGNETIC RESONANCE IMAGING OF THE VOCAL TRACT: TECHNIQUES AND APPLICATIONS

Sandra M. Rua Ventura
Área Científico-pedagógica da Radiologia, School of Allied Health Science – IPP
Rua Valente Perfeito 322, 4400-330 Vila Nova de Gaia, Portugal
smr@estsp.ipp.pt

Diamantino Rui S. Freitas, João Manuel R. S. Tavares
FEUP – Faculty of Engineering of University of Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal
dfreitas@fe.up.pt, tavares@fe.up.pt

Keywords: Magnetic Resonance Imaging, Image Processing, 3D modeling, Vocal Tract Study, Speech Production.

Abstract: Magnetic resonance (MR) imaging has been used to analyse and evaluate the vocal tract shape through different techniques and with promising results in several fields. Our purpose is to demonstrate the relevance of MR and image processing for the vocal tract study. The extraction of contours of the air cavities allowed the set-up of a number of 3D reconstruction image stacks by means of the combination of orthogonally oriented sets of slices for each articulatory gesture, as a new approach to solve the expected spatial under sampling of the imaging process. In result these models give improved information for the visualization of morphologic and anatomical aspects and are useful for partial measurements of the vocal tract shape in different situations. Potential use can be found in Medical and therapeutic applications as well as in acoustic articulatory speech modelling.

Magnetic Resonance (MR) improvements, in the past decades, allowed vocal tract imaging, making it currently one of the most promising tools in speech research.

Speech is the most important instrument of human communication and interaction. Nevertheless, the knowledge about its production is far from being complete or even sufficient to describe the most relevant acoustic phenomena that are conditioned at morphological and dynamic levels. The anatomic and physiologic aspects of the vocal tract are claimed to be essential for a better understanding of this process. The quality and resolution of soft-tissues and the use of non-ionizing radiation are some of the most important advantages of MR imaging (Avila-García et al., 2004; Engwall, 2003).

Several approaches have been used up to now for the study of the vocal tract based on MR images. Since the first study proposed by Baer et al. (1991), many MR techniques have been used (from static to dynamic studies, and more recently even done in real-time), starting by studies of vowel production (Badin et al., 1998; Demolin et al., 2000), followed by consonant production (Engwall, 2000b; Narayanan et al., 2004), for different languages such as French (Demolin et al., 1996; Serrurier & Badin, 2006), German (Behrends et al., 2001; Mády et al., 2001), and Japonese (Kitamura et al., 2005; Takemoto et al., 2003).

The work presented in this paper, consisting basically in the static description of the vocal tract shape during sustained vowels and consonants and in the dynamic description of some syllables, is the first to report the application of MR imaging for the characterization of European Portuguese (EP). This study started in 2004, having attained a first series of results published in 2006 (Rua & Freitas, 2006). Our approach can be seen as a contribution to the wide area of articultory speech modeling, since it provides geometrical data to the acoustic modeling phase or research.

In the articulatory speech research of EP a few studies of nasal vowels have been carried through, at the acoustic production and perceptual levels based on acoustic analysis and electromagnetic

articulography (Teixeira et al., 2001, 2002, 2003). More recently, another MR study of EP presents some results relative to oral and nasal vowels exploring contours extraction from 2D images, articulatory measures and area functions (Martins et al., 2008).

In former studies, vocal tract modelling has been limited to the midsagittal plane (Engwall, 2000a; Takemoto et al., 2003), but improvement of MR imaging equipment system capabilities allowed the expansion into this domain of research and made it possible to obtain three-dimensional (3D) modelling (Badin & Serrurier, 2006). The more realistic models of the vocal tract shape that nowadays are possible to obtain, are hugely needed in the research towards improved speech synthesis algorithms and more efficient speech rehabilitation.

The main purpose of this paper is to present some 3D models of the vocal tract based on MR data of some relevant sustained articulations of EP in a static study. From the point of view of image processing, a new approach for 3D modelling by means of the combination of orthogonal stacks, to describe the vocal tract shape in different articulatory positions is presented. We also demonstrate an MR technique to capture useful image sequences during speech (dynamic study). In addition, some preliminary results of this dynamic study are presented.

The remaining of this paper is organized in four sections. The next section is dedicated to the methods and describes the equipment, corpus and subjects, as well as the procedures used for the speech study, namely for morphologic and dynamic imaging of the vocal tract. The results are presented in following section, through the exhibition of some three-dimensional models built of the vocal tract and an image sequence obtained during speech. Finally the conclusions of the work described are presented.

2 METHODS

This study was performed in two phases: 1) exploration of MR techniques applied to the vocal tract imaging; 2) the use of image processing techniques that can aid the analysis of vocal tract.

2.1 Equipment, corpus and subjects

An MR Siemens Magneton Symphony 1.5T system and a head array coil were used, with subjects in supine position. The image data were acquired from two subjects (one male and one female) for the static study, and from four subjects for the dynamic study, all without any speech disorder.

The corpus of the static study consisted in twenty five sounds of European Portuguese: oral and nasal vowels, and consonants. For the dynamic study, the subjects produced several repetitions of sequences of three consonant-vowel syllables (/tu/, /ma/, /pa/) during the acquisition.

Because of the MR acoustic noise produced during image acquisitions, the acoustic recording of the produced speech was not yet possible.

2.2 Techniques

According the safety procedures for MR, subjects was previously informed about the exam and instructed about the procedures during the acquisitions. A consent informed was obtained from each subject involved.

Furthermore, the training of the subjects was performed to ensure the proper production of the intended sounds for the static study, and to achieve good speech-acquisition synchronization for the dynamic one.

2.2.1 Static Study

A set of MR WT1-images using Turbo Spin-Echo (TSE) sequences was acquired in sagittal and coronal orientations. The subjects sustained the articulation during 9 seconds for the acquisition of three sagittal slices and 9.9 seconds for the four coronal slices. The acquisition time was a compromise between image resolution and the duration of the sustained articulation allowed by the subject.

Initially, a single midsagittal T1-weighted image was acquired with subjects instructed to rest with mouth close and the tongue in full contact with the teeth. This reference image was used for teeth space identification and contour extraction (Figure 1).

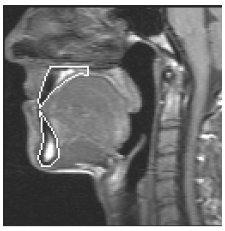

Figure 1: Midsagittal reference image for teeth identification and contours extraction.

The used protocol set includes the parameters shown in Table 1.

Table 1: MR protocol for vocal tract imaging.

Sagittal slices	Coronal slices
TR = 443 ms	TR = 470 ms
TE = 17 ms	TE = 15 ms
ETL = 7	ETL = 7
SNR = 1	SNR = 1.03
3 mm thickness	6 mm thickness and 10 mm gap
Three slices	Four slices
Matrix size 128 x 128	
Resolution = 0.853 px/mm	
Field of View (FOV) = 150 mm	

2.2.2 Dynamic Study

The dynamic study was performed following the same principle of MR cardiac analysis, with the modification of a FLASH-2D sequence using the patient's heart beat as a trigger signal, a 300 mm field of view and the acquisition parameters: TR = 60 ms and TE = 4.4 ms. The subjects tried to synchronize the utterance of the syllables to their own cardiac rhythm by means of the acoustic monitoring of their own simple electrocardiogram (ECG) through a synchronous sound emission conveyed to the subject by an earphone.

Each set of images from a single-slice (midsagittal) of 6 mm thickness was collected during 12 to 22 seconds. For each sequence, a variable number of images (4-6 images) were acquired with a regularly increasing shift in synchrony from the start of the cardiac cycle (Figure 2).

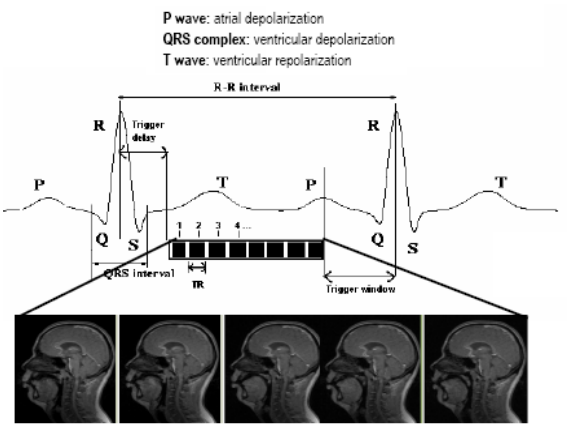

Figure 2: Diagram of the dynamic MR acquisition based on ECG monitoring and synchronization.

The RR interval is the time duration between two consecutive R waves (in ECG graph), and it is usually the reference interval for programming the slice acquisitions. It should be noted that depicted the images were acquired in a single cardiac cycle for the sake of representation, this is an under-sampling method, assuming that the phenomenon is stationary, which is quite good in cardiac analysis but no so in repeated speech production. In fact, the images were collected distributed along the time with a period as short as possible, but always longer than a cardiac cycle due to machine limitations.

2.2.3 Image Processing and Analysis

Few segmentation methods have been described in speech studies for vocal tract contours extraction from MR images. Briefly, those methods are based on manual edition of curves, such as Bézier curves, and threshold binarizations (Badin et al., 2000; Engwall, 2004; Soquet et al., 2002). Soquet et al. (1998) compared different approaches on the same data in order to assess the accuracy of some manual segmentation methods, and concluded that the methods considered give comparable results and that the threshold method is the one that presents lower dispersion.

Here, image analysis and 3D model reconstruction were accomplished in two stages:

- Image segmentation using the S*egmenting Assistant*, a 3D editing plug-in of *Image J* the image processing software developed by the National Institute of Health and subsequent 3D reconstruction;
- Graphic representation and combination of orthogonal stacks using the *Blender* software for 3D graphics creation.

The histogram-derived *threshold* technique was chosen for the segmentation of the airway from the surrounding tissues. The extraction of the contours of the vocal tract was then obtained by the following sequence of procedures:

(a) Identification and closure of the vocal tract area of interest, mandatory closure of the mouth, larynx, vertebral column and velum, through the manual superimposition of opaque objects;

(b) Manual overlapping of teeth image (done only on the sagittal stacks), after extraction of the teeth contours from the initially acquired sagittal anatomic reference image;

(c) Extraction of the contours of the vocal tract, for each image of 2D slices using the *Image J* semi-automatic *threshold* technique.

The Figure 3 depicts the image segmentation procedure used in sagittal and coronal slices during a sustained nasal vowel. The closure of the vocal tract area was necessary to avoid contours to "escape" and complicate the segmentation task.

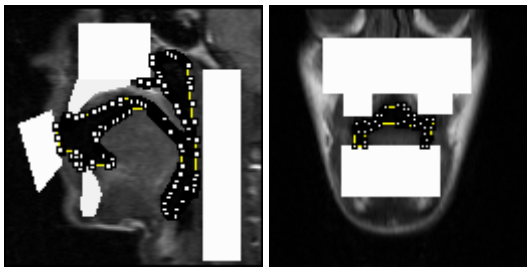

Figure 3: Contours extraction from sagittal (right) and coronal (left) slices after closure of the vocal tract area and manual overlapping of teeth.

The contour extraction process resulted in a total of 175 planar contours (maximally 7 contours for each sound).

Outlines were subsequently used to generate a 3D surface, 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 was required that the outlines be well aligned – this process is usually known as image registration. In Computational Vision, the term image registration means the process of transforming the different sets of images into one common coordinate system, what was here necessary in order to be able to compare or integrate the data obtained from different measurements.

3 RESULTS AND DISCUSSION

The static study was designed to obtain the morphologic data of most of the range of the articulators' positions aiming the imaging characterization of Portuguese sounds. A variable number of images were obtained by dynamic MR, according to the cardiac cycle of each subject, followed by the assembly of all images for sequence visualization.

3.1.1 3D models

The following images (Figure 4) represent different perspectives of the 3D model obtained for the vowel [u]. In the presented images, the blue surface represents the union of the three outlines extracted from the sagittal stack. By other hand, the red surface represents the union of the four outlines extracted from the coronal stack.

The Figure 5 depicts two 3D models of the vowels: corresponding to an oral sound (above) and to a nasal sound (below). The different viewpoints presented allow the identification of the velum lowering, and especially the partial closure of the oral cavity, comparatively to the oral sound. In Portuguese there is a special interest in nasal sounds due to their frequent use in common speech.

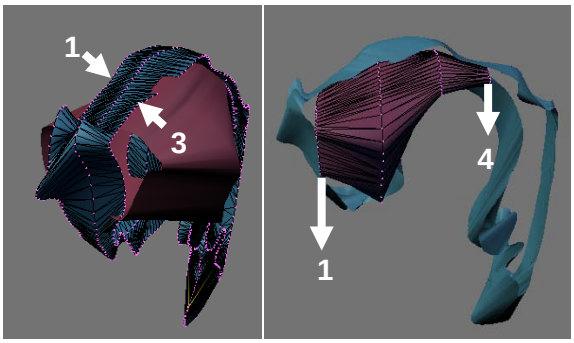

Figure 4: Surfaces representations for the 3D model of the vowel [u].

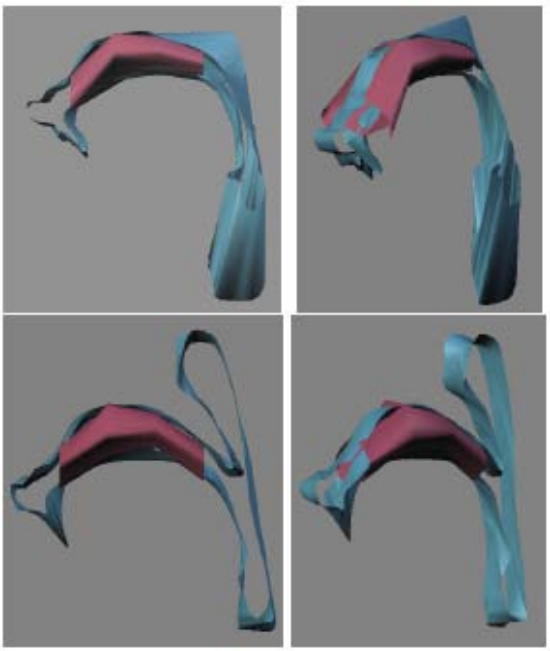

Figure 5: Three-dimensional models of the oral (top) and nasal (bottom) vowel [a] of European Portuguese.

In the 3D models obtained, some differences in the vertical lengths between sagittal and coronal stacks in some sounds were observed resulting in some registration errors. This could reflect the specific variability of the speaker in sound production, due in this case, to the fact that acquisitions of different orientations were separately done (first sagittal images, and subsequently coronal images), to minimize the subjects' effort needed for an extra long utterance. On the other hand, the segmentation process used had also some

implications for the determination of the vocal tract area contour.

Furthermore, the coronal data is important for 3D modelling of the vocal tract, because some articulatory situations lead to occlusions in the midsagittal plane, while lateral channels are maintained open (e.g. lateral consonants, nasals vowels). The models shown in Figure 6 intend to demonstrate the relevance of coronal stacks for the characterization of lateral consonants of Portuguese.

Although the 3D models built are not yet completely closed, it can be observed several essential features needed for the articulatory description of speech.

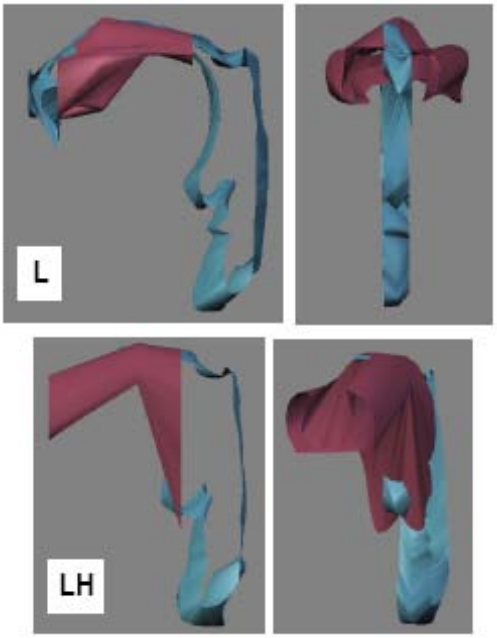

Figure 6: Three-dimensional models of the laterals consonants [l] and [lh] of European Portuguese.

3.1.2 Image sequence during speech

A variable number of images (sagittal slices) were obtained in dynamic studies, according to the cardiac cycle (heart frequency) of each subject, followed by the assembly of all images for image sequences visualization. The best image features were obtained for the syllable /tu/, as illustrate in Figure 7, because the articulatory positions are very different between single sounds, compared with the other syllables.

When considering the differences verified amongst the subjects, by image comparison, the dynamic studies demonstrated the actual variability in sounds production between subjects, not only due to anatomic differences, but also because each subject uses different strategies for motion control and articulation.

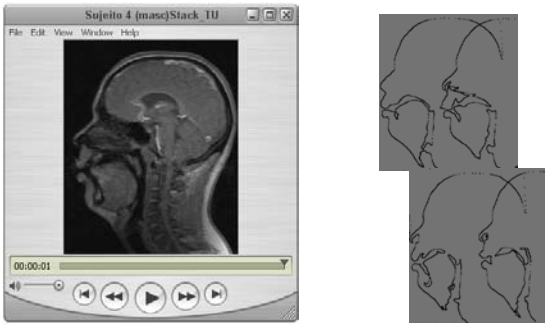

Figure 7: Contours extracted from midsagittal images obtained in the dynamic study by the repetition of the sequence /tu/.

4 CONCLUSIONS AND FUTURE WORK

In our study, a considerable number and diversity of images were acquired aiming at not only morphological but also a dynamic characterization, by exploring various MR techniques. We tried to acquire a number of images with enough anatomic resolution, maximum vocal tract extension of representative speech gestures, minimizing speaker effort (reducing hyperarticulation). The image data was analysed and processed resulting in the reconstruction of 3D models for the entire corpus (3D geometrical database).

For almost all 3D models obtained for Portuguese sounds the morphologic data showed that both orientations slices (sagittal and coronal) are useful for the knowledge of the vocal tract shape during speech production. Articulators' positions are better demonstrated in sagittal images, and the coronal images allow the observation of the lateral dimension of oral cavity.

The completion of the construction of the surfaces for the hybrid models made from sagittal and coronal stacks is the next step in the way to obtain a complete 3D anatomical model of the vocal tract, prepared for the subsequent prediction of the acoustic output.

The extension of the dynamic sequences obtained to other sequences is also important in terms of coverage of the study, and will be done in the near future.

Other problems related with the image registration and with acoustic recording of speech are being investigated until now, aiming to be solving in a future work.

ACKNOWLEDGEMENTS

The images considered were acquired at the Radiology Department of Hospital S. João, Porto, with the collaboration of Isabel Ramos (Professor from Faculdade de Medicina da Universidade do Porto and Department Director) and the technical staff, which are gratefully acknowledged.

REFERENCES

Avila-García, M.S., Carter, J.N., Damper, R.I., 2004. Extracting Tongue Shape Dynamics from Magnetic Resonance Image Sequences. *Transactions on Engineering, Computing and Technology V2*, *December*, 288-291.

Badin P., Bailly G., Raybaudi M., Segebarth C., 1998. A three-dimensional linear articulatory model based on MRI data. *3rd ESCA / COCOSDA Int. Workshop on Speech Synthesis*, Australia, 249-254.

Badin, P., Borel, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., 2000.Towards an audiovisual virtual talking head: 3D articulatory modeling of tongue, lips and face based on MRI and video images. *5th Speech Production Seminar*, Germany, 261-264.

Badin, P., Serrurier, A., 2006. Three-dimensional Modeling of Speech Organs: Articulatory Data and Models. *IEICE Technical Committee on Speech*, Japan, 29-34.

Baer, T., Gore, J.C., Gracco, L.C., Nye, P.W., 1991. Analysis of Vocal Tract Shape and Dimensions using Magnetic Resonance Imaging: Vowels. *J. Acoust. Soc. Am*., 90, 799-828.

Behrends J., Wismuller A., 2001. A Segmentation and Analysis Method for MRI data of the Human Vocal Tract, *FIPKM-37*, 179-189.

Demolin D., Metens T., Soquet A., 1996. Three-dimensional Measurement of the Vocal Tract by MRI. *4th Int. Conf. on Spoken Language Processing (ICSLP 9*6), USA, 272-275.

Demolin, D., Metens, T., Soquet, A., 2000. Real time MRI and articulatory coordinations in vowels. *5 th Speech Production Seminar*. Germany.

Engwall, O. , 2003. A revisit to the Application of MRI to the Analysis of Speech Production - Testing our assumptions. *6th Int. Seminar on Speech Production*. Sydney.

Engwall, O., 2000a. A 3D Tongue Model based on MRI data. *6th Int. Conf. on Spoken Language Processing* (ICSLP), China, 901-904.

Engwall, O., 2000b. Are static MRI representative of dynamic speech? Results from a comparative study using MRI, EPG and EMA. *6th Int. Conf. on Spoken Language Processing* (*ICSLP*), China, 17-20.

Engwall, O., 2004. From real-time MRI to 3D tongue movements. *ICSLP 2004*, vol. II, October, Korea, 1109-1112.

Kitamura T., Takemoto H., Honda K., Shimada Y., Fujimoto I., Syakudo Y., Masaki S., Kuroda K., Oku-uchi N., Senda M., 2005. Difference in vocal tract shape between upright and supine postures: Observations by an open-type MRI scanner. *Acoustical Science and Technology*, 26(5), 465-468.

Mády, K., Sader, R., Zimmermann, A., Hoole, P., Beer, A., Zeilhofe, H., Hannig, C., 2001. Use of real-time MRI in assessment of consonant articulation before and after tongue surgery and tongue reconstruction. *4th Int. Speech Motor Conf.* Netherlands, 142-145.

Martins, P., Carbone, I.C., Pinto, A., Silva, A., Teixeira, A.J., 2008. European Portuguese MRI based speech production studies. *Speech Communication*, 50, 925-952.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., Byrd, D., 2004. An Approach to Real-time Magnetic Resonance Imaging for Speech Production. *Journal Acoustical Society of America*, 115(4), 1771-1776.

Rua, S.M., Freitas, D.R., 2006. Morphological Dynamic Imaging of Human Vocal Tract. *Computational Modelling of Objects Represented in Images: Fundamentals, Methods and Applications (CompIMAGE)*, Portugal, October 20-21, 381-386.

Serrurier, A. & Badin, P., 2005. A Three-dimensional Linear Articulatory Model of Velum based on MRI data. *Interspeech 2005: Eurospeech, 9th Europ. Conf. on Speech Communication and Technology*, Portugal, 2161-2164.

Soquet, A., Lecuit, V., Metens, T., Demolin, D., 2002. Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI. *Speech Communication*, 36, 169-180.

Soquet, A., Lecuit, V., Metens, T., Nazarian, B., Demolin, D., 1998. Segmentation of the Airway from the Surrounding Tissues on Magnetic Resonance Images: A comparative study. *ICSLP*. Sydney, 3083-3086.

Takemoto, H., Honda, K., 2003. Measurement of Temporal Changes in Vocal Tract Area Function during a continuous vowel sequence using a 3D Cine-MRI Technique. *6th Int. Seminar on Speech Production*, Australia, 284-289.

Teixeira, A. et al., 2002. SAPWindows – Towards a Versatile Modular Articulatory Synthesizer. *Proceedings of 2002 IEEE Workshop on Speech Synthesis*. Portugal, 31-34.

Teixeira, A., Moutinho, L.C., Coimbra, R.L., 2003. Production, Acoustic and Perceptual Studies on European Portuguese Vowels Height. *15th Int. Congress of Phonetic Sciences*, Barcelona, 3033-3036.

Teixeira, A., Vaz, F., 2001. European Portuguese Nasal Vowels: An EMMA Study. *Eurospeech 2001.* Aveiro, Portugal, 1483-1486.