Application of MRI and Biomedical Engineering in Speech Production Study

S. R. Ventura[a], D. R. Freitas[b], João Manuel R. S. Tavares[c 1]

[a]*Escola Superior de Tecnologia da Saúde do Porto - IPP*

Rua Valente Perfeito, n.º322

4400-330 VILA NOVA DE GAIA

PORTUGAL

[b]*Departamento de Engenharia Electrotécnica e de Computadores*

[c]*Departamento de Engenharia Mecânica e Gestão Industrial*

[a, b, c]*Faculdade de Engenharia da Universidade do Porto*

Rua Dr. Roberto Frias, s/n

4200-465 PORTO

PORTUGAL

Telephone: +351 22 508 1487 Fax: +351 22 508 1445

Email: sandra.rua@eu.ipp.pt, {dfreitas, tavares}@fe.up.pt

[1] Corresponding author. Email: tavares@fe.up.pt

Application of MRI and Biomedical Engineering in Speech Production Study

Abstract

The speech production has always been a subject of interest both at morphologic and acoustic levels. This knowledge is useful for a better understanding of all the involved mechanisms, and for the construction of articulatory models. Magnetic resonance imaging (MRI) is a powerful technique that allows the study of the whole vocal tract, with good soft tissues contrast and resolution, and permits the calculation of area functions toward a better the understanding of this mechanism. Thus, our aim is to demonstrate the value and application of MRI in speech production study and its relationship with engineering, namely with biomedical engineering. After vocal tract contours extraction, data was processed for three-dimensional reconstruction culminating in models construction of some sounds of the European Portuguese. MRI provides useful morphologic data about the position and shape of the different speech articulators, and the biomedical engineering computational tools for its analysis.

Keywords: Biomedical Engineering, Magnetic Resonance, European Portuguese Sounds, Vocal Tract, Speech Production, Three-dimensional Models.

1. Introduction

Verbal communication is one of the most common, familiar and frequently used forms of human interaction, which result of an organized and synchronized work of a set of anatomic organs. The articulation is a result of the activity of a set of organs – the vocal tract, that modify their position and shape during air expulsion (expiration), producing different sounds and so different acoustic representations.

The speech production understanding has been, therefore, widely studied by means of several techniques, but the information is still scarce, namely for the European Portuguese (EP) Language (Teixeira et al., 2001a).

This issue has a significant and multidisciplinary interest, which involves Medicine (anatomic and functional study of the vocal tract organs), Engineering (in particularly the Biomedical Engineering for acoustic analysis and speech processing), Phonetics (study of the production and perception of speech and sounds), Speech Therapy (assessment of anatomic and physiologic aspects related with communication disorders, language and speech) and Medical Imaging (improvement and application of image techniques for the vocal tract study).

Our aim is to demonstrate the value and application of MRI in speech production study of EP language, and its relationship with Engineering, namely with Biomedical Engineering. From the point of view of image processing, a new approach for 3D modeling through the combination of orthogonal stacks, to describe the vocal tract shape in different articulatory positions is presented. This knowledge contributes to improved speech synthesis algorithms and more efficient speech rehabilitation.

This paper is organized as follows: in the next section we will present MRI protocol and image analysis and reconstruction techniques; then we will present the obtained three-dimensional models for EP characterization; finally, some conclusions and suggestions for future work are made.

1.1 Vocal tract anatomic aspects

The vocal tract consists of a set of cavities where sounds are produced, similar to a tube, with a non-linear shape and a considerable length that extends from the vocal folds to the lips, with a side branch leading to the nasal cavity. Vocal tract organs (or articulators) include the tongue, lips, teeth, alveolar ridge, hard palate, velum (soft palate), and the pharynx, as shown in Figure 1.

The speech production mechanism depends on the respiratory system, where the lungs are the functioning engine of a set of cavities (larynx, pharynx, oral and nasal), producing the necessary air flow. The bronchi and

trachea make the connection between lungs and the larynx cavity. The human vocal tract acts like an acoustic filter, by changing the characteristics of the sound during air passage and producing the speech.

The variability is a typical feature of the speech mainly because of the dynamic character of this mechanism (Faria et al., 1996). Furthermore it can occur an overlapping of adjacent articulations by the non-synchronization of the movements of organs, or two articulators can move at the same time for different phonemes. This phenomenon is known by coarticulation.

According to the source-filter theory of speech proposed by Fant (1960), the vocal tract is considered an acoustic filter of the sounds produced in the larynx, where the vocal folds are the sound source, and the vocal tract shape (resonances cavities) is the filter. The vocal tract shape determines how the sound is modified; different phonemes can be distinguished by the properties of their source and their spectral shape according these tube variations. This model is widely used in a number of applications, e.g. speech synthesis and speech analysis, because of its relative simplicity.

1.2 European Portuguese Language

The standard European Portuguese system consists of nine oral and five nasal vowels, three diphthongs and nineteen consonants (six plosives, six fricatives, three nasals and four liquids). The Portuguese Language has one of the richest vowel phonologies of all Romance languages, and like Catalan, uses vowel height to contrast stressed syllables with unstressed syllables; the vowels /a ɛ e ɔ o/ tend to be raised to /ɐ e i ɨ o u/ (although /ɨ / occurs only in EP) when they are unstressed.

Vowels are long and somewhat continuous sounds that are similar but hardly ever match perfectly across languages; consonants usually either match closely enough to permit easy transference or, in some cases, do not match at all. Consonants are classified according to the places at which the articulators come close together and obstruct the vocal tract - articulation points - being moderately easy to identify the distinctive features of the sounds produced.

Consonants have almost the same “value” as in other Romance languages, with some variations from region to region. One example of the most important variations is the phoneme /rr/ that is generally alveolar in Portugal and frequently uvular in France. Another example is be the phoneme /lh/ that corresponds to the Spanish /ll/ (as pronounced in Latin America) and to the Italian /gl/, another one is the phoneme /nh/ that corresponds to the Spanish /ñ/ and to the Italian and French /gn/. Both phonemes /ch/ and /j/ are pronounced

as in French. The dental character of the consonants /d t n l/ is more pronounced in Portuguese than in English, because in Portuguese pronunciation the tongue tends to touch the base of the upper teeth.

Nasality is a complex feature, given rise to a large number of perception and production studies. This gesture is defined by the lowering of the velum to open the velopharyngeal port, which induces strong and complex changes in the vocal tract acoustical behaviour. We can say that the EP Language is rich in nasal sounds – both vowels and consonants.

1.3 Measuring and analysis of the speech production

A wide variety of tools has been devised to measure and analyze speech production, thanks to technological advances based on the assessment of muscular activity, on movement and shape articulators studies and in the acoustical analysis of the speech.

Cineradiography was first introduced by Russel (1928) to examine movements of the vocal tract, and since then, a considerable amount of research on different language has been conducted. This image technique allows the observation, using frame-by-frame analysis, of the trajectories of metal markers attached to lips, jaw or tongue for speech production study, or with the subject swallowing a contrast media substance for dysphagia study. X-ray data is been proved very useful to study the dynamic characteristics of speech movements especially concerning tongue's shape and position during vowels and the pharynx. Nevertheless, is has been very difficult to get accurate measurements of the cross-sectional morphology from midsagittal profiles, and the need of long time exposure to ionizing radiation has been considered as the main limitation.

In order to reduce some x-ray harmful effects and frame-by-frame analysis the x-ray microbeam technique was introduced. This technique uses narrow x-ray beams to track gold pellets attached to the tongue, jaw and lips. However, this speech-specific technique does not enable the tongue root and pharynx imaging.

Other image technique, widely used in speech research, is ultrasound. It uses sound waves to collect real time data, showing tongue surface motion during speech. Without health hazards, ultrasound technique allows tongue contours extraction with enough accuracy. However, it shows some drawbacks when facing the lack of absolute spatial references in the signal and it is unable to acquire the tongue tip. On the other hand, ultrasound is relatively inexpensive (comparing to other techniques), can be portable, and is more comfortable to subjects.

Electromagnetic articulography (EMA) is by far the most used system for tracking movements within the vocal tract among speech researchers. The movements during speech production are tracked by measuring

induced current from receiver sensors moving in a magnetic field. The most important advantages of EMA include the fast tracking rate and the ability to study multiple articulators simultaneously.

Both ultrasound and EMA systems can demonstrate tongue shape or positions, but are not able to measure precisely the tongue/palate contact. Through electropalatography, a custom “pseudo-palate” is created for the subject, with tiny pressure sensors. This instrumental technique collects contact data from the palate and sends it into a computer for contact patterns displaying, in real time or off-line, for analysis. Electropalatography limitations include: 1) the extremely limited availability, 2) it is unsuitable for younger childrens, 3) initial set-up cost, and 4) no large scale studies to prove the efficacy available.

Computed tomography scanners allow obtaining a comprehensive amount of data on the vocal tract dimensions, with good temporal resolution, but due to the hazards associated with the use of high x-rays dosage, this technique tends to be less and less used.

To obtain more detailed information between vocal tract shape and speech sounds, three-dimensional data is required. For this purpose, only magnetic resonance imaging (MRI) technique provides excellent structural differentiation of the whole vocal tract organs and without harmful effects. Ridouane (2006) presented a comparison of some techniques to investigate speech production and evaluate weaknesses and strengths of each one.

MRI has been used in many different studies, with the purpose of obtaining morphological measurements in static situations during the production of vowels (Badin et al., 1998; Demolin et al., 2000; Engwall, 2004; Serrurier & Badin, 2005) and some consonant (Badin et al., 2006; Mády et al., 2001), and more rarely, because of device limitations, in dynamic situations (Avila-Garcia et al., 2004; Narayanan et al., 2004; Shadle et al., 1999). The use of MRI has been expanded from two-dimensional to three-dimensional imaging, and is becoming one of the most powerful tools to understand speech production.

Regarding the sounds previously subject of interest, found in our literature revision, vowels have been the most frequently focus of study in different languages such as French, Swedish, Germany, English and Japanese, mainly, because these sounds are somewhat easier to produce and to sustain, and the segmentation process is less complex (air flows freely through the whole vocal tract). Consonants produced in symmetric vowels contexts (Badin et al., 2000; Engwall, 2000) were more rarely studied. To understand the characteristics of EP, only nasal vowels studies have been carried through, regarding the acoustic production and perceptual levels, based on acoustic analysis and electromagnetic articulography (Teixeira et al., 2001b; Teixeira et al., 2002; Teixeira et al., 2003). More recently, another MRI study of EP presents some results relative to oral and nasal vowels and consonants exploring contours extraction from 2D images, articulatory

measures and area functions (Martins et al., 2008). In this study both 2D and 3D MRI data were collected from one speaker in order to get a quantitative analysis of articulatory points by means of image analysis processing; mainly, the 2D and 3D segmentation of the vocal tract and area extraction of the sections were performed through a resliced midsagittal from volumetric data. However, the 3D vocal tract modeling was not performed and quantitative measures were obtained through one sagittal image only.

Our study is the first report of MRI application for EP production study by means of the development of a new protocol-imaging, for the collection of more 2D images in different orientations and achieved in a reasonable time. Furthermore, we present the first 3D models of the vocal tract – a morphological database, in different articulatory positions for a considerable number of EP sounds.

1.3.1 MRI value and drawbacks in speech research

MRI is a reference technique to obtain multiplanar high quality imaging of soft tissues, with enough subject and personnel safety, allowing the simultaneous study of the whole vocal tract extension. In addition, the airway area and volume can be directly calculated. Even so, the currently available information to understand speech production can be considered as insufficient or even non-existing, particularly for European Portuguese sounds.

It is well known that unfortunately MRI devices produce quite loud and disturbing noise during normal operation. Although acceptable for the regular clinical practice causing only subject's annoyance, this situation is radically limiting for speech studies. The simultaneous recording of subject's speech and the imaging of slices of the vocal tract is therefore extremely difficult and useless for analysis. One tentative can be done through the usually existing intercommunicator but even with enthusiastic signal processing the resulting speech is, up to now, generally non-usable. It must be noted that no pick-up devices are allowed, in general, near the magnet/coil system of the equipment.

Another attempt for making the recording of the speech signal can be done using acoustic transmission through an acoustic wave guide (acoustic pipe) in order to reach a pick-up device placed far enough from the forbidden volume, but no better results are obtained.

In Figure 2, the depicted waveform was recorded via the intercommunicator at the end of an acquisition of an MRI sequence. It can be seen that the speech waveform on the right side is about -12dB in amplitude relative to the noise part on the left. A small AGC effect is observed, showing that the SNR is even worse.

More recently, new possibilities were open for speech studies by NessAiver et al. (2006) and Bresch et al. (2006). According these studies it is possible to obtain high quality speech recording using an optical microphone during the MR acquisition, allowing spectral analysis of speech signal.

Other drawbacks of MRI in speech production research related with the subject, previously described by Engwall (2003) is the need to sustain the articulation artificially for long times (30 sec. or more) due to very long acquisition times, and the positioning imposed (corporal position lying on their back or supine position) due to device constraints. This study presented an evaluation of the effects of these two situations, and suggests that the artificially prolonged articulations seemed to be hyperarticulated and more difficult to hold by the subject (backward movement of the tongue and lack of velum control), and the corporal posture does affects position and shape tongue, especially when articulation has to been sustained. Consequently, the acquisition time should be kept as short as possible.

Also similar results were obtained by Kitamura et al. (2005); they studied the gravitational effect in articulations, comparing the shapes of the vocal tract and the articulators between two subjects' positions: supine and upright (by open-type MRI equipment).

1.3.2 MRI acquisition particularities

MRI is one of the most interesting fields of medical imaging with its ability to produce images almost equivalent to a slice of the body parts anatomy, and making possible an infinite number of image planes through the body. Image acquisition is obtained from the detected radio frequency signals emitted by the nuclei of hydrogen atoms, after proper excitation done by a set of coils placed near the part under observation in presence of the high-valued magnetic field produced by the large superconducting magnet that surrounds the patient.

The equipment submits the nuclei of hydrogen atoms present in a certain volume to a radio frequency signal that temporarily "knocks" them out of their magnetic alignment. When the excitation signal stops, these nuclei tend to return to their previously aligned position, in a precession movement radiating their own radiofrequency contributions which are picked-up by the receiving coil working as a radio antenna. These signals are then detected and computer-processed to produce spatial nuclei concentration information that when plotted results in very detailed image-like data of the anatomic structure under observation.

The sets of images obtained by MRI are usually called stacks, because they are designed to be parallel and regularly spaced in order to describe the object and represent the spatial sampling of an anatomic volume along a certain axis, or, posed in another way, by slices oriented in a desired direction.

Unfortunately, some structures such as the bones and teeth have small concentrations of mobile hydrogen, and therefore produce very subtle signals that make them almost invisible in MRI. Because of this, for our purpose, the non-identification of the teeth leaves a serious problem to solve, because similar pixel intensity values appear in the oral cavity and on the teeth making the extraction of contours largely overestimated in the directions of teeth if no measures are taken.

Various methods have been used to extract information about the teeth volume; as for example, by Yang et al. (cited by Takemoto et al., 2004) who used scanned casts in water to obtain by contrast the three-dimensional shape of the teeth. A few different approaches have also been attempted, using other imaging techniques like, for example, dental crow plates, requiring time and effort for a manual segmentation of the images. Takemoto et al. (2004) described a new method to superimpose three-dimensional dental images on MRI volume data, in which subjects were asked to hold blueberry juice inside the mouth during image acquisition in prone position. The vocal tract shape was extracted after overlapping of the dental and orofacial images (in supine positions during sustained vowels production), using defined anatomical landmarks.

The accurate measurement of the vocal tract shape during phonation is otherwise the most important part of speech production studies aiming speech articulatory modeling. For a very long time, the midsagittal plane was used to measure the vocal tract (and to obtain area functions from these data). This approach, according to Badin et al. (2000), lead to a number of problems: 1) the need for a model converting midsagittal contours to area functions, 2) the difficulty of modeling lateral consonants and 3) the limitation of acoustical simulations to the plane wave mode only.

Therefore, three-dimensional data is obviously needed in order to get a more realistic representation that may lead to better models. Demolin et al. (1996) proposed a new MRI technique to make measurements of sagittal, coronal, coronal oblique and transversal planes; the extracted areas were placed on a flux line of the vocal tract from a mid-sagittal slice. These data offer, according to the authors, promising perspectives for the making of true 3D measures of the vocal tract. Currently, with the increasing availability of MRI devices and image processing means it is more and more feasible to obtain these 3D vocal tract data with a reasonable acquisition speed.

However, no specific-MRI protocol for the vocal tract study was found in the current literature. Only some general acquisition considerations have been reported:

- Acquisition time: must be a compromise between image resolution and time for sustained articulation allowed by the subjects;

- Articulation: can be repeated many times followed by posterior image reconstruction (however this introduces variability effects), or sustain articulation throughout the whole acquisition with or without breath hold (artifacts appear due to subject's fatigue). The time of sustained articulation must be reduced;
- Number and thickness of slices: is a compromise between the resolution needed and acquisition time. More slices mean a better spatial resolution but a longer acquisition time. These parameters depend on the covered area needed for the study.
- Orientation of slices: contours extraction is more accurate when the slice is approximately perpendicular to the region surface that is intended to be detected;
- Slice planes: planned according to the volume under study. Sagittal plane provides information of whole vocal tract extension. Coronal plane provides orthogonal information in the lateral dimension and allows lips and tongue study. Transversal plane is useful in pharynx study, but the segmentation process is many times much more difficult;
- Field of view (FOV): must be adjusted according to the area under study, range from 150 to 200mm. However, for dynamics studies the FOV must be increased to 300mm due to the higher amount of artifacts that appear (due to motion and aliasing);
- MRI sequences: fast techniques are required; this is only possible with high magnetic field equipment (up to 1.5T or even 3.0T if available);
- Corpus: is a compromise between researcher's and subject's contentment.

1.3.3 Subjects training and MRI acquisition begin

The image collection is defined in accordance with acquisition-specific parameters (briefly explained above), and is planned over a first survey (reference) image taken with similar parameters.

Because of all the particularities of MRI, subjects must be previously informed about the imaging exam, and instructed in order to avoid anxiety and to have their total collaboration. For vocal tract imaging, the subjects are asked to produce different speech sounds, during one determined and required time. Furthermore, the entire corpus must be explained and they must be prepared to sustain a specific sound at a specific time. Subjects' training is therefore need a for good speech-acquisition synchronization, and to ensure the production of the intended sound.

1.4 Speech and image processing

Speech processing involves the study of speech signals and the processing methods. The first speaking machine was created in 1771 by Wolfgang von Kempelen, in the attempt to create speech. This machine consists of a bellows and a sound tube that simulate vocal tract, along with the reed of a musical instrument installed at one end of the sound tube.

A different approach was proposed in 1930 by Homer Dudley – the Vocoder (Voice Operated reCOrDER). This machine recreates the sound resonance characteristics of the sound waves using the shape of the vocal tract and electrical circuits as resonance filters.

Speech synthesis systems are today widely used in order to simulate the process of human speech production (artificially), by means of the construction of realistic models. The vocal tract model is created as a non-uniform sound tube, from the image data provided by MRI or x-rays techniques, with differing cross-sectional areas, and the transmission of sounds waves inside the sound tube is expressed by a digital filter.

Speech synthesis is an assistive technology tool useful to people with a wide range of disorders, and although the data processing tool and measurement methods have made great improvements, but some limitations to explain all mechanisms involved in the speech production process still to be solving.

Image processing usually refers to digital image processing and consists in treating the image as 2D signal and applying standard signal-processing techniques, such convolution, Fourier analysis and statistical descriptions, and manipulative tools. Image segmentation is one of the most important and difficult tasks of digital image processing and analysis systems, particularly for vocal tract segmentation. This process is used to find the objects of interest separating them from all the rest. The most common technique is thresholding, because of its semi-automatism, flexibility in contours adjustment, and the lesser dependency by the user when compared to manual methods.

Due to the high susceptibility of MRI technique, the images contain some background noise, addressed to the random variation in signal intensity which degrades image information. Many factors related with the system measurement (coils, electronics) may also contribute for these signal interferences, but the main source of noise in the image is the patient's body. This issue makes segmentation task more difficult, namely for speech production study, because air-tissue boundaries of the vocal tract are hard to extract and the segmentation is less accurate especially when using automatic methods.

2. Methods

2.1 The equipment, the speech corpus and the subjects

Experiments for image acquisition were performed on a Siemens Magneton Symphony 1.5T system, with subjects lying in supine position. A head array coil was used. The speech corpus consisted of a set of images collected during sustained articulations of eight EP phonemes (five oral vowels and three nasals consonants).
This study was carried out with two young subjects (one male and one female), all volunteers, healthy without previous oral or facial pathology, and skilled in Speech Therapy. This subjects' knowledge was essential for obtaining the best possible results, because the acoustic recording of the produced speech just was not possible during the MRI acquisition. In addition, the required subject's training time was substantially reduced. Previous to this study a proper informed consent was obtained for each subject, and also the safety procedures for MRI examinations.

2.2 Static study

In order to determine the acquisition time of MRI sequences, the average sustained time of some sounds for each subject was measured and voice was recorded, in order to assess their respiratory capability. This procedure was performed with the intention to design a protocol for the vocal tract study, and to decrease subjects' effort in long sustaining phonation, with a most probable degree of hyperarticulation.
Because of the importance of an accurate segmentation of the sagittal image data, the teeth shape and placement identification was obtained by means of a midsagittal reference image (due to its higher resolution), collected with the subjects in "rest" position: lips closed and full contact of the tongue with the teeth. Because soft tissues are clearly demonstrated by MRI, this procedure was realized on the assumption that the low signal (dark area) observed between lips and tongue corresponds to the teeth (Figure 3).
To ensure the right speech production, before each acquisition sequence, each subject was informed about the sound intended to be sustained, and guided for speech initiation. This was possible using the available intercommunicator system, by the following procedure: subjects started articulation after a full breath inspiration and a numerical counting instruction at number two. The numerical counting was used to allow subjects to be prepared, and to initiate MR acquisition during speech production (that was done on the count of three) thereby reducing subjects' reaction times.
The first acquisitions, the conventional ones, consisted in complete sagittal stacks, from right-to-left sides of the vocal tract during sustained articulation. Thirteen slices with 5mm thickness were acquired during 33sec.

With this approach we verified that medial articulatory positions are more relevant and only half of the slices were useful for the segmentation process. Furthermore, as these results in a high subjects' effort, it was possible to specify the best balance for the number of slices to acquire, leading to a much smaller number that inevitably leaves out a lot of detail at the wider parts of the vocal tract.

Given the extension and non-linear shape of the vocal tract, two basic slice orientations were used, eventually at specific sections of the vocal tract: the sagittal (images of the whole vocal tract) and the coronal (images of oral cavity).

The acquisition was then done using Turbo Spin Echo Sequences, T1-weighted images, with a recording duration approximately 10s and a 150mm-sized FOV. Three contiguous sagittal slices with 5mm thickness and four coronal slices with 6mm thickness, 16mm spaced were collected.

2.3 Image analysis and 3D vocal tract models

Image analysis and 3D model reconstruction was accomplished in two stages, namely: 1) image segmentation, using the Image J image processing software (developed by the National Institute of Health) and a 3D editing plug-in, and 2) a 3D reconstruction of each stack, using the Blender software for 3D graphics creation, version 2.41.

The segmentation process remains complex and time consuming, because a few related problems still persist due to the MRI technique characteristics and to the anatomic aspects of the vocal tract:

- Airway contours delimitation problems were verified looking for eventual non-identification of some organs or anatomic parts or artificial inclusion of non-existing parts;
- Non-identification of teeth in MRI;
- The back-end of the vocal tract is not well defined;
- Air-tissue boundaries of the vocal tract are hard to extract by similarity of anatomic structures around it.

A few segmentation methods have been described, in speech studies, for vocal tract contours extraction from MR images, based on manual edition, such as Bézier curves, and threshold contours. Soquet et al. (1998) compared different approaches to the same body data in order to assess the accuracy of three segmentation methods: manual, threshold and elastic. This study shows that the different methods give comparable results with a somewhat lower dispersion for the threshold method, and the settings of the parameters of each method

(contrast, threshold level, free-form curve) have an impact on the resulting area. For this study, we selected the threshold method for the following additional reasons:

- Contours extraction is almost automatic, easy to adjust and flexible to the vocal tract structures;
- It takes less time consumption and effort comparing with others tools of manual edition.

The image segmentation of the airway from the surrounding tissues for extraction of the contours of the vocal tract was then performed by the following sequence procedures:

(a) Identification and closure of the area of interest of the vocal tract, with mandatory closure of the mouth, larynx, vertebral column and velum;

(b) Manual pasting of teeth image (only over sagittal stacks), after extraction of the teeth contours from the 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.

Figure 4 depicts the result of the manual identification and closure of vocal tract of interest with solid boxes, and the pasting of the contours of the teeth. The extraction of contours of the vocal tract was then performed by outlining the area of interest, slice by slice. Each outline is defined by the minimum and maximum levels in the area, moving and controlling histogram values. The boundary is given after adjusting levels parameters to match the outline with the area of interest.

3. Results and discussion

The images shown in Figure 5 represent different views of the 3D model obtained for the vowel /u/. The blue skin represents the union of the three outlines extracted from sagittal stack (Figure 5a). The red skin represents the union of the four outlines extracted from coronal stack (Figure 5b). It should be observed that these reconstructions are not yet closed as should be for a whole vocal tract reconstruction. This closing of the skin by unification of the different stacks is the next step in the processing.

The 3D models presented in Figure 6 for oral vowels among male and female subjects, demonstrate articulators positions and shapes. The main differences observed comparing vowels are addressed to the tongue and lips shapes. Anatomic differences among subjects are also well demonstrated, but the main positions of the articulators are similar. 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 centre and from centre to back for

the vowels [i, e, a] and [a, o, u] respectively. The lateral dimension of oral cavity demonstrated by the coronal data is larger for the production of the vowels [a, e, o] and for the male subject.

The models presented in Figure 7 intend to demonstrate nasal consonants production in different views, and show the velum lowering and especially the partial closure (for the sounds /m/ and /n/) of the vocal tract in the oral cavity. These 3D models are apparently similar, but the points of articulation are different. For the production of the consonant /nh/, we can observe the full tongue/palate contact, closing oral cavity, and forcing air to flow through nasal cavity; for this reason coronal data was not useful. For the construction of this 3D model only some cavities of the vocal tract are demonstrate in sagittal stack.

The 3D models presented demonstrate several essential features needed for the articulatory description of speech, specifically:

- shape of the tongue;
- position of the tongue in oral cavity, and relation the palate;
- mouth opening degree and shape;
- lateral dimension of oral cavity;
- length of the vocal tract;
- velum position (open or close).

5. Conclusions and future work

The use of MRI can provide very useful and precise morphological information about the position and shape of the different articulators involved in speech production and also of their dynamics, although, as shown in the present work, still has some speed restrictions due to the limitations of present day's available equipment and protocols.

The image material was carefully analysed and processed. A set of 3D reconstruction data of the vocal tract shape in a wide variety of EP sounds, obtained from careful and registered combination of orthogonal sagittal and coronal contour stacks was produced and presented. These data show that with the proposed approach a better coverage of important vocal tract regions was clearly achieved. The acquired and processed data can be used for several studies in articulatory phonetics, speech therapy, vocal tract modeling for speech synthesis, etc.

The next required step of this work is the completion of the construction of the 3D models (closing skins laterally by geometric processing of sagittal and coronal stacks), in order to get complete and real anatomical

data for speech synthesis. In the present phase, the basic data is already useful for many articulatory phonetic studies.

Acknowledgment

The images here considered were acquired at the Radiology Department of the 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, which are gratefully acknowledged.

References

Avila-García MS, Carter JN, Damper RI. 2004. Extracting Tongue Shape Dynamics from Magnetic Resonance Image Sequences. Transactions on Engineering, Computing and Technology Enformatika V2, December, 288-291.

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. In Proceedings of the 5th Speech Production Seminar; München; Germany, 261-264.

Badin P, Pouchoy L, Bailly G, Raybaudi M, Segebarth C, Lebas JF, Tiede MK, Vatikiotis-Bateson E, Tohkura Y. 1998. Un modèle articulatoire tridimensionnel du conduit vocal basé sur des données IRM. Actes des 22èmes Journées d'Etude sur la Parole; Martigny, 283-286.

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

Bresch E, Nielsen J, Nayak K, Narayanan S. 2006. Synchronized and noise-robust. audio recordings during realtime MRI scans. J Acoust Soc Am. October; 120(4): 1791–-1794.

Demolin D, Metens T, Soquet A. 1996. Three-dimensional Measurement of the Vocal Tract by MRI. Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP); Philadelphia; USA, 272-275.

Demolin D, Metens T, Soquet A. 2000. Real time MRI and articulatory coordinations in vowels. Proceedings of the 5 th Speech Production Seminar; München; Germany.

Engwall O. 2000. Are static MRI representative of dynamic speech? Results from a comparative study using MRI, EPG and EMA. Proceed. of the 6th International Conference on Spoken Language Processing (ICSLP), Beijing, China, 17-20.

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

Engwall O. 2004. From real-time MRI to 3D tongue movements. Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH 2004-ICSLP)ICSLP, vol. II; October. Jeju Island; Korea, 1109-1112.

Fant G. 1960. The Acoustic Theory of Speech Production. The Hague: Mouton.

Faria IH, Pedro ER, Duarte I, Gouveia CAM. 1996. Introdução à Linguística Geral e Portuguesa. Lisboa: Caminho.

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, vol. 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. Proceedings of the 4th International Speech Motor Conference; Nijmegen; Netherlands, 142-145.

Martins P, Carbone IC, Pinto A, Silva A, Teixeira AJ, 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): 1771-76.

NessAiver M, Stone M, Parthasarathy V, Kahana Y, Kots A, Paritsky A. 2006. Recording High Quality Speech during tagged Cine MRI studies using a fiber optic microphone. Journal of Magnetic Resonance Imaging 23: 92-97.

Ridouane R. 2006. Investigating speech production: A review of some techniques. [Online Document] Available at: http://lpp.univ-paris3.fr/equipe/rachid_ridouane/Ridouane_Investigating.pdf.

Russel GO. 1928. The vowel: In physiological mechanism as shown by x-ray. Columbus: Ohio State University Press.

Serrurier A & Badin P. 2005. 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, 2161-2164.

Shadle CH, Mohammad M, Carter JN, Jackson PJB. 1999. Multi-planar dynamic magnetic resonance imaging: new tools for speech research. Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS); S. Francisco; USA.

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. Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 98); Sydney, 3083-3086.

Takemoto H, Kitamura T, Nishimoto H, Honda K. 2004. A method of tooth superimposition on MRI data for accurate measuremente of vocal tract shape and dimensions. Acoust. Sci. & Tech. 25(6): 468-473.

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

Teixeira A, Moutinho LC, Coimbra RL. 2003. Production, Acoustic and Perceptual Studies on European Portuguese Vowels Height. Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, CD-Rom (ISBN: 1-876346-48-5), 3033-3036.

Teixeira A, Vaz F, Martinho L, Coimbra RL. 2001a. Articulatory Synthesis of Portuguese. In III Encontro do Forum International de Investigadores Portugueses; Institut EETA; Aveiro; Portugal.

Teixeira A, Vaz F. 2001b. European Portuguese Nasal Vowels: An EMMA Study. Proceedings of the Eurospeech 2001. Institut EETA; Aveiro; Portugal, 1483-1486.

FIGURE CAPTIONS

Figure 1. Vocal tract organs in a MRI midsagittal slice.

Figure 2. Noise and speech waveform recorded at the end of the acquisition of an MRI sequence, showing an SNR value smaller than -12dB. The right part of the waveform is representative of the speech.

Figure 3. A midsagittal reference image for teeth identification and contours extraction.

Figure 4. An example of image segmentation of the airway.

Figure 5. Three-dimensional model of the vowel /u/ sustained by a male subject.

Figure 6. Three-dimensional models of five EP oral vowels, of male (i) and female subjects (ii).

Figure 7. Three-dimensional models of EP nasals consonants, of male (i) and female subjects (ii).

FIGURES

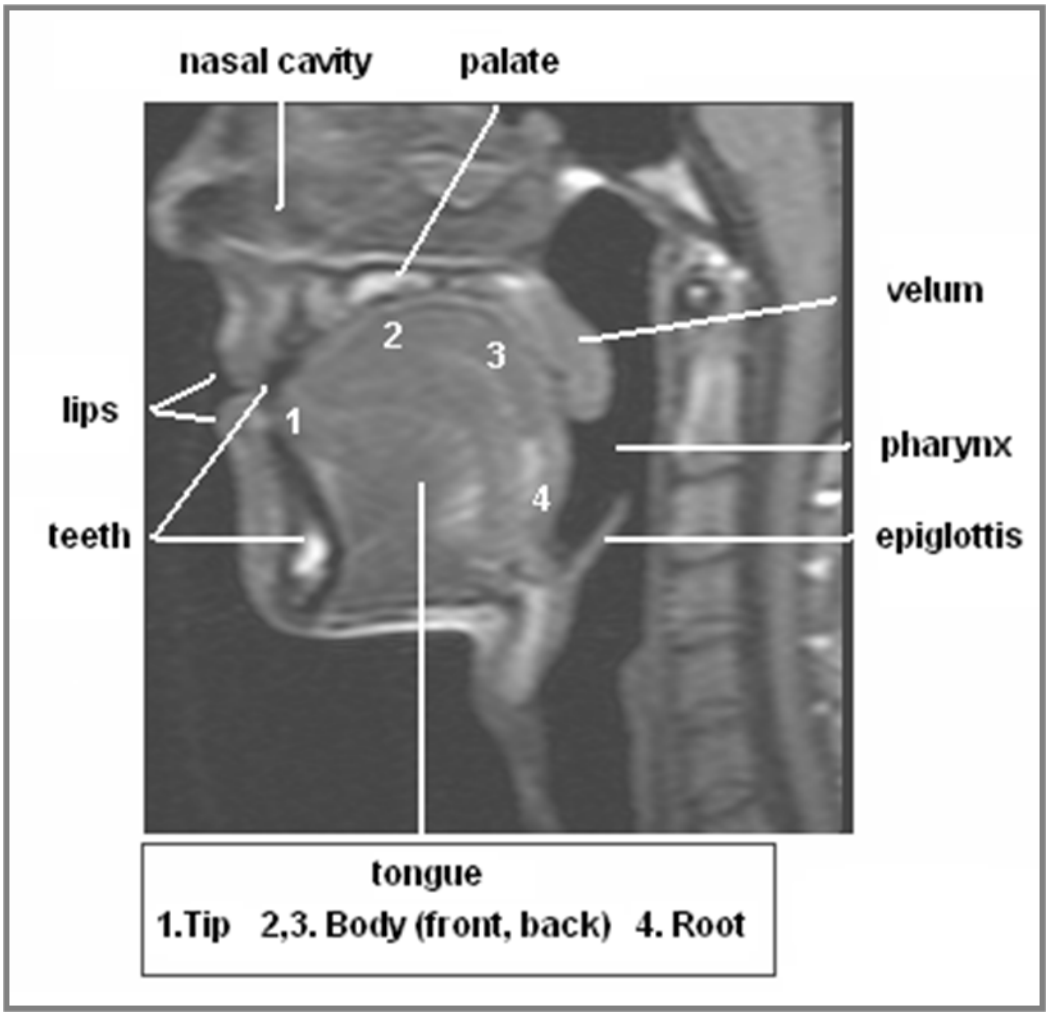

Figure 1

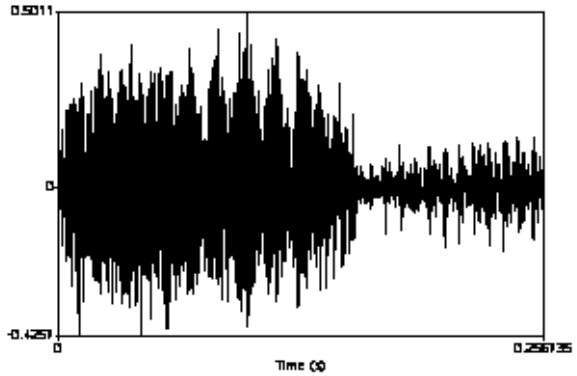

Figure 2

Figure 3

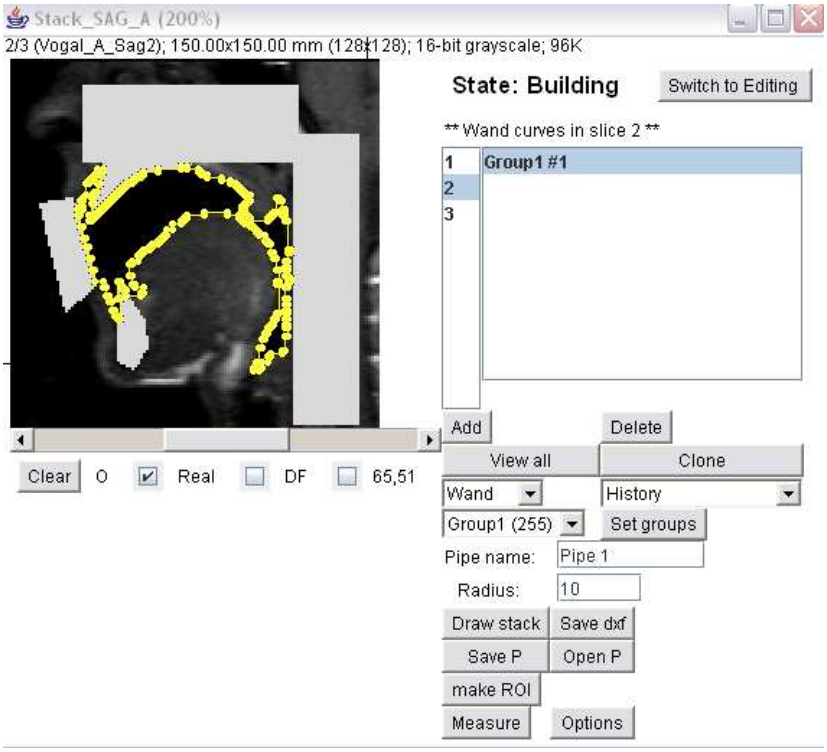

Figure 4

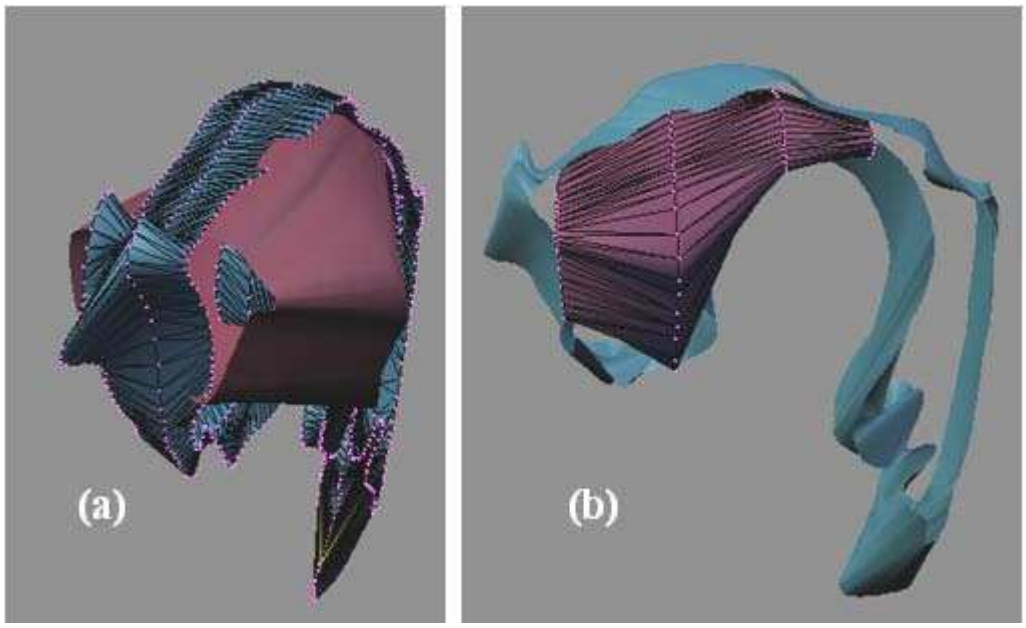

Figure 5

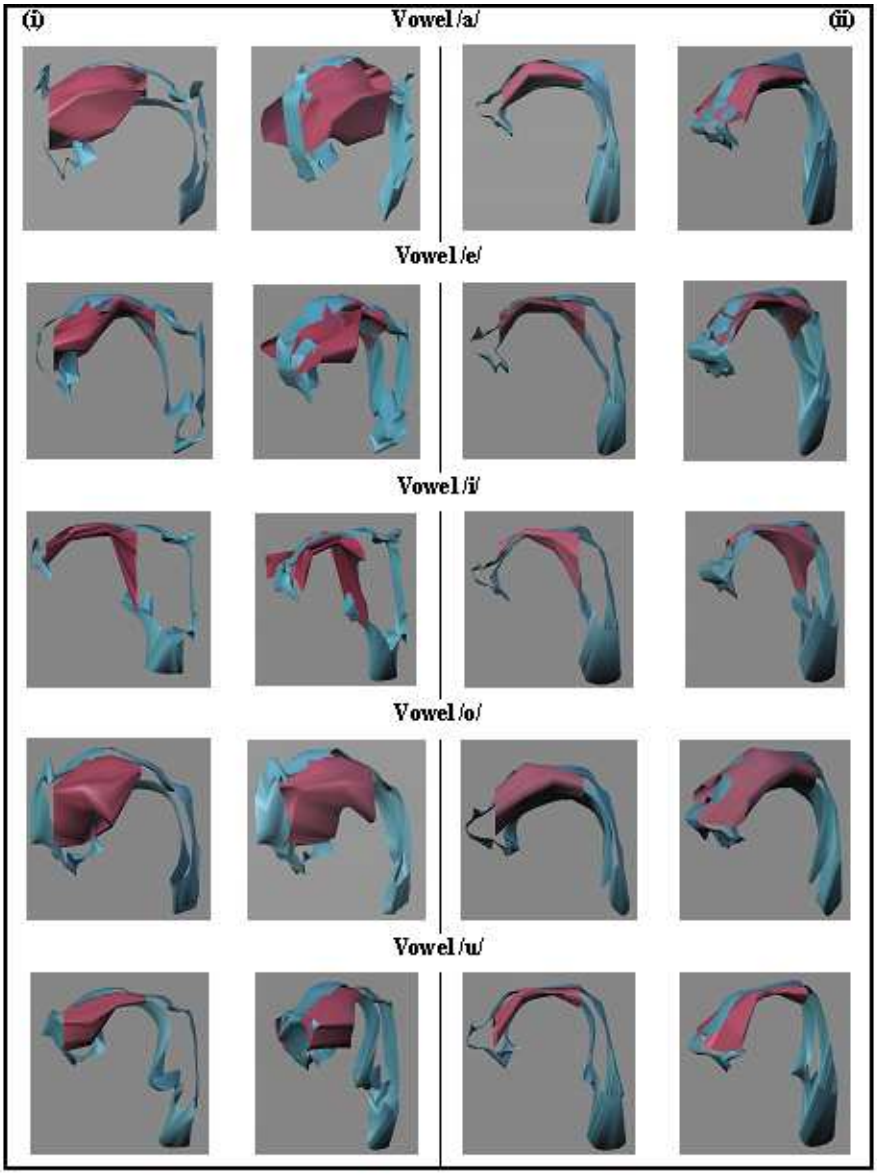

Figure 6

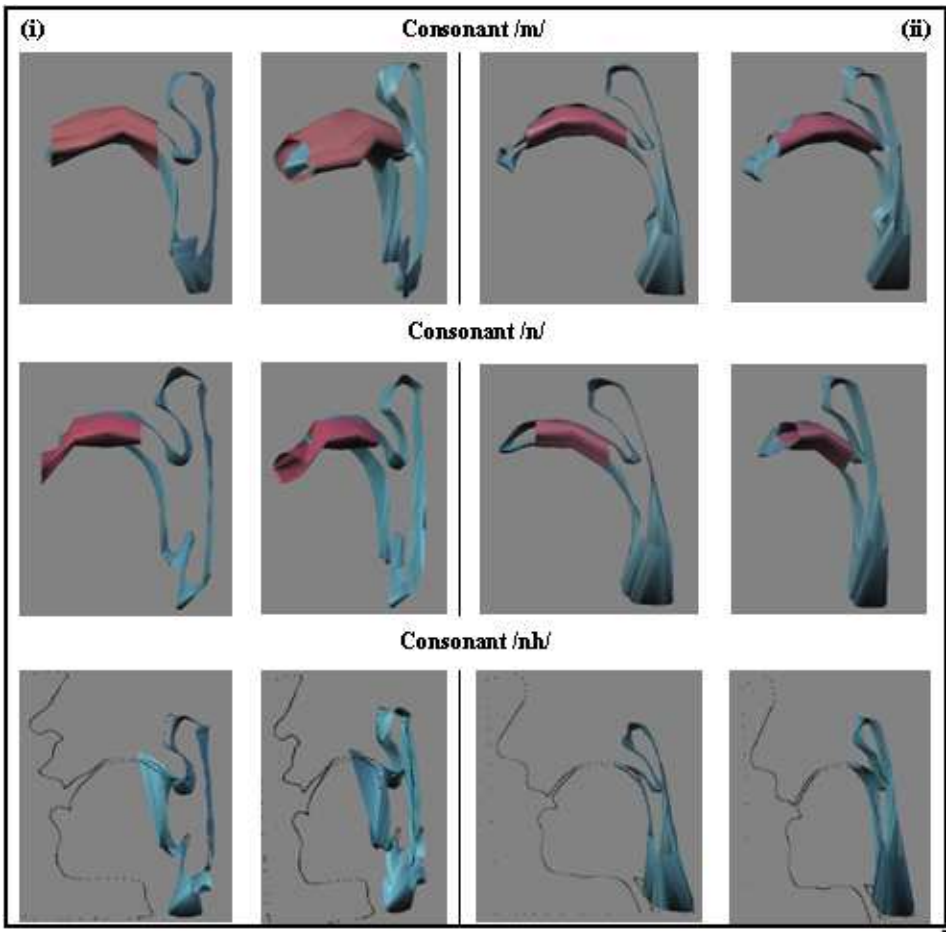

Figure 7