3D Vocal Tract Reconstruction using MAGNETIC RESONANCE IMAGING DATA TO STUDY FRICATIVE CONSONANT PRODUCTION

Sandra M. Rua Ventura[1], Diamantino Rui S. Freitas[2], Isabel Maria A. P. Ramos[3], João Manuel R. S. Tavares[4]

[1]Área Técnico-Científica da Radiologia, Escola Superior de Tecnologia da Saúde do Porto, Instituto Politécnico do Porto, Portugal

[2]Departamento de Engenharia Electrotécnica e de Computadores, Faculdade de Engenharia, Universidade do Porto, Portugal

[3]Serviço de Radiologia do Centro Hospitalar de S. João, EPE, Faculdade de Medicina, Universidade do Porto, Portugal

[4]Instituto de Engenharia Mecânica e Gestão Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidadedo Porto, Portugal, tavares@fe.up.pt

Abstract The development of Magnetic Resonance Imaging (MRI) has grown rapidly in clinical practice. Currently, the use of MRI in speech research provides useful and accurate qualitative and quantitative data of speech articulation. The aim of this work was to describe an effective method to extract vocal tract and compute their volumes during speech production from MRI images. Using a 3.0 Tesla MRI system, 2D and 3D images of the vocal tract were collected and used to analyze the vocal tract during the production of fricative consonants. These images were also used to build the associated 3D models and compute their volumes. This approach showed that, in general, the volumes measured for the voiceless consonants are smaller than the counterpart voiced consonants.

Keywords Magnetic Resonance Imaging (MRI) • Fricative Consonants • Vocal Tract Imaging • 3D Models • Volumetric Measurements

1 Introduction

This chapter has a particular interest for professionals related to speech study and rehabilitation, medical imaging and bioengineering. It is organized as follows: the Introduction section starts with a review from the literature about the use and feasibility of MRI to study the vocal tract, in particular, during speech production of fricative consonants. In addition, some articulatory phonetic concepts of Portuguese sounds are introduced. In the second section, the methodology adopted for the MRI acquisition and image data assessment is described. In the Results and Discussion sections the key aspects of the 2D and 3D visualization of the vocal tract during the production of sustained consonants are illustrated and discussed. In the final section, the conclusions are presented and some suggestions for future works are indicated.

1.1 Magnetic Resonance Imaging of the vocal tract

The first Magnetic Resonance Imaging (MRI) scan of the entire human body was carried out by Raymond Damadian, in 1977, to diagnose cancer. Since then the use of this imaging modality has grown rapidly in clinical practice. Its use is of interest in various medical areas and at the same time has opened up new fields of research.

Magnetic Resonance Imaging is a recognized and powerful non-ionizing diagnostic tool, employed for the diagnosis of various disorders, such as cancer, soft tissue damage, cardiology and neurology. The technique uses a magnetic field and radiofrequency waves to create detailed images of the organs and tissues, and relies on the nuclear magnetic resonance of the hydrogen nucleus. Compared with images from other techniques, MRI

provides excellent anatomical details [19] and is superior to computed tomography (CT) in distinguishing various tissue characteristics [13].

The images produced have good signal-to-noise ratio and contrast; however, the temporal resolution of MRI is very low when compared with radiographic techniques. In addition, some safety and contraindication issues must be taken into consideration, namely metallic fragments, clips or patients with magnetically activated implants and devices.

The vocal tract consists of a set of open air-cavities surrounded by soft-tissue, where sounds are produced. This region is similar to a long tube with a non-linear shape that extends from the vocal folds to the lips, with a side branch leading to the nasal cavity. Vocal tract organs (also called articulators) include the tongue, lips, teeth and alveolar ridge, hard palate, velum (soft palate), and the pharynx. All these organs (except the teeth) are well defined on MR images due to their good signal-to-noise ratio.

The main resonance cavities of the vocal tract include: nasal cavity (and nasopharynx), oral cavity, oropharynx and hypopharynx as shown in Fig. 1.

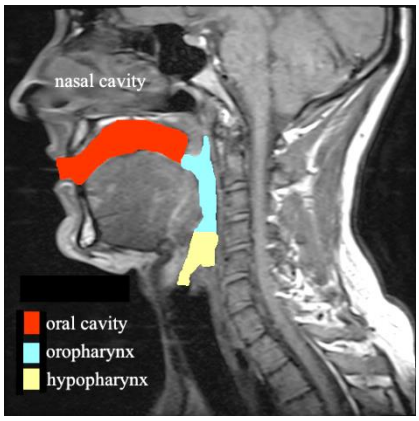


Fig. 1 Sagittal MR image of the vocal tract and its main resonance cavities.

As can be seen in Fig. 1, air has a low signal in MR images compared with the high signals of the fat and surrounding soft tissues of the vocal tract. This air conducting tube that extends along the lips and the glottis presents an irregular shape, controlled by the movement of the articulators during speech production.

The tongue is the most important articulator, mainly because it is the largest one, and performs a wide range of movements during speech production.

1.2 Measuring and modelling speech: a review

The task of modelling the vocal tract for articulatory synthesis systems aims to obtain the complete geometrical information concerning the vocal tract measured by different techniques, especially MRI [11, 14]. Currently, the use of MRI in speech research provides useful and accurate qualitative and quantitative data of speech articulation [21, 22, 23, 24].

Modelling of the vocal tract has traditionally been limited to two-dimensional (2D) images [2, 17] although, with the improvement of MRI acquisition techniques, three-dimensional (3D) modelling has become available [3, 5, 16]. Between 1986 and 1996, most of the vocal tract models to study speech production and articulatory synthesis relied on the extraction of midsagittal distances and cross-sectional area functions. These previous MRI studies were based on 2D multi-slice acquisitions requiring artificially sustained sounds or multiple repetitions of the same sound during long scan acquisition times. With this data, it was theoretically possible to try to reconstruct the vocal tract shape, but some important problems still remained, such as:

1. Air-tissue boundaries were hard to segment, i.e. to identify, and were time consuming;
2. Several image artefacts due to the long scan acquisition time;
3. Low signal from the teeth;

4. Hardware constraints: supine position imposed on the subjects and MRI system noise that inhibits high-quality audio recording.

To obtain more detailed information about the relation between vocal tract shape and speech sounds, 3D data are required [15, 18, 25]. For this purpose, only the MRI technique provides excellent structural differentiation of all the vocal tract organs and without harmful effects. At present, MRI is the most reliable and powerful imaging tool to acquire the full geometry of the vocal tract and to provide quantitative volumetric data.

Although multiplanar scanning may have been performed, only one MR image stack in one orientation (transversal, sagittal or coronal) is usually used in vocal tract reconstruction. Several researchers also used this approach, due to the common MRI-related constraints: acquisition times, costs and fatigue of the subjects under study. In order to improve vocal tract modelling, some techniques have been proposed that combine manually [1, 4, 14] or even automatically [26] orthogonal image stacks. On the other hand, the first application of compressed sensing to high-resolution 3D upper airways was provided in [9]. The authors demonstrated that it is possible to acquire 3D imaging data of the vocal during a single sustained sound production stack (without sound repetitions). In line with this work, other studies have been performed, particularly for European Portuguese sounds [12, 24].

1.3 Articulatory phonetics: sounds of Portuguese speech

The standard European Portuguese speech system consists of nine oral and five nasal vowels, three diphthongs and nineteen consonants (six plosives, six fricatives, three nasals and four liquids). There are two main classes of consonants: plosives and fricatives. Plosives are sounds in which the air streams from the lungs are interrupted by a complete closure in some part of the vocal tract. In fricative sounds, the air passes usually through a narrow constriction that causes the air to flow turbulently and

thus create a loud sound [7]. The other classes of consonants that are found in the majority of languages (nasals, "liquids" and vowel-like approximants) are voiced in the overwhelming majority of cases.

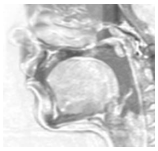
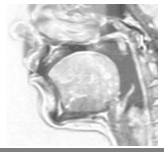
Consonants can be classified according to three major features:

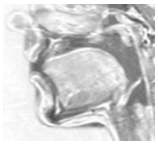
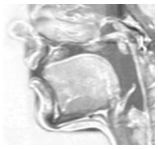
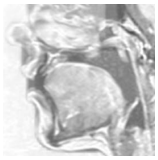
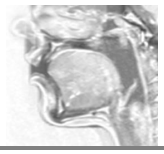
- Place of articulation - specifies where the constriction in the vocal tract is;
- Manner of articulation - how narrow the constriction is, whether air is flowing through the nose, and whether the tongue is dropped down on one side;
- Voicing - specifies whether the vocal folds are vibrating.

One of the major ways that consonants differ from each other is in the accompanying action of the larynx; most larynx settings allow air to flow freely between the vocal folds versus others when the vocal folds vibrate to produce regular voicing.

For the European Portuguese language, the distribution of the sounds between voiced and voiceless counterparts in the fricative consonants is /v, z, ʒ/, and /f, s, ʃ/, respectively. The main configuration of the vocal tract during the production of fricative consonants is illustrated in Table 1.

Table 1. Vocal tract configuration associated to the European Portuguese fricative consonants.

Fricative consonant	Description	Vocal tract configuration
[f]	Voiceless labiodental Reference word: /f̲é/ (faith)	
[v]	Voiced labiodental Reference word: /v̲ê/ (see)	

[s]	Voiceless alveolar Reference word: /sol/ (sun)	
[z]	Voiced alveolar Reference word: /casa/ (home)	
[ʃ]	Voiceless post-aveolar Reference word: /já/ (already)	
[ʒ]	Voiced post-aveolar Reference word: /chave/ (key)	

2 Methodology

2.1 Magnetic resonance imaging: protocol and procedures

The MRI data used in this study were acquired at the Radiology Department of the Centro Hospitalar de S. João, Porto, in Portugal, using a MAGNETOM Trio 3.0 Tesla MRI system (Siemens AG, Germany) and two integrated coils: a 32-channel head coil and a 4-channel neck matrix coil, with the subjects in the supine position.

The experiments were performed with two young native European Portuguese (EP) subjects, one male and one female and according to the MRI safety procedures. In addition, a questionnaire was carried out before any procedure to screen any contraindications. The subjects were previously informed and instructed about the study, and their informed

consents were collected. Verbal communication was maintained with the subjects through an intercom.

The speech corpus included the sustained sounds of the EP language: three pairs of fricative consonants /f v/, /s z/ and /ʃ ʒ/ according to the International Phonetic Alphabet (IPA).

The first MRI dataset was collected using a turbo spin echo 2D sequence, to be used as the sound articulation references, and consisted of one T1-weighted 4 mm thick midsagittal slice for each sound. The following imaging acquisition parameters were adopted: a repetition time of 400 ms, an echo time of 10 ms, an echo train length of 5, a square field of view of 240 mm, a matrix size of 512 x 512 pixels, and a pixel size of 0.469 x 0.469 mm^2. The acquisition time lasted around 8.07 seconds for each sound. Afterwards, a 3D volumetric MR acquisition of the vocal tract resonance cavities was carried out using a flash gradient echo sequence. A 120 mm thick transversal slab was acquired according to the following parameters: repetition time of 5.8 ms, echo time of 2.17 ms, flip angle = 10º, a square field of view of 270 mm, a matrix size of 256 x 256 pixels and a pixel spacing of 1.055 x 1.055 mm^2. The imaging acquisition lasted around 16.12 seconds for each sound. From this slab, 60 2-mm-thick slices were reconstructed automatically.

2.2 Image processing and analysis: tasks and techniques

The segmentation of the vocal tract, for the image processing and analytical tasks, was automatically performed in each slice followed by image-based reconstruction of the vocal tract shape using the ITK-SNAP software (version 2.1.4-rc1), which was developed at the Penn Image Computing and Science Laboratory (PICSL) in USA. Similar to other imaging software packages for 3D analysis, this software provides reliable and accurate volume segmentation of the oropharynx for upper airway assessment [20].

The three-dimensional models were built from the subset of sagittal images defining the whole vocal tract, the oral cavity and the pharynx. The original set of 60 2-mm-thick slices, and without any gap between them, was segmented using the active contour method, usually known as snakes, proposed by [8], which has revealed robustness against image noise and efficiency in images with low signal-to-noise ratio as is the case of MR images. Active contours are curves that are moved and deformed by internal and external forces until they reach the object's boundary in the image to be segmented.

In order to segment the vocal tract, a pre-processing of the original images was performed in order to adjust the range of the airway voxel intensities (threshold interval selection). After the segmentation, a 3D model of the airway of each resonance cavity was automatically built from the contours segmented in each slice. As a result of the automated segmentation and labelling of each resonance cavity, the software could compute the volume of the vocal tract.

3 Results

The complete 2D MR image dataset is depicted in Fig. 2. The images have good signal-to-noise ratio and resolution, which allows a clear visualization of the different vocal tract shapes produced for each EP sound. These reference images were then used to define the slices acquired and also to confirm the proper production of the intended sound.

For the segmentation of each consonant imaged, first the vocal tract was segmented as a whole, and then the oral cavity and pharynx were segmented individually.

	Male	
Voiceless		**Voiced**

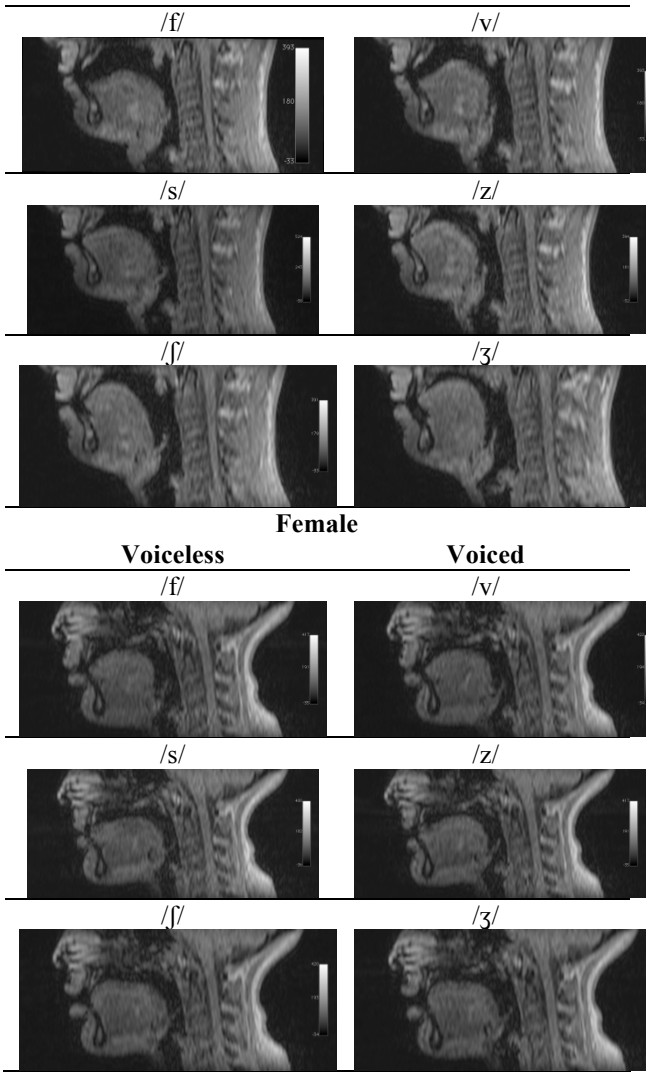
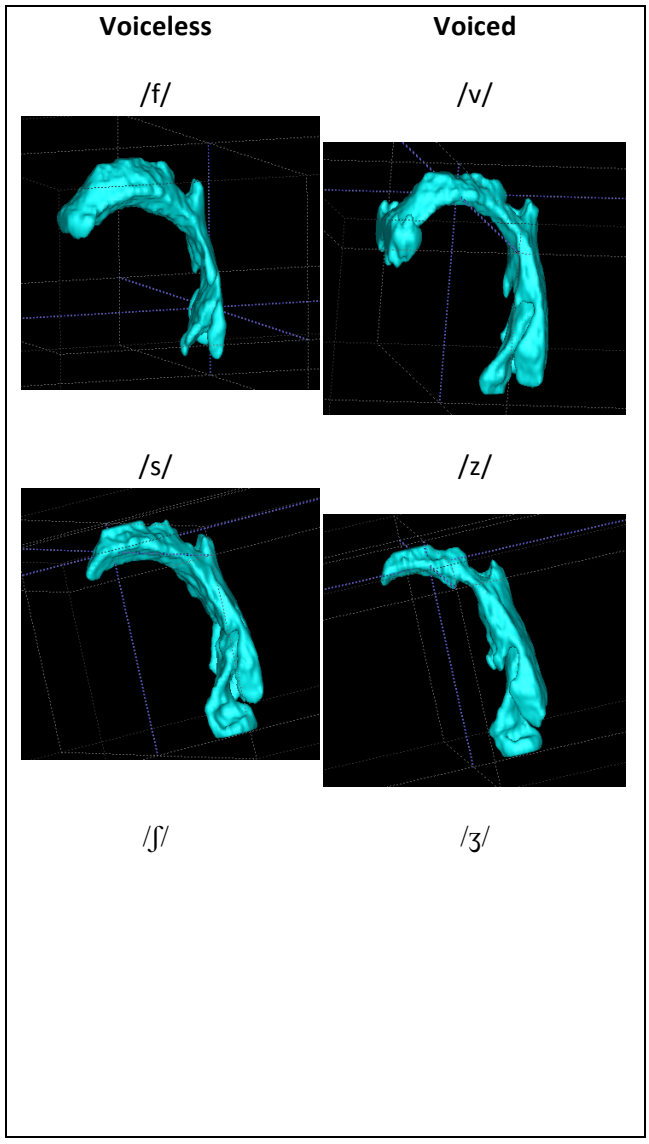


Fig. 2 Midsagittal MR slices of the vocal tract of a male and a female during EP speech production.

The analysis of the whole vocal tract by means of the 3D models built confirms the occurrence of important articulatory changes, as shown in Fig. 3. The 3D models show that the main difference found is the increase of the overall volume of the vocal tract, especially in the pharynx cavity.

In order to get measurable data concerning the articulatory changes found in the morphology of the vocal tract and the major resonance cavities, the volumes were calculated and then compared. The results obtained are shown in Table 2 according to each speaker.



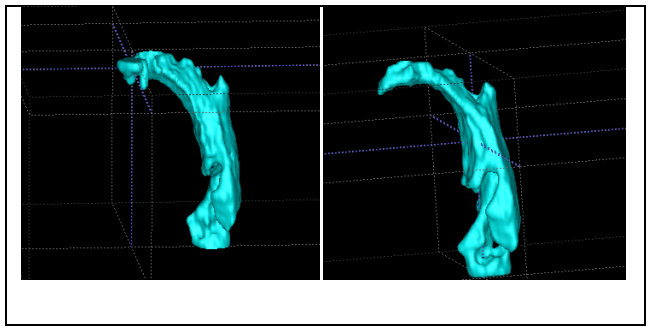


Fig. 3 3D vocal tract models built for the voiceless fricative consonants from the image dataset of the male subject.

Table 2. Volumetric measurements of the major resonance cavities computed for each pair of fricative consonants (voiceless versus voiced) for each subject under study.

	Volumes (mm^3)	/f/	/v/	/s/	/z/	/ʃ/	/ʒ/
Male	Total vocal tract	52007	64717	62014	75075	86704	107420
	Oral cavity	28432	22614	19711	24287	26383	48770
	Pharynx	23575	42103	42303	50788	60321	58650
Female	Total vocal tract	47777	50207	52142	54830	37361	44055
	Oral cavity	12240	13192	17159	17488	12062	12718
	Pharynx	35537	37015	34983	37342	25299	31337

4 Discussion

In this chapter, a morphological assessment of the vocal tract during sustained production of EP by MRI is presented. From the approach adopted, morphologic information can be extracted in both 2D and 3D MRI sequences concerning the shape and dimensions, respectively, of the vocal tract and the two major resonance cavities during the production of EP fricative consonants.

On the 2D sagittal images acquired, only subtle differences can be observed among each pair of fricative consonants, especially for the male subject. The vocal tract shapes associated to the female subject were

more difficult to establish in the MR images acquired due to the smaller anatomy.

The results obtained using the 3D models built and the quantitative measurements made, confirmed the important articulatory changes among the fricative consonants and for each subject. The main difference found between the subjects was that the volumes measured for the voiceless consonants were smaller than their counterpart voiced consonants. Excluding the male subject for the consonant /f/, all pharynx volumes measured were superior to the oral cavity volumes. The study in [14] also found larger pharyngeal volumes in voiced fricatives by means of vocal tract lengths and area function measurements. Equivalent results were also reported in [16] where the authors observed some supraglottal volume changes of pharyngeal articulation during the production of voiced and voiceless fricative consonants.

The volumes of the resonance cavities of the male subject are larger for the alveolar and post-alveolar sounds; additionally, the highest volumetric value was attained during the production of the consonant /ʒ/.

Comparing the volumes obtained for the female subject, larger dimensions were observed associated to labiodental and alveolar sounds, and the largest volume measured was during the production of the consonant /z/.

One issue that should be taken into account in this study is the artificiality of the speech task. MRI requires a supine posture and unnaturally long times to sustain sounds for 3D imaging sequences. Several authors have discussed these problems but without finding a solution [6, 10]; however, they should not affect the fundamental findings of this work. The larger pharyngeal volume, found in this study, in voiced fricatives is consistent with the study in [16]. In addition, the MRI acquisition times in this work were considerably lower (around 16 seconds) than the sustain times used in that study (around 36 seconds).

This study has addressed only the morphological aspects of the vocal tract and major resonance cavities during the production of EP fricative consonants; however, there are many aspects of fricative production and voicing, which remain to be investigated.

5 Conclusion

The approach used to visualize the vocal tract during the production of fricative consonants in MR images, and posterior building of the associated 3D models proved to be very effective. Quantitative measurements of volumes for each pair of fricative consonants (voiceless versus voiced) for each subject were also successfully attained.

The 3D models built for the vocal tract showed that the volumes measured for the voiceless consonants are smaller than the counterpart voiced consonants.

Applying the present technology, volumetric MRI can be successfully applied to obtain 3D models, as well as measurable data concerning the vocal tract resonance cavities and all the articulators involved. More phonetic data are required to endorse the differences in the volume measures of each fricative consonant, and to reduce some inconsistencies observed between subjects and at different places of articulation.

Due to the developments that have recently taken place in MRI systems, namely the use of 3.0 Tesla magnetic fields, new applications and image refinements are expected, and consequently, significant improvements in the quality of the data acquired for articulatory events during speech production. This 3D enhanced knowledge about 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 the acoustic theory in speech production.

Until today, the automatic and robust interpretation of biomedical images is still a major goal. Thus, a future development concerning medical image processing and analysis will be the increased integration of such algorithms and their applications in clinical practice.

Acknowledgments The images were acquired at the Radiology Department of the Hospital S. João, Porto, in Portugal, with the collaboration of the technical staff, to whom we are most grateful. The first author would like to thank the support and contribution of the PhD grant from Escola Superior de Tecnologia da Saúde (ESTSP) and Instituto Politécnico do Porto (IPP), in Portugal. This work was partially done in the scope of the project with reference PTDC/BBB-BMD/3088/2012, financially supported by Fundação para a Ciência e a Tecnologia (FCT), in Portugal.

References

1. Baer T, Gore JC, Gracco LC, Nye PW (1991) Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels. Journal of Acoustical Soc. Am. 90(2): 799–828
2. Badin P, Bailly G, Raybaudi M, Segebarth C (1998) A three-dimensional linear articulatory model based on MRI data. In: Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 98), pp 417–420
3. Badin P, Serrurier A (2006) Three-dimensional modeling of speech organs: Articulatory data and models. IEIC Technical Report of the Institute of Electronics, Information and Communication Engineers 106(177): 29–34
4. Demolin D, Metens T, Soquet A (1996) Three-dimensional Measurement of the Vocal Tract by MRI. In: Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP 96), pp 272–275
5. Doel KVD, Vogt F, English R, Fels S (2006) Towards Articulatory Speech Synthesis with a Dynamic 3D Finite Element Tongue Model. In: Proceedings of the International Seminar on Speech Production, pp 59–66
6. Engwall O and Badin P (2000) An MRI study of Swedish fricatives: coarticulatory effects. In: Proceedings of the 5th Seminar on Speech Production: Models and Data, pp 297–300
7. Jackson P and Shadle C (2000) Aero-acoustic modelling of voiced and unvoiced fricatives based on MRI data. In: Proceedings of 5th Speech Production Seminar, pp 2–5
8. Kass M, Witkin A, Terzopoulos D /1987) Snakes: active contour models. International Journal of Computer Vision 1(4): 321-331
9. Kim Y-C, Narayanan SS, Nayak KS (2009) Accelerated three-dimensional upper airway MRI using compressed sensing. Magnetic Resonance in Medicine 61(6): 1434–1440
10. 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
11. Kröger B and Birkholz P (2009) Articulatory synthesis of speech and singing: State of the art and suggestions for future research. **E**sposito A, Hussain A, Marinaro M, Martone R (Eds.).

Multimodal Signals: Cognitive and Algorithmic Issues, Lectures Notes in Computer Science 5398 pp 306–319
12. Martins ALD, Mascarenhas NDA, Suazo CAT (2010) Temporal Resolution Enhancement of Vocal Tract MRI Sequences Based on Image Registration. In: 17th International Conference on Systems, Signals and Image Processing (IWSSIP 2010), pp 190–193
13. Muller NL (2002) Computed tomography and magnetic resonance imaging: past, present and future. European Respiratory Journal 19(35): 3s–12s
14. Narayanan SS, Alwan AA, Haker K (1995) An articulatory study of fricative consonants using magnetic resonance imaging. The Journal of the Acoustical Society of America 98(3): 1325–1347
15. Niikawa T, Matsumura M, Tachimura T, Wada T (2000) Modeling of a speech production system based on MRI measurement of three-dimensional vocal tract shapes during fricative consonant phonation. In: Interspeech, pp 174–177
16. Proctor MI, Shadle CH, Iskarous K (2010) Pharyngeal articulation in the production of voiced and voiceless fricatives. The Journal of the Acoustical Society of America 127(3): 1507–1518
17. Serrurier A and Badin P (2005) Towards a 3D articulatory model of velum based on MRI and CT images. Papers in Linguistics *40*(1): 195–211
18. Shadle C, Proctor M, Iskarous K (2008) An MRI study of the effect of vowel context on English fricatives. In: European Conference on Noise Control, pp 5101–5106
19. Symms M, Jäger HR, Schmierer K, Yousry TA (2004) A review of structural magnetic resonance neuroimaging. Journal of Neurology, Neurosurgery, and Psychiatry *75*(9): 1235–1244
20. Weissheimer A, Macedo de Menezes L, Sameshima GT, Enciso R, Pham J, Grauerf D (2012) Imaging software accuracy for 3-dimensional analysis of the upper airway. American Journal of Orthodontics and Dentofacial Orthopedics 142(6): 801-813
21. Vasconcelos MJ, Ventura SR, Freitas DR, Tavares JMRS (2011). Inter-speaker speech variability assessment using statistical deformable models from 3.0 Tesla magnetic resonance images. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine 226(3): 185–196
22. Ventura SR, Freitas DR, Ramos IM, Tavares JMRS (2011) Requisitos e condicionantes da imagem por ressonância magnética no estudo da fala humana. In: Congresso de Métodos Numéricos em Engenharia (CMNE), pp 1–12
23. Ventura SR, Freitas DR, Tavares JMRS (2011) Toward dynamic magnetic resonance imaging of the vocal tract during speech production. Journal of Voice 25(4): 511–518
24. Ventura SR, Freitas DR, Ramos IM, Tavares JMRS (2013) Morphological Differences in the Vocal Tract Resonance Cavities of Voice Professionals: An MRI- based Study. Journal of Voice 27(2): 132–140
25. Ventura SR, Freitas DR, Tavares JMRS (2009) Application of MRI and biomedical engineering in speech production study. Computer Methods in Biomechanics and Biomedical Engineering 12(6): 671–81
26. Zhou X, Woo J, Stone M, Prince JL, Espy-Wilson CY (2013) Improved vocal tract reconstruction and modeling using an image super-resolution technique. The Journal of the Acoustical Society of America 133(6): 439–445