In: Language Acquisition
Editor: Leslie M. Warfelt, pp.
ISBN 978-1-61209-569-1

Chapter 4

SPEAKER-SPECIFIC ARTICULATORY ASSESSMENT AND MEASUREMENTS DURING PORTUGUESE SPEECH PRODUCTION BASED ON MAGNETIC RESONANCE IMAGES

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

[1] Radiology Department, School of Allied Health Science –
Porto Polytechnic Institute (IPP) / Faculty of Engineering,
University of Porto – PORTUGAL
[2] Laboratory of Optics and Experimental Mechanics,
Institute of Mechanical Engineering and Industrial Management /
Department of Mechanical Engineering,
Faculty of Engineering, University of Porto – PORTUGAL
[3] Department of Electrical Engineering and Computers,
Faculty of Engineering, University of Porto – PORTUGAL
[4] Radiology Service, São João Hospital / Faculty of Medicine,
University of Porto – PORTUGAL

ABSTRACT

The development of two and three-dimensional magnetic resonance imaging (MRI) opened new options for a better understanding of speech production; in particular, for the articulation process, comparing with other imaging techniques, such as x-rays. Several MRI studies have been carried out considering different languages, but concerning European Portuguese the available information is far from being completely achieved. Recently, the knowledge gained with the application of deformable models in magnetic resonance images towards the automatic study of the vocal tract, has allowed an

[*] Email: smr@estsp.ipp.pt / sandra.rua@eu.ipp.pt

enhanced identification and description of the articulatory mechanism and its organs. Our aim is to extract and evaluate the main characteristics of the movements of vocal tract during European Portuguese speech production to achieve speaker-specific articulatory assessment from MRI. For this, we used active shape and active appearance models to identify, i.e. to segment, and simulate the vocal tract's shape in MR images and concluded that both are suitable for such tasks being the later more proficient. The findings obtained are believed to be useful for speech rehabilitation and simulation purposes, namely to recognize and simulate the compensatory movements of the articulators during speech production.

Therefore, this chapter gains particular interest within the communities of speech study and rehabilitation, medical imaging and bioengineering. It is organized as follows: the introduction section starts with a literature review concerning the use, application and challenges of MRI in speech production study, in particular for speech articulation. In addition, image analysis techniques based on deformable templates, more specifically by using geometrical shapes driven by parameterized functions, are introduced. In the second section, the adopted methodology of MRI acquisition and data assessment are described. Based on this image analysis approach, in the results section the key aspects of articulatory movements during the production of relevant European Portuguese speech sounds are addressed. In the final section of this chapter, conclusions are presented and further suggestions for future work are indicated.

1. INTRODUCTION

1.1. Review on Speech Production Studies

Speech research is an interesting multidisciplinary field for several areas, including medicine, engineering and phonetics. Since the X-Rays discovery by W. Roentgen in 1895, several imaging techniques have been developed giving new perspectives in morphologic and dynamic knowledge of the human anatomy. Concerning the articulatory organs, the first applications of Magnetic Resonance Imaging (MRI) proposed by Rokkaku et al. [1] and Baer et al. [2] have opened new perceptions of the speech production process, allowing the capture of two-dimensional (2D) and three-dimensional (3D) images of all organs of the vocal tract.

Several MRI studies have been carried out for few subjects in different languages, such as French [3, 4], German [5], Swedish [6] and Japanese [7]. Concerning European Portuguese (EP) the available information is far from being completely achieved [8, 9, 10]. First speech research studies have been based on the production of vowels and of some consonant groups [11, 12], from static [13] to dynamic imaging [14, 15, 16], and more recently in real-time imaging [17, 18, 19].

The surprising advances in the hardware and software of the MRI systems, which have been occurred in the last 25 years, permitted to get considerable results in speech research. According to Masaki et al. [20] the developmental history of MR images of speech organs can be divided into three stages; the first stage, in the 1980s, is the application of MR as still an imaging technique, where studies have to deal with the long acquisition times requiring long sustained articulations for capturing the morphology of the vocal tract from midsagittal slices. The second stage (1990s) is relative to the application of the cine sequences of MRI to perform dynamic studies by means of several repetitions of sounds. Currently, at the third

stage, we can observe the refinement of MRI data acquisition and visualization techniques for speech production studies.

1.2. Challenges in Magnetic Resonance Imaging and Vocal Tract Modeling

The human vocal tract is a part of both respiratory and digestive systems and has an essential role on speech production. Many of the associated organs, including the lips, the teeth, the tongue, the palate and the pharynx, outline a set of cavities with significant acoustic features. For speech production, the tongue is the articulator with higher mobility and with very flexible movements. In general, the study of the articulators is a difficult task due to their high degree of freedom during the speech production process, which causes problems in their control and observation.

Magnetic Resonance Imaging is one of the most exciting fields of medical imaging due to its ability to produce images almost equivalent to slices of the body anatomy, and this in 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; afterwards, proper excitation is done by a set of coils placed near the zone under observation in the presence of the high-valued magnetic field (produced by the large superconducting magnet and the gradient generation coils that surround the patient).

Multi-slice imaging provided by MR allows direct measurements of the vocal tract shape and the calculation of area functions. However, some MRI drawbacks must be compensated, namely 1) the teeth imaging (teeth produce low signals and are poorly seen in MR images), 2) the subjects body position (the equipment's geometry imposes to the subjects a non-usual position), and 3) the intense noise produced which deteriorates the quality of audio recording [21]. As Engwall [22] observes, MRI gives adequate information about the 3D shape of the vocal tract and articulators, but it must be taken into account that sustained sounds cause hiperarticulated speech and the subject's body position affects the tongue's shape and often with the decrease of the pharynx cavity.

The excellent contrast and spatial resolution of soft-tissues and the exclusion of risks for patients are some of the most important advantages of MRI, and are the major reasons for its wide use in speech research [17, 19, 22, 23, 24]. Moreover, as claimed by Badin and Serrurier [25], detailed 3D knowledge of the vocal tract's shape by MRI is important to improve aerodynamic models for speech production studies and is useful in the domain of speech rehabilitation.

The accurate measurement of the vocal tract's 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 the acquired data). This approach, according to Badin et al. [26], lead to a number of problems, such as: 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. [3] 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 midsagittal slice.

According to the authors, the data offer 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 more reasonable acquisition speed. For example, recently, Kim et al. [27] demonstrated the application of compressed sensing[1] MRI to high-resolution 3D imaging of the vocal tract during a single sustained sound production task, in 7 seconds of scanning time.

1.3. Statistical Modeling Methods in Speech Production Studies

Active contours, deformable templates, physical models and point distribution models are examples of the most relevant deformable models that have been used to extract objects' characteristics from images.

Active contours, more frequently called snakes [28], were the first deformable models used in image analysis. They consist in a set of points capable of adapting to the object under study through a combination of internal and external forces. Deformable templates, use similar shapes of the object under study (templates) described by parameterized functions in order to correctly identify it [29]. Physical modeling allows the incorporation of previous physical knowledge of the object in the developed model [30]. Finally, Point Distribution Models are built from a set of training images and extract the main characteristics of the object using statistical techniques [31].

The application of statistical methods to speech production data extends back to the mid seventies, where Harshman et al. [32] reported the use of component analysis to identify a set of articulatory features for tongue shapes. Shirai and Honda [33] also applied statistical analysis of real data to describe the position of the articulatory organs and Maeda [34] described the lateral shapes of the vocal tract through factor analysis. Another example is the work of Stone et al. [35] that employs principal component analysis to examine the sagittal tongue contours taken from ultrasound images.

2. METHODOLOGY

2.1. Magnetic Resonance Imaging Protocol and Procedures

Magnetic resonance scan protocols are efficient and useful standard tools for imaging any anatomic region or structure and are defined in function of the imaging technique. According to the usual regulated procedures for MRI, the subjects were previously informed about the undergoing imaging exam and subsequently instructed about the procedures to be adopted, whereupon they signed a consent form. For vocal tract imaging, the subjects are asked to produce different speech sounds, during one determined and required time. Furthermore, the entire corpus acquisition must be explained, and they must be prepared to sustain a specific

[1] Compressed sensing - Emerged in the literature of Information Theory and Approximation Theory as an abstract mathematical idea, and aims to reconstruct signals and images from significantly fewer measurements than the ones that were traditionally thought as required.

sound during the defined time. The training of the subjects is therefore needed for good speech-acquisition synchronization, and to ensure the production of the intended sound.

The acquisition of a T1-weighted sagittal slice with 5 mm thickness was done using Turbo Spin Echo Sequences, with a recording duration around 3.3 seconds. The remainder acquisition parameters used were: field of view of 150 mm, image matrix with 128x128 pixels and image resolution of 0.853 px/mm.

In this work, only midsagittal images were used for a proof of concept, but in the near future image stacks will be used in 3D reconstruction.

2.1.1. Equipment and Corpus

Image acquisition was performed using a 1.5T Siemens Magneton Symphony system and a head array coil. One young male volunteer, without speech disorders, was positioned in supine position and, to permit intercommunication during the image acquisition and to reduce MR acoustic noise, headphones were used. The speech corpus consisted of a set of images collected during sustained articulations of twenty-five European Portuguese sounds (i.e. nine oral and five nasal vowels, two lateral consonants, six fricative consonants and three nasal stop consonants). Due to the MR acoustic noise produced during the acquisition process, the speech recording had no enough quality, and therefore was discarded.

2.1.2. Midsagittal Data Set

A total of 25 MR images of the vocal tract were acquired during sustained articulation, one for each EP sound. A reference word was provided previously to the subject, for a better perception of the intended sound. The midsagittal slice was chosen because this slice orientation gives information about the length and shape of the vocal tract and namely the tongue's position in the oral cavity. Some of the midsagittal images of EP sounds are presented in Figure 1. As observed, air cavities of the vocal tract (represented with low signal intensities) are well differentiable from the surrounding soft-tissues (represented with higher signal intensities).

2.2. Statistical Deformable Modeling

The set of images acquired according to the protocol previously described were used to model the vocal tract's shape. The used statistical modeling procedures are explained in the following sections.

2.2.1. Labeling Process

In the building of a Point Distribution Model (PDM), the shape represented in each image of the training set is defined by a group of labeled landmark points, usually representing important features of the boundary or inner regions of the shape to be modeled, as can be seen in Figure 2. (To help the visualization of the landmark points, in this image and in the subsequent ones, the landmark points are represented connected by fictitious line segments.)

The manual process of labeling the structure to be modeled is usually the simplest one. Still, this considers the premise that the user has technical knowledge about the shape

involved so as to choose the best locations for the landmarks and consequently, be able to mark it correctly in each image of the training set.

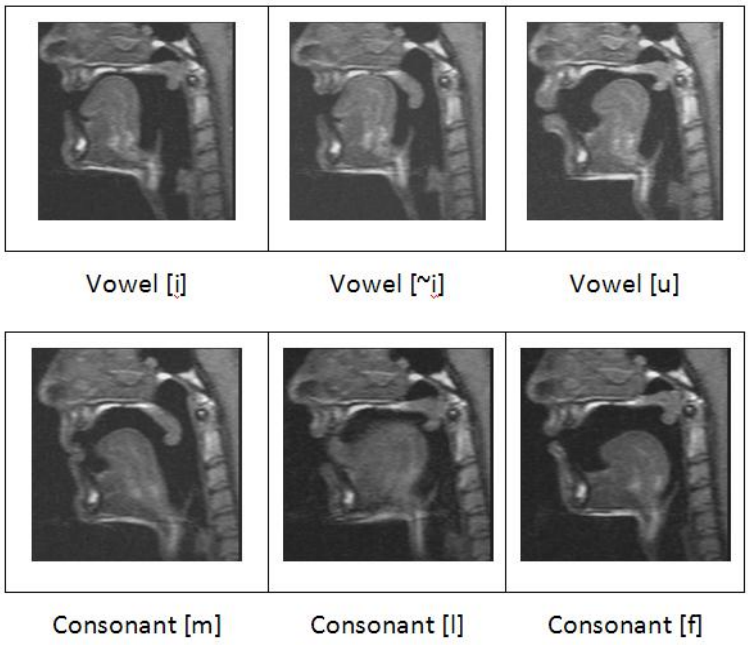

Figure 1. Midsagittal MR images of the vocal tract during vowels and consonants production of EP speech language.

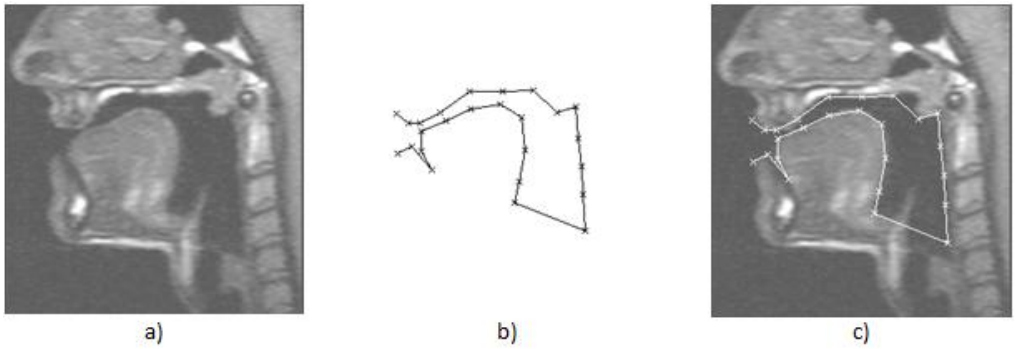

Figure 2. a) Training image, b) landmark points selected, c) original image labeled a) overlapped with the landmark points selected.

The manual tracing of the landmark points was carried out by one of the authors with medical imaging knowledge, according to the subsequent rules, and was realized on images sequentially displayed on the computer screen and later crosschecked by another author. The labeling method was performed according to the anatomic location of the vocal tract articulators, by the followed set of points:

- Four points at the lips (front and back of lips' borders);
- Three points corresponding to the lingual *frenulum* and tongue's tip;

- Seven points equally spaced along the surface of the tongue;
- Seven points along the surface of the hard palate (roof of the oral cavity) placed in symmetry with tongue points;
- One point at the velum (or soft palate);
- Three points equally spaced at the posterior margin of the oropharynx (behind the oral cavity).

Please note that the epiglottis was not taken into account during this task.

2.2.2. Statistical Modeling of the Vocal Tract

In order to study the variation of the coordinates of the landmark points of the training shapes it is necessary to align them, using, for example, dynamic programming [36]. Hence, given the co-ordinates $\left(x_{ij}, y_{ij}\right)$ at each feature j of shape i, the shape vector is:

$$x_i = \left(x_{i0}, x_{i1}, \ldots, x_{in-1}, y_{i0}, y_{i1}, \ldots, y_{in-1}\right)^T,$$

where $i = 1 \ldots N$, with N representing the number of shapes, i.e. configurations, in the training set and n the number of landmark points. Once the shapes are aligned, the mean shape and the variability of the modeled shape can be found. The modes of variation characterize the ways that landmarks of the modeled shape tend to move together and can be found applying principal component analysis to the deviations from the mean. So, each vector x_i can be rewritten as:

$$x_i = \bar{x} + P_s b_s, \tag{1}$$

where x represents the n landmark points of the new configuration of the modeled shape, $\left(x_k, y_k\right)$ is the position of landmark point k, $\bar{x}$ is the mean position of the landmark points, $P_s = \left(p_{s1} \quad p_{s2} \quad \ldots \quad p_{st}\right)$ is the matrix of the first t modes of variation, p_{si} correspond to the most significant eigenvectors in a Principal Component Analysis of the position coordinates, and $b_s = \left(b_{s1} \quad b_{s2} \quad \ldots \quad b_{st}\right)^T$ is a vector of weights for each variation mode of the modeled shape. Each eigenvector describes the way in which linearly correlated x_{ij} move together over the set, referred to as a mode of variation. The equation above represents the Point Distribution Model of a structure and can be used to generate new configurations of the shape modeled [37, 38].

The local grey-level behavior of each landmark point can also be considered in the modeling of a shape from images [37, 39]. Thus, statistical information is obtained about the mean and covariance of the grey values of the pixels around each landmark point. This information is used in the PDMs variations: to evaluate the match between landmark points in Active Shape Models and to construct the appearance models in Active Appearance Models, as explained afterwards. The active models can then be used to identify, i.e. to segment, the modeled shape in new images.

A) Active Shape Model

The combination of PDM and the grey level profiles at each landmark point of the shape to be modeled can be used to segment it in new images through the Active Shape Models, an iterative technique for fitting flexible models to shapes represented in images [37, 39].

The aforementioned technique is an iterative optimization scheme for PDMs allowing the refining in a new image of an initial estimated configuration of a modeled shape. The used approach can be summarized by the following steps: 1) at each landmark point of the model calculate the necessary movement to displace that point to a better position; 2) calculate the changes in the overall position, orientation and scale of the model which best satisfy the displacements; 3) finally, through calculating the required adjustments to the shape parameters, use residual differences to deform the model into the new configuration.

Active shape models can be implemented following a multiresolution approach, as it was done in Cootes et al. [40] by building a multiresolution pyramid of the input images, by applying a Gaussian mask, and then evaluate the grey level profiles on the various levels of the pyramid built, which originates faster models.

B) Active Appearance Model

This approach was first proposed in Cootes and Edwards [41] and allows the building of texture and appearance models of shapes from images. These models are generated by combining a model of shape variation (a geometric model) with a model of the appearance variations in a shape-normalized frame [37]. The statistical model of the shape that uses it is also described by Equation [1].

To build a statistical model of the grey level appearance, we deform each training image so that its landmark points match the mean configuration of the shape to be modeled by using a triangulation algorithm. Afterwards, we sample the grey level information, g_{im}, from the shape-normalized image over the region covered by the mean configuration. In order to minimize the effect of global light variation, we normalize this vector to obtain $_{g}$. Then, by applying a Principal Component Analysis to $_{g}$, we obtain a linear model called the texture model:

$$g = \bar{g} + P_g b_g , \tag{2}$$

where $_{\bar{g}}$ is the mean normalized grey level vector, $_{P_g}$ is a set of orthogonal modes of grey level variation and $_{b_g}$ is a set of grey level model parameters. Therefore, the geometrical configuration and appearance of any example of the modeled shape can be defined by the vectors $_{b_s}$ and $_{b_g}$.

Since correlation may exist between the geometrical and grey level variations, we apply a further Principal Component Analysis to the data. Thus, for each training shape we generate the concatenated vector:

$$b = \begin{pmatrix} W_s b_s \\ b_g \end{pmatrix} = \begin{pmatrix} W_s P_s^T \left(x - \bar{x} \right) \\ P_g^T \left(g - \bar{g} \right) \end{pmatrix}, \tag{3}$$

where W_s is a diagonal matrix of weights for each configuration parameter, allowing the adequate balance between the geometrical and the grey models. Then, we apply a Principal Component Analysis on these vectors, giving a further model:

$$b = Qc, \tag{4}$$

where Q are the eigenvectors of b and c is the vector of appearance parameters controlling both the geometrical configuration and the grey levels of the model. In this way, a new configuration of the modeled shape can be obtained for a given c by generating the shape-free grey level configuration, from the vector g, and be deformed it using the landmark points described by x.

2.2.3. Data and Assessment

A framework in MATLAB was developed to build statistical deformable models, namely PDMs and ASMs, which integrates the Active Shape Models software [42]. Additionally, for the appearance models the Modeling and Search Software [43] was used.

According to the International Phonetic Alphabet (IPA), the EP language consists in 28 sounds. In this work, 21 of these sounds were considered in the building of the statistical models of the vocal tract's shape by using a MR image for each one, where the sounds considered include the most representative sounds of the EP speech language. Additionally, 4 distinct MR images, related with other 4 EP speech sounds, were later used to evaluate the quality of the segmentation obtained by the Active Models previously built.

Due to the variability of the speech for each subject in addition to the considerable number of sounds studied (speech corpus), the set of 25 MR images was acquired from one young male subject in a similar manner to other works that use MRI to study the vocal tract during speech production. The training of the subject was performed to ensure the proper production of the intended EP speech sounds and to reduce speech subject variability. Moreover, the subject in question had a vast knowledge of EP speech therapy.

To enable the analysis on the sensibility of the Active Shape Models in terms of the percentage of the retained variance and on the dimensions of the profile adopted for the grey levels, ASMs were built with 95% and 99% of retained variance and with profiles for the grey levels of 7, 11 and 19 pixels. In the same way, Active Appearance Models were built with 95% and 99% of retained variance and considering 50000 and 10000 pixels for the texture model.

Following the construction of the vocal tract model, some sounds were chosen to be reconstructed by using the statistical deformable model built, namely four EP fricative consonants. Phonetically, the EP consonants are classified according to the obstruction of the vocal tract from the front of the mouth to the back; the EP vowels are regarded as being long and somewhat continuous sounds, classified from the front to the back of the mouth and from the higher to the lower tongue positions. EP consonants are classified according to the places where the articulators converge in order to obstruct the vocal tract – articulation points. In this

manner, it is relatively easy to identify the distinctive features of the sounds produced. As far as the EP fricative consonants are concerned, the articulation points are: labiodental (SAMPA) /f/, /v/, alveolar /s, z/ and post-alveolar /S, Z/. With regards to the production of the EP vowels /i, E/, the tongue moves to higher frontal positions and in the case of the EP vowels /o, u/, the tongue moves to more elevated backward positions. The EP sound /a/ is produced when the tongue is to be found in a central and mid-low position.

After building the Active Shape Models and Active Appearance Models from the training set of 21 MR images, they were used to segment the vocal tract's shape in 4 MR images that were not included in the training set. As stopping criterion of the segmentation process, a maximum of 5 iterations on each resolution level was considered. As 4 resolution levels were defined based on the dimensions of the images, this criterion means that, from the moment that the segmentation process starts to its end, a maximum of 20 iterations can occur. This maximum number of iterations was chosen because with the images considered it leads to excellent segmentation results. Additionally, it was verified that a lower value was not always sufficient to obtain satisfactory segmentations and a higher value constantly lead to the same segmentation results.

In order to assess the quality of the segmentations obtained in new MR images for the shape modeled by the Active Shape Models and Active Appearance Models built, the values of the mean and standard deviation of the Euclidean distances between the landmark points of the final configuration of the models and the desired segmentation shapes were calculated as well as the minimum and maximum values.

3. RESULTS

3.1. Modes of Variation

As aforementioned, the first task of the considered active models building process consists of the manual labeling of the MR images of the vocal tract. As expected, this task revealed itself difficult and extremely time consuming. However, the considerable noise presented in the MR images and the significant variability of the sounds under study make the use of automatic approaches difficult.

In Table 1, the first 15 modes of variation of the active shape model built and their retained percents are indicated. From the values presented one may conclude that the initial 7 modes, which correspond to 14% of the modes of variation, are capable of explaining 90% of all variance of the vocal tract's shape under study. Additionally, one may conclude that the first 10 modes, i.e. 20% of the modes of variation, represent 95% of all variance and that the first 15 modes, that is to say, 30% of the modes of variation, provide an explanation for 99% of all variance. This indicates that the ASM built is able to considerably reduce the data that is required to represent all shapes that the vocal tract assumes in the set of training images.

The effects of varying the first 6 modes of variation are depicted in Figure 3. This figure allows one to become aware that the first mode is associated with the movements of the tongue from the high front to the back positions at the oral cavity. With regards to the second mode of variation, it is possible to observe the vertical movement of the body of the tongue towards the palate. On one hand, the variations of the third mode are related with the opening

of the lips and tongue's backward movement. On the other hand, the fourth mode of variation reflects the tongue's tip movement from the central position of the tongue to the alveolar palate's ridge. Additionally, the fifth mode of variation translates the opening of the lips and the overall lateral enlargement of the vocal tract. Finally, the sixth mode is related with the movement of the tongue's body from back to front and down positions.

Table 1. Initial 15 modes of variation of the model built for the vocal tract's shape and their retained percentages

Mode of variation	Retained Percent	Cumulative Retained Percent
λ_1	45.349%	45.349%
λ_2	13.563%	58.912%
λ_3	9.672%	68.584%
λ_4	9.123%	77.707%
λ_5	6.716%	84.423%
λ_6	4.674%	89.097%
λ_7	2.262%	91.359%
λ_8	1.872%	93.231%
λ_9	1.442%	94.673%
λ_{10}	1.367%	96.040%
λ_{11}	0.978%	97.018%
λ_{12}	0.702%	97.720%
λ_{13}	0.507%	98.227%
λ_{14}	0.494%	98.721%
λ_{15}	0.397%	99.118%

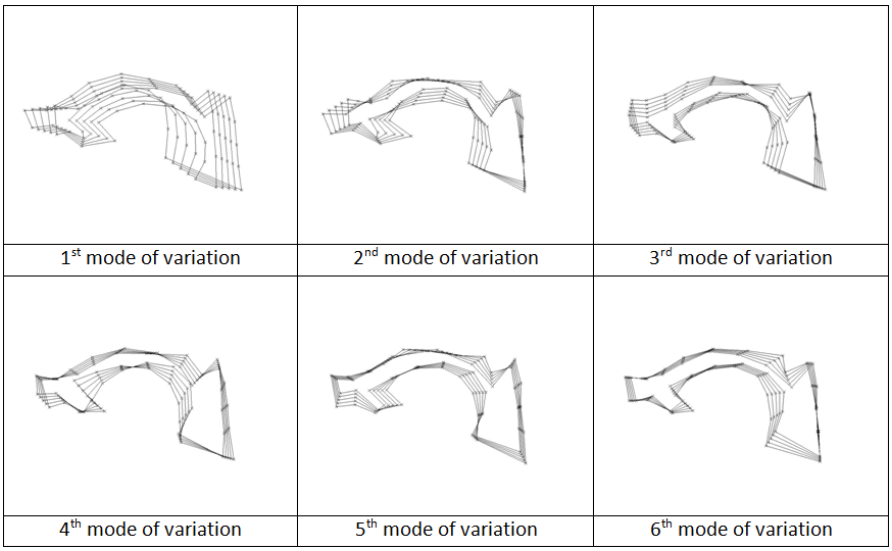

Figure 3. Effects of varying each of the first six modes of the model built for the vocal tract's shape $_{(\pm 2sd)}$.

3.2. Vocal Tract's Shape Reconstruction

Following the construction of the model for the vocal tract's shape, the previously selected sounds not belonging to the training set were reconstructed, i.e. simulated, by using the statistical deformable model built, namely the four EP fricative consonants.

The main goal of the present study in this phase was to conclude whether the modes of variation of the statistical deformable model built could be combined in order to successfully reconstruct, that is, reproduce, an EP speech sound. The sounds that were revealed to be the easiest to reconstruct were the consonants /f/ and /v/, as they only required the combination of two variation modes of the model built. Thus, in order to obtain the shape of the vocal tract when articulating the consonant /f/, it was necessary to merge the 1st and the 3rd modes. Similarly, the same modes, but with different weights were used to reconstruct the consonant /v/.

Through the union of four variation modes of the statistical deformable model built, it was possible to reconstruct the shape of the vocal tract when articulating the EP consonant /z/. Thus, the combination of the 1st, 2nd, 3rd and 6th modes permitted the reconstruction of the vocal tract's shape associated with the consonant /z/.

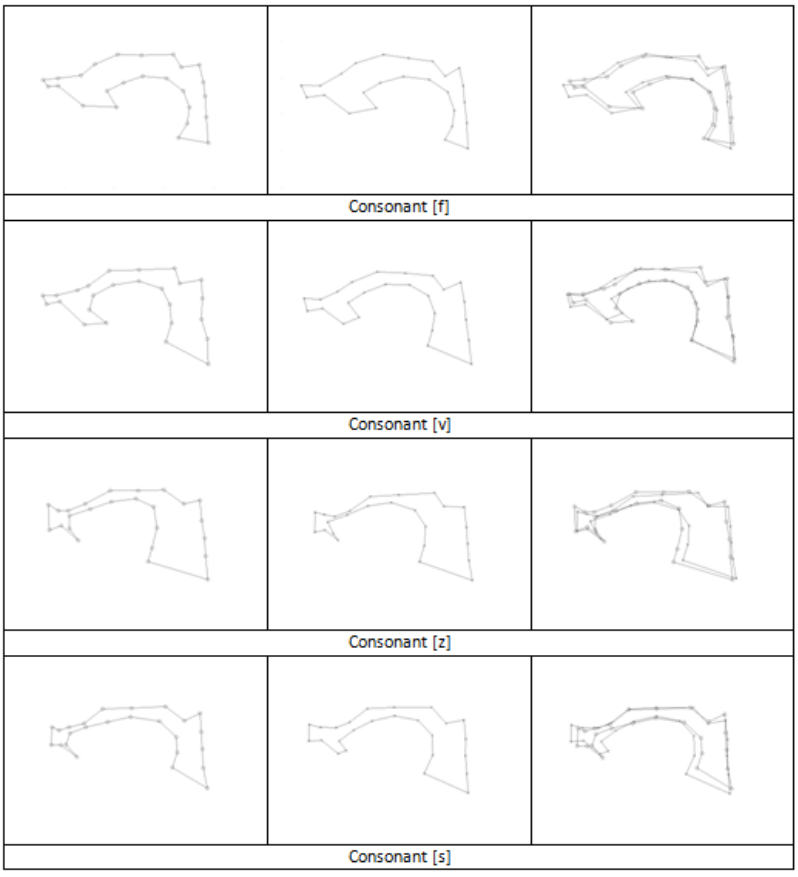

Figure 4. Reconstruction of the EP speech sounds /f/, /v/, /s/ and /z/: a) original shape, b) reconstructed shape and c) both shapes overlapped.

Table 2. Reconstructed shapes' errors

Phoneme	Minimum error [pixels]	Maximum error [pixels]	Mean error and standard deviation [pixels]
Consonant /f/	0.75	13.82	5.94 ± 3.00
Consonant /v/	1.29	15.10	6.10 ± 3.62
Consonant /z/	1.61	12.35	5.96 ± 2.75
Consonant /s/	1.26	18.27	6.72 ± 3.95

The fricative consonant more difficult to reconstruct turned to be the consonant /s/. In order to obtain the shape of the vocal tract when articulating the EP consonant /s/, it was necessary to bring together the 1st, 2nd, 3rd, 4th and 6th modes of variation of the statistical deformable model built. In Figure 4, the resultant reconstructions of the vocal tract's shape relating to the EP consonants /f/, /v/, /z/ and /s/ is depicted. In order to assess the quality of the reconstruction of the vocal tract's shape in the articulation of EP speech sounds, the minimum, maximum and mean errors and the standard deviation of the Euclidean distances between the landmark points of the original shape and that which is to be reconstructed must be calculated. Table 2 indicates these values for the reconstructions presented in Figure 4.

3.3. Vocal Tract's Shape Segmentation and Articulatory Assessment

Subsequently, 4 MR images of 4 distinct EP speech sounds, which were not considered in the set of training images used, were segmented by the active shape models built. In Figure 5, one of the segmentations obtained is depicted.

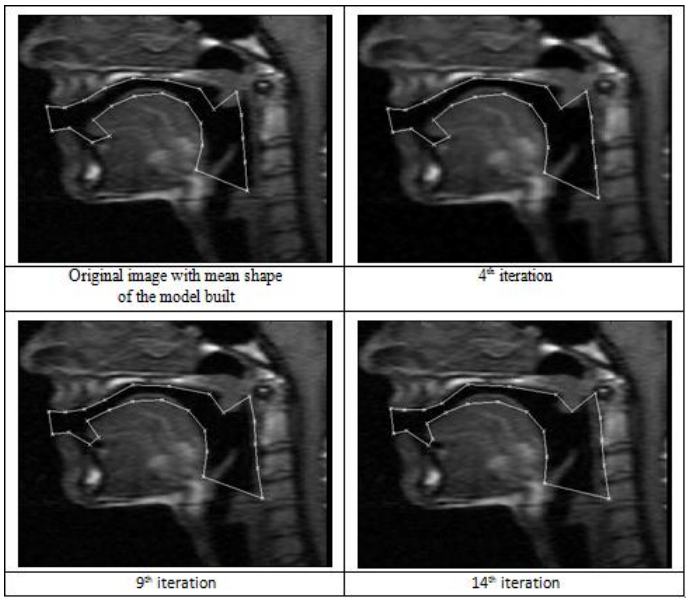

Figure 5. Testing image with the initial position of the overlapped mean shape of the model built and after 4, 9 and 14 iterations of the segmentation process through an active shape model built.

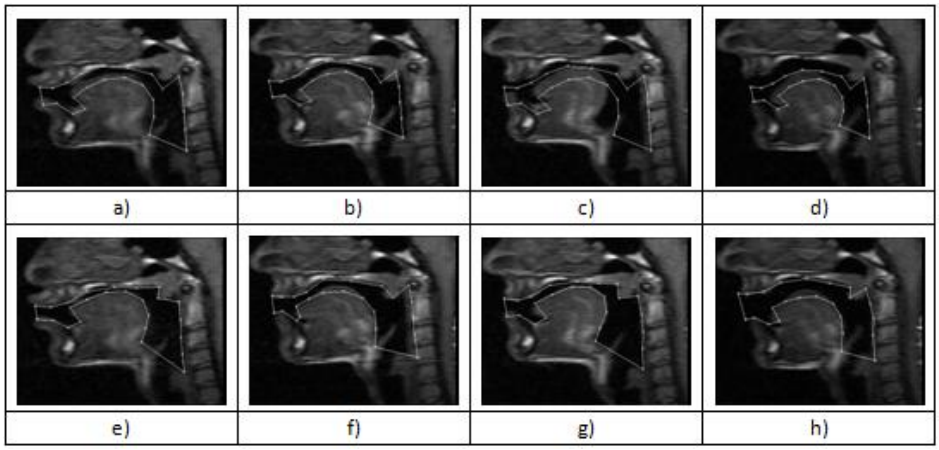

Figure 6. Testing images overlapped with the initial position of the shape model built (a-d) and the final results of the segmentation process through an active shape model built (e-h).

In Figure 5 it is possible to observe 4 of all the iterations of the segmentation process by the active shape model built: it starts with a raw estimation on the localization of the vocal tract in the image (1st iteration), downwards each multiresolution level (4th and 9th iterations represented) until convergence into the final the vocal tract's shape after 14 iterations. This segmentation was obtained considering an active shape model able to explain 95% of all variance of the vocal tract's shape under study and adopting a 7 pixels long grey level profile that is considering 3 pixels from each side of the landmark points. Likewise, the segmentation results using this model on the 4 testing MR images are shown in Figure 6.

In Table 3, the values of the mean and standard deviation that translate the quality of the segmentation obtained in each testing MR image by the active shape models built are presented. (For a better understand of the data indicated in this table, the models are named as: *Asm_varianceretained_pprofiledimension*).

Table 3. Average and standard deviation errors of the segmentations obtained from the testing images using the statistical models built

Model	Image 1	Image 2	Image 3	Image 4
Asm_95_p7	9.99 ± 5.76	9.89 ± 4.43	11.54 ± 6.36	14.23 ± 7.66
Asm_99_p7	9.97 ± 6.27	10.65 ± 3.45	fail	12.25 ± 5.86
Aam_95_5000	4.90 ± 2.42	10.21 ± 5.09	8.98 ± 4.80	9.91 ± 3.95
Aam_99_5000	6.77 ± 3.18	9.73 ± 4.56	8.80 ± 4.88	9.83 ± 4.48
Aam_95_10000	4.94 ± 2.45	10.19 ± 5.07	8.98 ± 4.78	10.56 ± 4.00
Aam_99_10000	4.35 ± 2.30	9.71 ± 4.60	8.80 ± 4.89	10.06 ± 4.58

As it was said earlier, active shape models with grey level profile of dimensions equal to 11 and 19 pixels were also built. However, these active shape models were not able to segment successfully the modeled shape in the testing images. This fail is due to the size of the images considered, which is relatively small: during the segmentation process, at each landmark point is considered a segment of 22 (or 38) pixels long in the active search and consequently the model built can easily diverge.

As already indicated, active appearance models can also segment shapes modeled in new images. By considering 95% of all the shape's variance and 10000 pixels in the building of the texture model, 9 modes of shape variation were extracted, that is 18% of the modes of variation, 17 texture modes, i.e. 34% of the modes of variation and 13 appearance modes, that is to say 26% of the modes of variation. Additionally, if an active appearance model is built using 99% of all the shape variance and considering the same number of pixels, then 15 shape modes (30%), 20 texture modes (40%) and 18 appearance modes (36%) are obtained.

The effects of varying the initial 3 modes of variation of texture and appearance of one of the active appearance models built are depicted in Figure 7. This figure allows one to become aware that the first mode is associated with tongue's movements from the high front to back positions. On the other hand, one can verify that the second mode of variation is related to the vertical movement of the tongue towards the palate. Finally, the third mode of variation seems to translate the lips' movement together with the tongue's movement to backward. It should be noticed that these modes of variation also contain information about the appearance, meaning that the intensity profiles associated with each configuration of the vocal tract are considered.

Figure 8 presents the segmentation result using one of the active appearance models built on a MR testing image. In this figure, it is possible to observe 4 of all the iterations of the active search needed to correctly segment the shape modeled: it starts with a raw estimation on the localization of the vocal tract's shape in the image (1st iteration), downwards each multiresolution level (7th and 12th iteration) until it converges into the desired vocal tract' shape after 20 iterations.

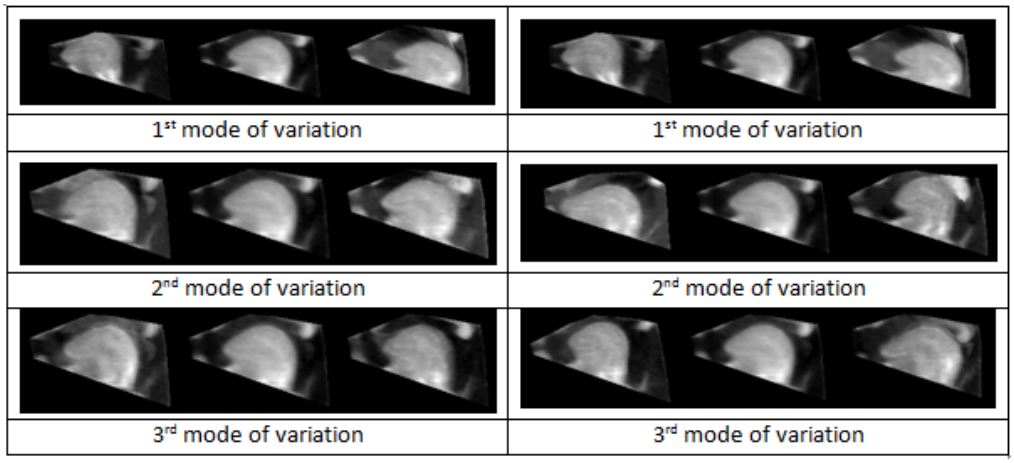

Figure 7. First three modes of the texture (left column) and appearance (right column) variation of an active appearance model built for the vocal tract's shape $(\pm 2sd)$.

Similarly, the segmentation results using the same model on all testing MR images are shown in Figure 9. The obtained values of the mean and standard deviation that translate the quality of the segmentation obtained in each testing MR image by the active appearance models built are presented in Table 3. (Again, for a better understanding of the data indicated, the models are named as: *Aam_varianceretained_npixelsused*).

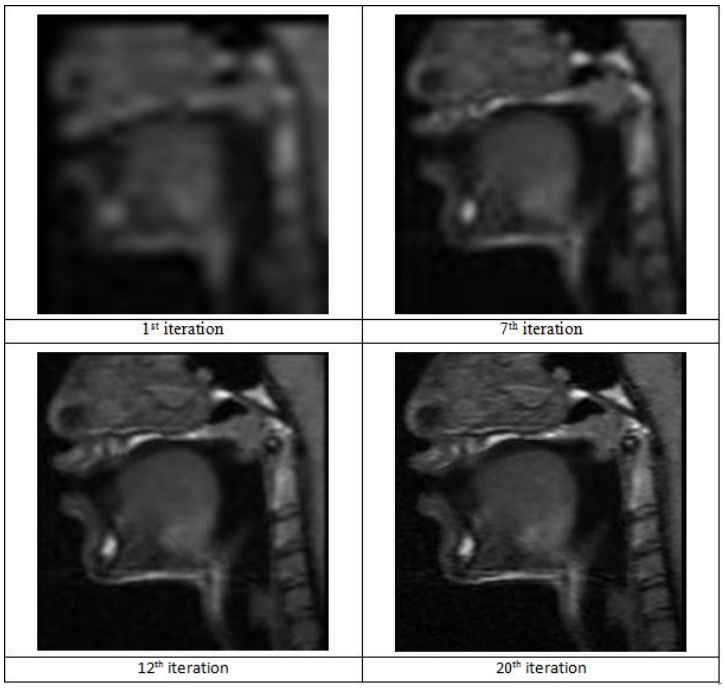

Figure 8. Results after the 1st, 7th, 12th and 20th iterations of the segmentation process using an active appearance model built for the vocal tract's shape.

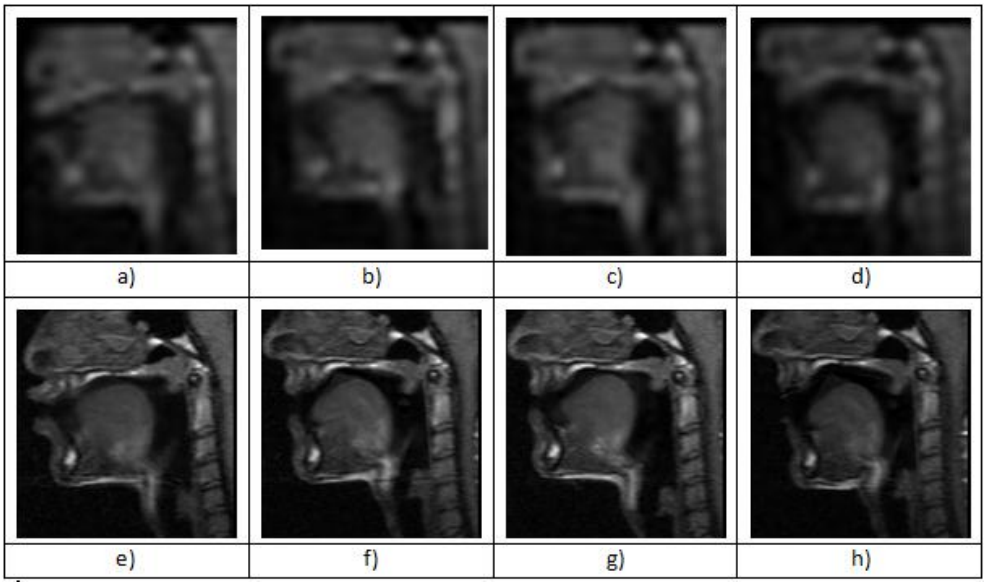

Figure 9. Testing images overlapped with the initial position of the mean shape model built (a-d) and the final results of the segmentation process obtained through an active appearance model (e-h).

Through an analysis of the data presented in the Table 3, one may conclude that the active appearance models obtain superior results when compared with the active shape models. Furthermore, the use of more modes of variation leads to better results when the

active appearance models are used, in contrast with the segmentation results obtained by using the active shape models, where the use of more modes of variation (retained percent) not always translates into improved results.

CONCLUSION

In this work, Deformable Models were applied in magnetic resonance images to study the shape of the vocal tract in the articulation of some European Portuguese sounds and further use the models built to segment the vocal tract's shape in new images.

The use of MR images permits the study of the whole vocal tract with enough safety, high quality imaging of soft-tissues and is also a non-invasive method, having as a drawback the amount of noise that is usually presented in the acquired images. However, the built models could be able to conveniently overcome the existent noise and obtain very good segmentation results.

From the experimental findings, it can be concluded that the statistical deformable models built are capable of efficiently characterize the behavior of the vocal tract modeled from the studied MR images. In fact, the modes of variation of the built model could provide an explanation of the actual actions involved in the EP speech sounds considered, such as: the movement of the tongue in the oral cavity, the lip's movements, or the approximation of the tip of the tongue to the alveolar region.

Additionally, it has been verified that the modeling performed could reduce the data set needed to characterize all variations of the vocal tract's shape during the production of the EP speech sounds, as 99 percent of these are explained by just 32 percent of the total of all the modes of variation.

The built statistical deformable models have also revealed to be good options to reconstruct proficiently the vocal tract's shape in the articulation of their speech sounds.

As future work, it is our intention to improve the built models of the vocal tract in order to obtain more geometric accuracy during EP speech production. This knowledge will allow a superior modeling of the vocal tract for speech synthesis and a better understanding of the speech production mechanisms for clinical purposes, namely for the rehabilitation of speech disorders.

The progresses in MRI acquisitions and the wider application of 3.0 Tesla scanners are promising advances expected for the imaging and analysis of the vocal tract during speech production that are going to be consider by us in future works regarding the EP language.

ACKNOWLEDGMENTS

This work was partially done under the scope of the following research projects "Methodologies to Analyze Organs from Complex Medical Images – Applications to the Female Pelvic Cavity", "Cardiovascular Imaging Modeling and Simulation - SIMCARD" and "Aberrant Crypt Foci and Human Colorectal Polyps: Mathematical Modelling and Endoscopic Image Processing", with the references PTDC/EEA-CRO/103320/2008,

UTAustin/CA/0047/2008 and UTAustin/MAT/0009/2008, respectively, financially supported by FCT - Fundação para a Ciência e a Tecnologia in Portugal.

The first author would like to acknowledge the support of the PhD grant SFRH/PROTEC/49517/2009 from the Escola Superior de Tecnologia da Saúde do Porto (ESTSP) and Instituto Politécnico do Porto (IPP). The second author would like to express her gratitude for the PhD grant SFRH/BD/28817/2006 from FCT.

The images considered in this work were acquired at the Radiology Department of the Hospital S. João, Porto, with the collaboration of the technical staff, which is gratefully acknowledged.

REFERENCES

[1] Rokkaku, M., Hashimoto, K., Imaizumi, S., Niimi, S., and Kiritani, S. (1986). Measurements of the Three-dimensional Shape of the Vocal Tract Based on the Magnetic Resonance Imaging Technique. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, 20, 47-54.

[2] Baer, T., Gore, J.C., Boyce, S., and Nye, P.W. (1987). Application of MRI to the analysis of speech production. *Magnetic Resonance Imaging*, 5, 1-7.

[3] Demolin D., Metens T., Soquet A. (1996). Three-dimensional Measurement of the Vocal Tract by MRI. Proc. 4th International Conference on Spoken Language Processing (ICSLP 96), Philadelphia, USA, 272-275.

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

[5] Behrends, J., and Wismuller, A. (2001). A Segmentation and Analysis Method for MRI data of the Human Vocal Tract. Proceedings of the Symposium on Human and Machine Perception in Acoustic and Visual Communication, Tutzing, Germany, 179-189.

[6] Engwall, O. (2000). A 3D tongue model based on MRI data. 6th Int. Conf. on Spoken Language Processing (ICSLP), China, 901-904.

[7] Takemoto, H., Honda, K., Masaki, S., Shimada, Y., and Fujimoto, I. (2003). Measurement of Temporal Changes in Vocal Tract Area Function during a continuous vowel sequence using a 3D Cine-MRI Technique. *Proceedings of the 6th International Seminar on Speech Production*, Sydney, Australia, 284-289.

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

[9] Rua, S., and Freitas, D. (2006). Morphological Dynamic Imaging of Human Vocal Tract. Proceedings of the International Symposium (CompIMAGE 2006). Porto, Portugal: Taylor and Francis.

[10] Teixeira, A., and Vaz, F. (2001). European Portuguese Nasal Vowels: An EMMA Study. 7th European Conference on Speech Communication and Technology (EuroSpeech). Scandinavia, 1843-1846.

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

[12] Serrurier, A. and 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.

[13] 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.

[14] 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, ISSN, 288-291.

[15] Kane, A.A., Butman, J.A., Mullick, R., Skopec, M., and Choyke, P. (2002). A new method for the study of velopharyngeal function using gated magnetic resonance imaging. *Plastic and Reconstructive Surgery*, 109(2), 472-481.

[16] Ventura, S.R., Freitas, D.R., and Tavares, J.M. (2010). Towards Dynamic Magnetic Resonance Imaging of the Vocal Tract during Speech Production. Journal of Voice, ISSN: 0892-1997, DOI: 10.1016/j.jvoice.2010.01.014 (in press).

[17] Demolin, D., Metens, T., and Soquet, A. (2000). Real Time MRI and Articulatory Coordinations in Vowels. Proc. 5 th Speech Production Seminar. München, Germany, 5-8.

[18] Mády, K., Sader, R., Zimmermann, A., Hoole, P., Beer, A., Zeilhofer, H., and Hannig, C. (2002). Assessment of Consonant Articulation in Glossectomee Speech by Dynamic MRI. Proceedings of 7th International Conference on Spoken Language Processing (ICSLP), Denver, USA, 961-964.

[19] Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. (2004). An Approach to Real-time Magnetic Resonance Imaging for Speech Production. *Journal Acoustical Society of America*, 115(4), 1771-76.

[20] Masaki, S., Nota, Y., Takano, S. Takemoto, H., Kitamura, T., and Honda, K. (2008). Integrated magnetic resonance imaging methods for speech science and technology. *Proceedings of Acoustics*, France, Paris, 5083-5088.

[21] Ventura, S.R., Freitas, D.R., and Tavares, J.M. (2009). Application of MRI and Biomedical Engineering in Speech Production Study. *Computer Methods in Biomechanics and Biomedical Engineering*, 12(6), 671-681.

[22] Engwall, O. (2003). A revisit to the Application of MRI to the Analysis of Speech Production - Testing our assumptions. *Proceedings of the 6th International Seminar on Speech Production*, Sydney, Australia, 43-48.

[23] Honda, K. (2002). Evolution of vowel production studies and observation techniques. Acoustical Science and Technology, *23*(4), 189-194.

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

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

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

[27] Kim, Y., Narayanan, S.S., and Nayak, K.S. (2009). Accelerated Three-Dimensional Upper Airway MRI Using Compressed Sensing. Magnetic Resonance in Medicine, 61, 1434-1440.

[28] Kass, M., Witkin, A., and Terzopoulos, D. (1987). Snakes: Active Contour Models. *International Journal of Computer Vision*, 1, 321-331.

[29] Yuille, A.L., Cohen, D., and Hallinan, P. (1992). Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8, 104-109.

[30] Gonçalves, P.C.T., Tavares, J.M.R.S., and Jorge, R.M.N. (2008). Segmentation and Simulation of Objects Represented in Images using Physical Principles. Computer Modeling in Engineering and Sciences. *Computer Modeling in Engineering and Sciences*, 32 (1), 45-55.

[31] Cootes, T.F., Taylor, C.J., Cooper, D.H. and Graham, J., (1992). Training Models of Shape from Sets of Examples. *Proceedings of the British Machine Vision Conference*, Leeds, 9-18.

[32] Harshman, R.A., Ladefoged, P., and Golstein, L. (1997). Factor analysis of tongue shapes. Journal of the Acoustical Society of America, 62, 693-707.

[33] Shirai, K. and Honda, M. (1978). Estimation of articulatory motion by a model matching method. *Journal of the Acoustical Society of America*, 64(S1), S42-S42.

[34] Maeda, S. (1988). Improved articulatory models. *Journal of the Acoustical Society of America*, 84(S1), S146-S146.

[35] Stone, M., Cheng, Y., and Lundberg, A. (1997). Using principal component analysis of tongue surface shapes to distinguish among vowels and speakers. *Journal of the Acoustical Society of America*, 101(5), 3176-3177.

[36] Oliveira, F.P.M. and Tavares, J.M.R.S. (2008). Algorithm of Dynamic Programming for Optimization of the Global Matching between Two Contours Defined by Ordered Points. *Computer Modeling in Engineering and Sciences*, 31 (1), 1-11.

[37] Vasconcelos, M.J.M., and Tavares, J.M.R.S. (2008). Methods to Automatically Built Point Distribution Models for Objects like Hand Palms and Faces Represented in Images. *Computer Modeling in Engineering and Sciences*, 36(3), 213-241.

[38] Vasconcelos, M.J.M., Ventura, S.M.R., Freitas, D.R.S., and Tavares, J.M.R.S. (2010). Using Statistical Deformable Models to Reconstruct Vocal Tract Shape from Magnetic Resonance Images. Proceedings of the Institution of Mechanical Engineers, Part H: *Journal of Engineering in Medicine*, 224, 1153-1163.

[39] Cootes, T.F., and Taylor, C.J. (1993). Active Shape Model Search using Local Grey-Level Models: A Quantitative Evaluation. in British Machine Vision Conference, Guildford: BMVA Press.

[40] Cootes, T.F., Taylor, C.J., and Lanitis, A. (1994). Active Shape Models: Evaluation of a Multi-Resolution Method for Improving Image Search. in British Machine Vision Conference, York, England: BMVA.

[41] Cootes, T.F., and Edwards, G. (1998). Active Appearance Models. *European Conference on Computer Vision*, Freiburg, Germany.

[42] Hamarneh, G. *ASM (MATLAB)*. 1999; Available from: http://www.cs.sfu.ca/~hamarneh/software/code/asm.zip.

[43] Cootes, T.F. *Build_aam*. 2004; Available from: http://www.wiau.man.ac.uk/~bim/software/am_tools_doc/download_win.html.