Inter-speaker speech variability assessment using statistical deformable models from 3.0 Tesla magnetic resonance images

Maria João M. Vasconcelos
Faculty of Engineering, University of Porto /
Institute of Mechanical Engineering and Industrial Management
Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
e-mail: maria.vasconcelos@fe.up.pt

Sandra M. Rua Ventura
Radiology Department, School of Allied Health Science –
Porto Polytechnic Institute / Faculty of Engineering, University of Porto
R. Valente Perfeito 322, 4400-330 Vila Nova de Gaia, Portugal
e-mail: smr@estsp.ipp.pt

Diamantino Rui S. Freitas
Department of Electrical Engineering and Computers, Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal
e-mail: dfreitas@fe.up.pt

João Manuel R. S. Tavares
Department of Mechanical Engineering, Faculty of Engineering, University of Porto /
Institute of Mechanical Engineering and Industrial Management
Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
e-mail: tavares@fe.up.pt

Corresponding author:

João Manuel R. S. Tavares
Department of Mechanical Engineering
Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
Phone: +351 22 508 1487, fax: +351 22 508 1445
e-mail: tavares@fe.up.pt, url: www.fe.up.pt/~tavares

Inter-speaker speech variability assessment using statistical deformable models from 3.0 Tesla magnetic resonance images

Abstract: The morphological and dynamic characterization of the vocal tract during speech production has been gaining greater attention due to the motivation of the latest improvements in Magnetic Resonance (MR) imaging; namely, with the use of higher magnetic fields, such as 3.0 Tesla. In this work, the automatic study of the vocal tract from 3.0 Tesla MR images was assessed through the application of statistical deformable models. Therefore, the primary goal focused on the analysis of the shape of the vocal tract during the articulation of European Portuguese sounds, followed by the evaluation of the results concerning the automatic segmentation, i.e. identification of the vocal tract in new MR images. In what concerns speech production, this is the first attempt to automatically characterize and reconstruct the vocal tract shape of 3.0 Tesla MR images by using deformable models; particularly, by using active and appearance shape models. The achieved results clearly evidence the adequacy and advantage of the automatically analysis of the 3.0 Tesla MR images of these deformable models in order to extract the vocal tract shape and assess the involved articulatory movements. These achievements are mostly required, for example, for a better knowledge on speech production, mainly of patients suffering from articulatory disorders, and to build enhanced speech synthesizer models.

Keywords*:* Medical Imaging; Image Analysis; Modelling and Segmentation; Active and Appearance Shape Models; Portuguese Speech Language; Vocal Tract Shape; Morphological Study.

1. Introduction

Since the first applications of Magnetic Resonance (MR) imaging to speech production assessment, in the 1980s [1, 2], many studies have been performed, namely for French [3], English [4], Swedish [5], Japanese [6], and European Portuguese (EP) [7-9] languages.

Due to the lengthy data acquisition time of the early MR imaging systems, the first studies were restricted to vowels and some consonants [10, 11]. However, with the emerging development of rapid imaging techniques, such as synchronized sampling methods [12] or tagged cine-MR [13, 14], the acquisition of image data regarding articulatory movements became possible. Nowadays, the acquisition of three-dimensional (3D) MR image sequences has been steady [15] and, consequently, enormous expectations have been made about the attainment of image data on speech production in a more efficient and repeatable manner.

Various organs have important roles in the production of numerous speech sounds, functioning in an organized, i.e. articulated, manner in order to change the shape and length of a set of air cavities - the vocal tract. Most of these organs, named articulators, are soft-tissues that execute active movements during the speech production, such as the lips, tongue and velum. The tongue is a muscular organ capable of moving in nearly every direction, expanding, compressing and displaying a fine degree of articulation in the oral cavity.

Nonetheless, individual differences in the vocal tract morphology turn speech production into a unique motor activity. Inter-speaker variability of the acoustic speech signal can confound the process of any movement data to evaluate theories of speech movement control [16]. Thus, this challenges the capability of Magnetic Resonance Imaging for the morphological description of the acoustical inter-speaker variability [17]. With the cutting-edge MR improvements, a proper 3D description on the vocal tract geometry of the speakers can be reached, both in terms of good image contrast and temporal resolution. In addition, useful and accurate morphological and dynamic information can also be attained, as to the positions and shapes of the involved articulators during speech production [9, 18-20].

The use of deformable models in image analysis has been generating remarkable results in innumerable and distinct applications [8, 21, 22]. Active contours, deformable templates, physical models and statistical models can be considered as the most well-known deformable models to extract object features from input images [23]. Active contours were introduced in [24], by considering the segmentation contour as a “snake”. Hence, the segmentation contour consists of an elastic set of points that are adjusted to the border of the object to be segmented, driven by the combination of internal and external forces, in order to minimize the energy of the model. On the other hand, deformable geometrical shapes (templates), built considering the shape of the object to be segmented, are parameterized by appropriated functions in order to segment the modelled object in new images based on its characteristic image features [25]. Recently, enhanced physical modelling approaches integrate the previously acquired knowledge about the objects, making the models used in the image segmentation process more realistic [26]. Finally, statistical models, particularly Point Distribution Models (PDMs) are built from a set of training shapes of the object under study in order to extract its main characteristics through statistical modelling [21, 27]. Then, the PDMs can be used to segment the modelled object in new images by considering the image intensity information, resulting in the Active Shape Models (ASMs) [21, 28], or by considering the image texture information, following in the Active Appearance Models (AAMs) [19, 21, 29].

In this work, deformable models, in particular, PDMs, ASMs and AAMs, were applied in the automatic study of the vocal tract from 3.0 Tesla Magnetic Resonance Images; mainly, to evaluate the shape of the vocal tract during the articulation of European Portuguese sounds and later to automatically segment the vocal tract in new images.

This paper has been organized as follows: First, the adopted MR imaging protocol is described. Then, the focus will be on PDMs, ASMs and AAMs, alongside the data used and the assessment adopted regarding the segmentation quality. Afterwards, the models built and their application in the segmentation of the vocal tract in new images representing EP speech sounds will be presented and discussed. The paper ends by pointing out the main conclusions.

2. Methods

In this section, the methodologies adopted to characterize the vocal tract during the production of EP speech sounds of 3.0 Tesla MR images are described. Thus, the used MR imaging protocol as well the procedures adopted in the image acquisition process are indicated. Afterwards, an explanation regarding the modelling of objects in images with PDMs is provided, as well as the building process of ASMs and AAMs that were employed to segment the shape of the vocal tract in new images. Finally, the data set used and the assessment addressed are presented.

2.1. Magnetic Resonance Imaging Protocol and Procedures

According to the safety procedures for MR, a questionnaire was performed for screening patients before any procedure. In addition, patients were previously informed and instructed about the study to be performed and informed consents were obtained.

The image data was acquired using a MAGNETOM Trio 3.0 Tesla MR system and two integrated coils (a 32-channel head coil and a 4-channel neck matrix coil), with the subjects in supine position. The two young volunteers (one male and one female) were trained before the MR exam to ensure the proper production of the intended sounds. The speech corpus consisted of 25 sounds of European Portuguese language, including oral and nasal vowels, and consonants.

Using turbo spin echo 2D sequence, and adopting the following parameters: 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 cm, a matrix size of 512x512 pixels, a resolution of 2.133 pixels per mm and a 0.469x0.469 pixel size, 1 T1-weighted midsagittal slice of 3 mm thickness was acquired for each sound. In order to reduce intra-speaker variability and to ensure consistency of results, 3 measurements (i.e. 3 slices per sound) were performed during the sustained sound with an overall acquisition time of approximately 8.07 seconds, resulting in 75 images for each subject.

Examples of the MR images acquired are depicted in Figure 1. From these images, one may observe different vocal tract configurations for EP vowels and consonant production, as well as for some oral and nasal sounds. Comparing the several vocal tract configurations of the subjects during the articulation of the EP sounds, individual differences of vertical length and of organs morphology were revealed, although the main movements were similar.

2.2. Vocal Tract Modelling

Some of the images acquired according to the imaging protocol described in the previous section were used to statistically model the shape of the vocal tract and the remainder to evaluate the models built. Hence, in the following sections, the process adopted to label the shape of the vocal tract, i.e. to define the landmark points to be addressed by the model, is described, and then the statistical modelling techniques are introduced.

2.2.1. Shape Landmark Points

In the building process of a PDM, each shape of the vocal tract that is presented in the image training set should be described by a group of labelled landmark points conveying important anatomical aspects of the structure, Figure 2. (In the current and subsequent images, the landmark points appear connected by fictitious line segments so as to enhance their visualization.) Consequently, the manual identification of these points in all training images requires a comprehensive knowledge of the structure in question, as the resultant model behaviour greatly depends on the landmark points selected.

The manual selection of the landmark points was carried out by one of the authors who has excellent knowledge on MR imaging and on the anatomy of the vocal tract, in addition to being cross-checked by another co-author in accordance with the following criteria:

- 4 points in the lips (front and back of the lips' margins);
- 3 points corresponding to the lingual *frenulum* and tongue's tip;
- 7 points equally spaced along the surface of the tongue;
- 7 points along the surface of the hard palate (roof of the oral cavity) placed in symmetry with the tongue points;
- 1 point at the velum (or soft palate);
- 3 points equally spaced at the posterior margin of the oropharynx (behind the oral cavity).

It should be noted that, during this task, the epiglottis was not taken into account.

Thus, in each of the 150 acquired MR images, 25 landmark points were defined according to these criteria.

2.2.2. Statistical Modelling

In order to study the admissible variation of the coordinates of the landmark points of the training shapes, it is initially necessary to align them by using, for instance, dynamic programming [30]. Therefore, given the co-ordinates $\left(x_{ij}, y_{ij}\right)$ of each landmark point j of the shape i of the modelled structure, 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 in the image training set and n the number of landmark points used. Once the training shapes are aligned, the mean shape and the admissible variability of the modelled structure may be found. The modes of variation characterize the manner in which the landmarks of the modelled structure tend to move together, the result of which may be obtained by applying a Principal Component Analysis (PCA) to the deviations from the mean. Hence, it is possible to rewrite each vector x_i as:

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

where x_i represents the coordinates of the n landmark points of the new shape of the modelled structure, $\left(x_k, y_k\right)$ are the coordinates of the landmark point k, $\overline{x}$ is the mean position of all 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} corresponds to the most significant eigenvectors in a PCA applied to the coordinates of all landmark points, 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 modelled structure. Each eigenvector describes the manner in which linearly correlated x_i move together over the training set, and due to this, is commonly known as a mode of variation. Thus, equation (1) represents the PDM of the modelled structure and may be used to generate new shapes that it can undertake. Further details about the construction of PDMs can be found in [21].

The local grey-level environment of each landmark point may also be considered in the statistical modelling of objects from images [21, 28]. Thus, statistical information is obtained in relation to the mean and covariance of the grey values of the image pixels around each landmark point. Additionally, this information can be used to evaluate the matching between landmark points, resulting into the ASMs, in addition to considering the information on image texture ensuing to the AAMs, as explained in the following.

a) Active Shape Models

The consideration of a PDM in addition to the grey level profiles of each landmark point used in its building can be used to segment the modelled structure in new images through Active Shape Models, which are based on an iterative technique for fitting flexible models to structures represented in images [21, 27]. Hence, this technique is an iterative optimisation scheme that refines the mean shape, $\overline{x}$, given by the PDM built for the structure under study, according to associated modes of variation, in a new image, i.e. this refining process segments the modelled structure in the new image. The refining process adopted may be summarized by the following steps: 1) The displacement required to dislocate the model to a more appropriate position, that is, closer to the final shape, is calculated at each landmark point; 2) The calculus of the changes in the overall shape position, orientation and scale that most adequately satisfy the local displacements found in 1); 3) The obtainment of the required adjustments in the parameters of the model, by analysing the residual differences between the shape of the model and the final desired shape.

The image segmentation process with the aid of Active Shape Models was improved in [31] due to the adoption of a multiresolution approach that may be summarised as follows: First, a multiresolution pyramid of the input images is built by applying a Gaussian mask; following this, the grey level profiles at the various levels of the pyramid built are studied. Consequently, the ASMs are capable of segmenting the input images in a more efficient and trustworthy manner.

b) Active Appearance Models

The segmentation of structure in images by AAMs was initially proposed in [29], based on the building of texture and appearance models for the structures to be segmented. These models are generated by combining a shape variation model, i.e. a geometric model, with an appearance variation model in a shape-normalised framework [19, 21].

The geometric models integrated into the AAMS are the PDMs described by Equation **Error! Reference source not found.**. Conversely, to build the statistical models of the grey level appearances of the structures represented in the training images, one needs to deform each training image in order that the landmark points match the mean shape of the structure to be modelled, using a triangulation algorithm. Then, the grey level information, i.e. the intensity values, g_{im}, from the shape-normalised image over the region covered by the mean shape is sampled. In order to minimize the effect of global intensity variation in the training images, the average vector of the grey levels (g_{im}) is once again normalized, thereby resulting in vector g. Following the application of a PCA to the previous vector g, a new linear model, called the texture model, is obtained:

$$g = \bar{g} + P_g b_g , \tag{1}$$

where $\bar{g}$ is the mean normalised grey level vector, P_g is a set of orthogonal modes related to the grey level variations and b_g is a set parameters of the grey levels model. Therefore, the shape and appearance of any configuration of the modelled structure can be defined by vectors b_s and b_g.

Given that a correlation may exist between the variations of shape and of grey levels, a further PCA is applied to the data of the structure. Thus, for each training image a concatenated vector is generated:

$$b = \begin{pmatrix} W_s b_s \\ b_g \end{pmatrix} = \begin{pmatrix} W_s P_s^T \left(x - \bar{x} \right) \\ P_g^T \left(g - \bar{g} \right) \end{pmatrix}, \tag{2}$$

where W_s is a diagonal matrix of weights for each parameter of the global model built, allowing for the adequate balance between the models of shape and grey levels. Next, a PCA is applied to these vectors, which results in a novel model:

$$b = Qc , \tag{3}$$

where Q is the eigenvectors of b and c is the vector of the appearance parameters that control the shape in addition to the grey levels of the model built. In this manner, a new shape of the modelled structure can be obtained for a given vector c by generating the shape-free grey level structure from vector g and then deforming it by considering the landmark points provided by x.

2.2.3. Data set and Assessment

A computational framework was developed in MATLAB to build statistical deformable models, namely PDMs and ASMs, which integrates the Active Shape Models software [32]. Additionally, the Modelling and Search Software [33] was used to build AAMs.

According to the International Phonetic Alphabet (IPA), the EP speech language consists of a total of 30 sounds. In this work, 25 of these sounds have been considered in the building of the statistical models of the vocal tract by using 3 3.0 Tesla MR images per each one. The considered sounds include the most representative sounds of the EP speech language. Additionally, 2 new 3.0 Tesla MR images for the EP speech sounds /v/, /f/ and /a/ for each subject (making a total of 12 new images) were used to evaluate the quality of the segmentations obtained by the models built. These sounds were selected since: the associated sounds are easy to sustain, require slight efforts to the subjects and ensure the steadiness of the vocal tract shape; and include the two classes of sounds under study (two fricative consonants, one voiced and another one voiceless, respectively, and one vowel (/a/)).

In order to analyse the sensibility of the Active Shape Models in terms of the percentage of retained variance and of the dimensions of the profile adopted for the grey levels in the modelling, ASMs were built adopting 90%, 95% and 99% of retained variance and profiles of 7, 11 and 15 pixels [20]. Similarly, Active Appearance Models were built adopting equal values of retained variance and the following values of 5000 and 10000 pixels were considered for building the texture model [19]. These parameters were defined based on the authors' previous experience concerning the statistically modelling of vocal tract using these models [9, 19-21].

Following the building of the ASMs and AAMs from the training set constituted by 138 images, the models were then used to segment the vocal tract in 12 new images. As a stopping criterion of the segmentation process, a maximum of 6 iterations on each resolution level was taken into consideration. Due to the fact that 5 resolution levels were defined based on the dimensions of the images under study, this criterion means that from the beginning of the segmentation process to its end, a maximum of 30 iterations could occur [21]. This maximum number of iterations was chosen as a result of the fact that in the experiments done it led to excellent segmentation results. In fact, it was observed that an inferior value was not always sufficient to attain satisfactory results and a superior value constantly caused similar results.

In order to assess the quality of the segmentations obtained for the vocal tract in new MR images by the models built, the values of the mean and standard deviation of the Euclidean distances between the landmark points of the final shape computationally obtained and the correspondent ones manually defined in the same images were calculated.

3. Results and Discussion

From Table 1, one may observe that the initial 11 modes of variation of the Active Shape Model built, that is, 22% of the modes of variation, are capable of explaining 90% of all variance of the vocal tract. Moreover, one may conclude that the first 17 modes, i.e. 34% of the modes of variation, provide an explanation for 95% of all variance and the initial 33 modes, which means that 66% of the modes of variation illustrates 99% of all variance. Consequently, these findings clearly indicate the ability of the built ASM to considerably condense the data that is required to represent all configurations that the vocal tract assumes in the image training set.

If one takes into consideration the first 7 modes of variation, it is possible to observe that a wide range of movements, including wide range ones to more refined and particular movements of the articulators, had been successfully addressed. The effects on varying the first 6 modes of variation of the built models are depicted in Figure 3. From this figure, one can realize that the first mode is related to the movements of the tongue from the front to the back in the oral cavity associated with the rise of the larynx. With regard to the second mode of variation, it is possible to observe the movements of the tongue from the front-high to the back-down in the oral cavity associated with the lips opening and narrowing. The third mode of variation describes the velum’s lowering associated with the enlargement/narrowing of the pharynx cavity and the tongue’s tip movement. The vertical movement of the body of the tongue towards the palate is revealed by the fifth mode of variation. In contrast, the variations of the sixth mode illustrate the open/close of the lips associated with the vertical movement of the tongue. After this mode of variation, all the remainder modes represent more particular movements, such as the larynx height adjustment, the tongue’s tip movement, the opening and closing of the lips, the vertical rise of the tongue’s body towards the palate and the pharynx narrowing.

After the analysis on the ability of the built statistical models to render the real behaviour of the vocal tract during the production of EP language sounds, 12 new MR images of the 3 distinct EP speech sounds previously selected (/f/, /v/ and /a/), i.e. of images not included in the used training image set, were automatically segmented by the same models. In Figure 4, one MR image of each subject articulating the EP speech sound /f/ is presented as well as the evolution of the correspondent segmentation by the active shape model built: the segmentation begins with a rough estimate for the vocal tract in the input image and then deforms it towards the desired segmentation. These results were obtained considering an ASM capable of explaining 90% of all variance of the vocal tract under study and adopting a grey level profile length of 11 pixels, that is by considering 5 pixels from each side of the landmark points [21]. Analogously, the segmentation results obtained by using this model on other 4 new MR images are presented in Figure 5, where the first two images concerns one subject and the last image concerns the other subject articulating the EP speech sounds/v/ and /a/, respectively.

In Table 2, the values of the mean and standard deviation that reflect the quality of the segmentations obtained by the built active shape models in each testing MR image are indicated. (For a more comprehensive understanding of the data included in this table, the models are named as *Asm_varianceretained_profiledimension* and cases of segmentation failures are indicated by a dash.) The results concerning the built ASMs considering 99% of all variance were not included in this table, since the models were not able to successfully segment the modelled organ in most of the testing images. This failure is precisely due to the percentage of retained variance used, 99%, which led to an extremely rigid model and, because of that, with a very low ability to be adapted to new configurations.

As aforementioned, active appearance models are also proficient in modelling objects in images and to segment the modelled objects into new images. Texture and appearance modes of variation are more difficult to analyse because some motion artefacts ("blur effect") are presented as a result of some inconsistencies of the female subject to sustain the sound, and also because of the inter-subjects differences of vocal tract morphologies. The effects of varying the initial 3 modes of variation in terms of texture and appearance of one of the active appearance models built are depicted in Figure 6. In this figure, it is possible to observe a few slight movements, which are mostly related to the tongue. The first mode of texture depicts the movement of the lower lips and tongue’s enlargement in the oral cavity. Whereas, the second mode of variation describes the tongue’s tip movement to the alveolar region, and the same movement is observed in association with a backward movement of the tongue in the third mode of variation. On the other side, the first mode of variation of appearance describes the tongue’s enlargement in vertical and horizontal directions in the oral cavity. Conversely, the variation of the second mode demonstrates the forward and backward movements of the tongue associated with the rise of the larynx. Finally, the third mode of variation depicts the forward and backward movements of the tongue in direction to the palate. These results were obtained considering an AAM capable of explaining 95% of all variance of the vocal tract under study and using 10000 pixels in the construction of the texture model.

Figure 7 presents the segmentation result obtained using one of the active appearance models built on one testing MR image of each subject articulating the consonant /f/. In this figure, one may observe the evolution of the segmentation process through the same active appearance model: the process begins with a rough estimate of the vocal tract in the input image and then deforms it into the final vocal tract configuration. Similarly, the segmentation results obtained by using the model on other 4 testing MR images are depicted in Figure 8, where the first two images concerns to the female subject and the last image to the male subject during the articulation of the EP speech sounds /v/ and /a/, respectively. Additionally, the values obtained for the mean and standard deviation in order to translate the quality of the segmentation obtained in each testing MR image by the active appearance models built are included in Table 2. (Again, for a clearer understanding of the data indicated, the models have been named as *Aam_varianceretained_npixelsused* and cases of segmentation failures are indicated by a dash.) Similarity as had occurred with the active shape models used, the active appearance models built considering 99% of all variance were not able to successfully segment the modelled organ in most of the testing images and, hence, their results were not included in Table 2.

Through the analysis of the data presented in Table 2, one may conclude that in comparison to the active shape models, the active appearance models obtained better results, in other words, inferior errors of segmentation. Furthermore, it is possible to realize that the use of more modes of variation do not always assure the best results. While ASMs presented enhanced performance when 90% of all variance was addressed, AAMs addressing 95% of all variance had a superior performance when compared with the ones attaining 90% of the variance. Another significant result is that the use of 99% of modes regarding all variance translates in an extraordinary rigid model that it is not capable of be adapted to different configurations, and consequently leading to fail in the segmentation of new images.

The experimental findings are also depicted in Figures 10 and 11, from which one may verify that the active appearance models built performed better than the active shape models used. The mean errors obtained for the female subject by the active shape models varied from 7.25 (Asm_90_p11 Image 2) to 17.72 (Asm_95_p11 Image 3) pixels, 2 situations had occurred in which the segmentation failed. However, the mean errors obtained by the active appearance models varied from 5.34 (Aam_95_10000 Image 6) to 13.63 (Aam_90_5000 Image 3) pixels, and 3 unsuccessfully segmentation had occurred. The mean errors obtained for the male subject using the active shape models varied from 6.25 (Asm_95_p15 Image 1) to 15.35 (Asm_90_p7 Image 4) pixels, and one unsuccessfully case had occurred; while using the active appearance models, the mean errors varied from 4.91 (Aam_95_10000 Image 1) to 11.99 (Aam_90_10000 Image 3) pixels and the model failed to successfully segment one image.

4. Conclusions

In this work, the automatic study of the vocal tract from 3.0 Tesla MR images was assessed through the application of statistical deformable models, namely active shape models and active appearance models. The primary goal focused on the analysis of the vocal tract during the articulation of European Portuguese sounds, followed by the evaluation of the results concerning the automatic segmentation of the modelled vocal tract in new images.

While active shape models consider the information around each landmark point of the modelled structure, active appearance models also use also the grey level information of the structure. Consequently, the former type of models tends to be less efficient than the latter, and it is this information which is confirmed in this work. Nevertheless, both active shape models and active appearance models obtained remarkable results, either in terms of translating the movements and configurations involved in speech production, as well as in the segmentation of the vocal tract in new images.

One of the premises for acquiring an efficient deformable model, and consequently obtaining good results concerning the segmentation of the modelled structure, is extremely related to the quality of the images to be studied. In this work, the images studied were acquired by a 3.0 Tesla MR system and, with the higher signal-to-noise ratio and resolution, it was expected that better segmentation results can be obtained when compared to the ones achieved in 1.5 Tesla MR images [9, 18-20]. Indeed, in our previous works mean errors rounding 10 pixels were achieved when 256 x 256 pixels 1.5 Tesla MR images were used, whilst the segmentation results using the 3.0 Tesla MR images led to similar mean errors but in double sized images (512 x 512 pixels). Hence, the errors obtained by the same models that were built adopting the same modelling conditions have been around 50% reduced with the improvement of the image quality.

When compared to previous works [9, 18-20], another major contribution accomplished by this work concerns the amount of data studied. Here, 25 out of 30 possible EP speech sounds were modelled for two subjects, three measurements (slices) were used for each sound. Thus, using a training image set of 150 MR images, with more efficient and accurate models than the ones built so far could achieve, as was verified by the experimental findings obtained.

To conclude, from the work here described, one should emphasize that the recent MR imaging systems, in particular the 3.0 Tesla, and the use of the adopted statistical modelling technique have made possible the automatically and realist simulation of the vocal tract during speech production as well as the efficient segmentation of vocal tract in new images. Therefore, the assessment of the articulators' positions and movements can be facilitated, contributing, for example, to a better knowledge of the speech production, especially in patients with articulatory disorders, and to build improved computational speech models and devices.

Acknowledgments

This work was partially carried out under the scope of the project entitled "Methodologies to Analyze Organs from Complex Medical Images – Applications to the Female Pelvic Cavity", with reference

PTDC/EEA-CRO/103320/2008, financially supported by Fundação para a Ciência e Tecnologia - FCT, in Portugal.

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

Declaration of Conflicting Interests

None Declared.

References

1. Rokkaku, M., Imaizumi, S., Niimi, S., and Kiritani, S. *Measurement of the three-dimensional shape of the vocal tract on the magnetic resonance imaging technique.* Ann. Bull RILP, 1986, **20**, 47-54.
2. Baer, T., Gore, J.C., Boyce, S., and Nye, P.W. *Application of MRI to the Analysis of Speech Production.* Magnetic Resonance Imaging, 1987, **5**(1), 1-7.
3. Soquet, A., Lecuit, V., Metens, T., and Demolin, D. *From sagittal cut to area function: an RMI investigation*. in *4th International Conference on Spoken Language Processing (ICSLP 96)*, 1996, Philadelphia, USA, 1205-1208.
4. Masaki, S., Akahane-Yamad, R., Tiede, M., Shimada, Y., and Fujimoto, I. *An MRI-based Analysis of the English /r/ an /l/ Articulators*. in *4th International Conference on Spoken Language Processing (ICSLP 96)*, 1996, Philadelphia, USA, 1581-1584.
5. Engwall, O. and Badin, P. *An MRI study of Swedish fricatives: Coarticulatory effects*. in *5th Seminar On Speech Production: Models And Data*, 2000, Munchen, Germany, 297-300.
6. Takemoto, H., Kitamura, T., Nishimoto, H., and Honda, K. *A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions.* Acoust. Sci. & Tech, 2004, **25**(6), 468-473.

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

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

9. Ventura, S.R., Vasconcelos, M.J.M., Freitas, D.R., Ramos, I.M.A.P., and Tavares, J.M.R.S. *Speaker-Specific Articulatory Assessment and Measurements during Portuguese Speech Production based on Magnetic Resonance Images*, in *Language Acquisition*, 2011, Nova Science Publishers, Inc.

10. Baer, T., Gore, J.C., Gracco, L.W., and Nye, P.W. *Analysis of vocal tract shape and dimensions using Magnetic Resonance Imaging: Vowels.* Journal of the Acoustic Society of America, 1991, **90**(2), 799-828.

11. Narayanan, S., Alwan, A., and Haker, K. *An articulatory study of fricative consonants using Magnetic Resonance Imaging.* Journal of the Acoustic Society of America, 1995, **98**(3), 1325-1347.

12. Bresh, E., Nielsen, J., Nayak, K., and Narayanan, S. *Synchronized and noise-robust. audio recordings during realtime MRI scans.* Journal of the Acoustic Society of America, 2006, **120**(4), 1791-1794.

13. Parthasarathy, V., Prince, J.L., Stone, M., Murano, E.Z., and Nessaiver, M. *Measuring tongue motion from tagged cine-MRI using harmonic phase (HARP).* Journal of the Acoustic Society of America, 2007, **121**(1), 491-504.

14. Stone, M., Davis, E., Douglas, A., Nessaiver, M., Gullapalli, R., Levine, W., and Lundberg, A. *Modeling Tongue Surface Contours from Cine-MRI images.* Journal of Speech, Language, and Hearing Research, 2001, **44**(5), 1026-1040.

15. Masaki, S., Nota, Y., Takano, S., Takemoto, H., Kitamura, T., and Honda, K. *Integrated magnetic resonance imaging methods for speech science and technology.* The Journal of the Acoustical Society of America, 2008, **123**(5), 3734.

16. Nieto-Castanon, A. and Guenther, F.H. *Constructing Speaker-Specific Articulatory Vocal Tract Models for testing Speech Motor Control Hypotheses.* in *14th International Congress of Phonetic Sciences (ICPhS 99)*, 1999, San Francisco, USA, 2271-2274.

17. Apostol, L., Perrier, P., Raybaudi, M., and Segebarth, C. *3D geometry of the vocal tract and inter-speaker variability*. in *14th International Congress of Phonetic Sciences (ICPhS 99)*, 1999, San Francisco, USA, 443-446.

18. Ventura, S.R., Freitas, D.R., and Tavares, J.M.R.S. *Towards Dynamic Magnetic Resonance Imaging of the Vocal Tract during Speech Production.* Journal of Voice, 2010, **doi:10.1016/j.jvoice.2010.05.002 (in press)**.

19. Vasconcelos, M.J.M., Ventura, S.R., Freitas, D.R., and Tavares, J.M.R.S. *Towards the Automatic Study of the Vocal Tract through Magnetic Resonance Images.* Journal of Voice, 2010, **doi:10.1016/j.jvoice.2010.05.002 (in press)**.

20. Vasconcelos, M.J.M., Ventura, S.R., Freitas, D.R., and Tavares, J.M. *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, 2010, **224**(10), 1153-1163.

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

22. Schaap, J. *3D Point Distribution Models and their application in medical image segmentation using Active Shape Models: A pilot study*. 1999, MSc thesis, Delft University of Technology.

23. Ma, Z., Tavares, J.M.R.S., Jorge, R.N., and Mascarenhas, T. *A Review of Algorithms for Medical Image Segmentation and ther Applications to the Female Pelvic Cavity.* Computer Methods in Biomechanics and Biomedical Engineering, 2010, **13**(2), 235-246.

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

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

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

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

28. Cootes, T.F. and Taylor, C.J. *Active Shape Model Search using Local Grey-Level Models: A Quantitative Evaluation*. in *British Machine Vision Conference*, 1993, Guildford: BMVA Press, 639-648.

29. Cootes, T.F. and Edwards, G. *Active Appearance Models*. in *European Conference on Computer Vision*, 1998, Freiburg, Germany, 2, 484-498.

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

31. Cootes, T.F., Taylor, C.J., and Lanitis, A. *Active Shape Models: Evaluation of a Multi-Resolution Method for Improving Image Search*. in *British Machine Vision Conference*, 1994, York, England: BMVA, 1, 327-336.

32. Hamarneh, G. *ASM (MATLAB)*. 1999 [accessed in 2011], Available from: http://www.cs.sfu.ca/~hamarneh/software/code/asm.zip.

33. Cootes, T.F. *Build_aam*. 2004 [accessed in 2011], Available from: http://www.wiau.man.ac.uk/~bim/software/am_tools_doc/download_win.html.

FIGURES CAPTIONS

Figure 1: Midsagittal MR images of the vocal tract from a female subject (top row) and one of a male subject (bottom row) during production of EP vowels and consonants.

Figure 2: Training image (a), chosen landmark points (b), original image overlapped with the chosen landmark points (c).

Figure 3: Effect on the vocal tract by varying $(\pm 2sd)$ each of the first 6 modes of variation (λ_i) of the model built.

Figure 4: Test image of female (top row) and male (bottom row) subjects overlapped with the mean shape model built and after some iterations of the segmentation process of the active shape model built.

Figure 5: Four test images overlapped with the mean shape model built (top row) and after the conclusion of the segmentation process by the active shape model built (bottom row).

Figure 6: Influence of the first 3 modes of texture (left) and appearance (right) variation (λ_i) of the active appearance model built$(\pm 2sd)$.

Figure 7: Segmentation process of two test images by the active appearance model built for the vocal tract.

Figure 8: Four test images overlapped with the mean shape model built (top row), final results of the segmentation process by the active appearance model built (middle row) and correspondent original images (bottom row).

Figure 9: Mean errors (in pixels) and standard deviations of the segmentations obtained by the deformable models built for the vocal tract of the female subject.

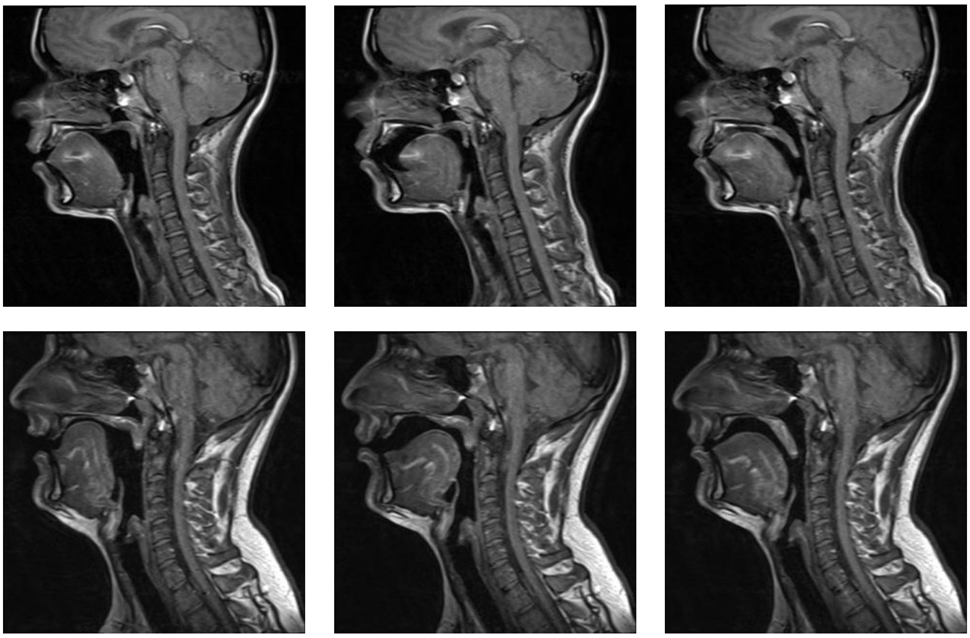
Figure 10: Mean errors (in pixels) and standard deviations of the segmentations obtained by the deformable models built for the vocal tract of the male subject.

TABLE CAPTIONS

Table 1: Retained percentages along the initial first 17 modes of variation of the model built for the vocal tract.

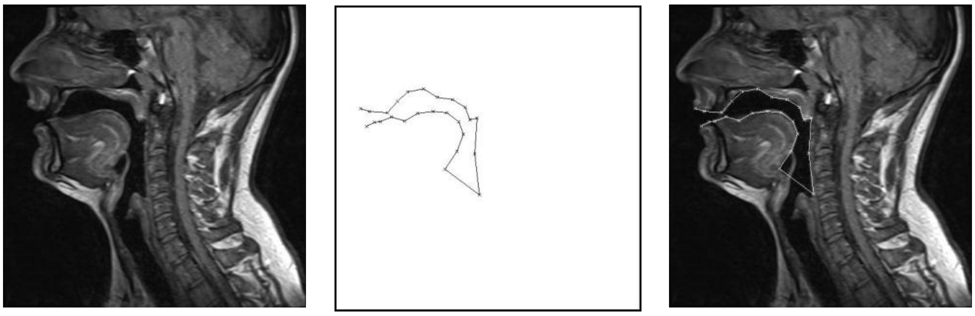
Table 2: Errors (in pixels) of the shapes segmented by the deformable models built (mean and standard deviation: $mean \pm sd$).

FIGURES



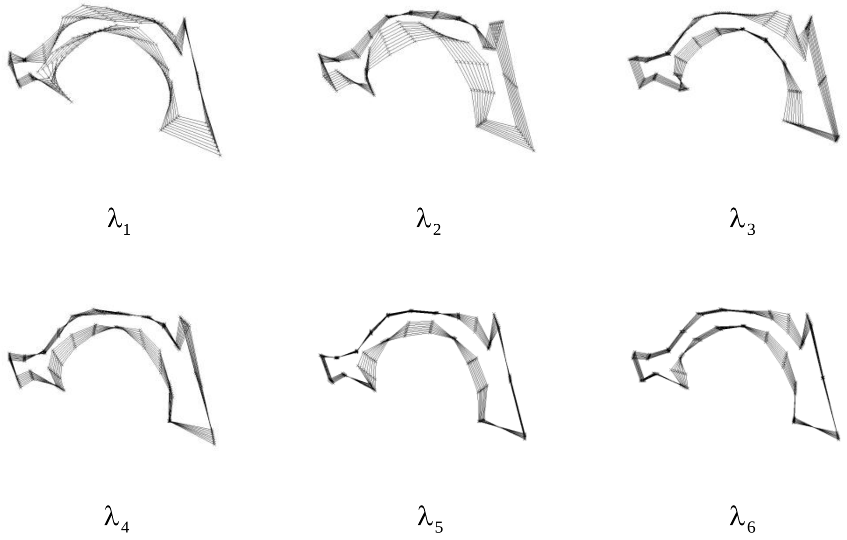
Vowel [i] Vowel [u] Consonant [m]

Figure 1



a) b) c)

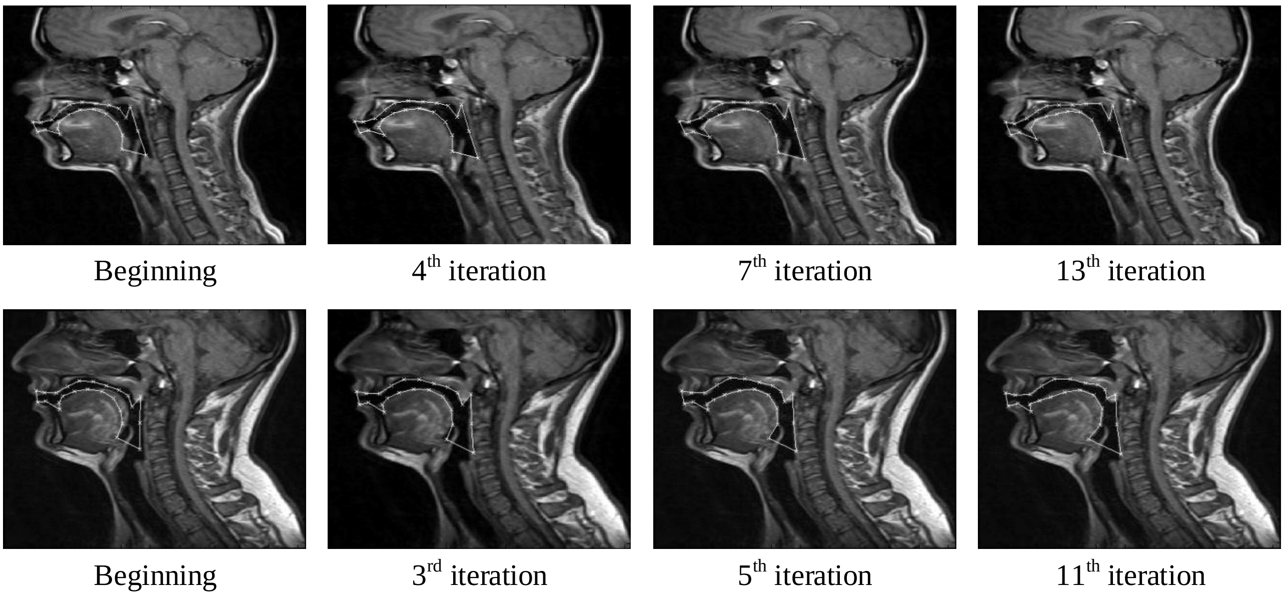
Figure 2



λ_1 λ_2 λ_3

λ_4 λ_5 λ_6

Figure 3



Beginning 4th iteration 7th iteration 13th iteration

Beginning 3rd iteration 5th iteration 11th iteration

Figure 4

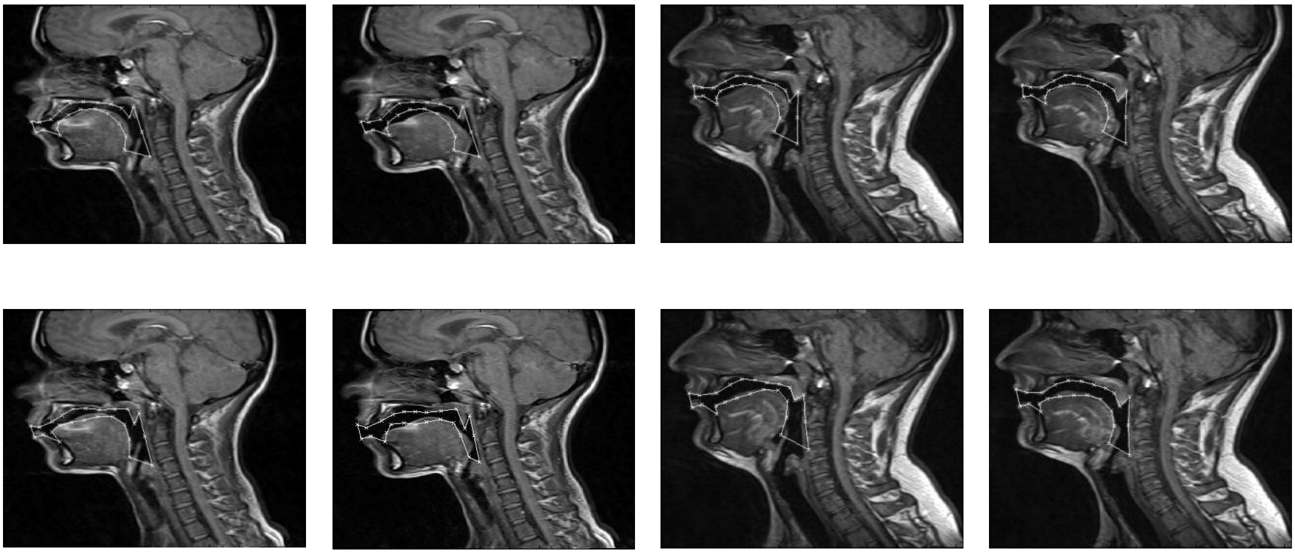
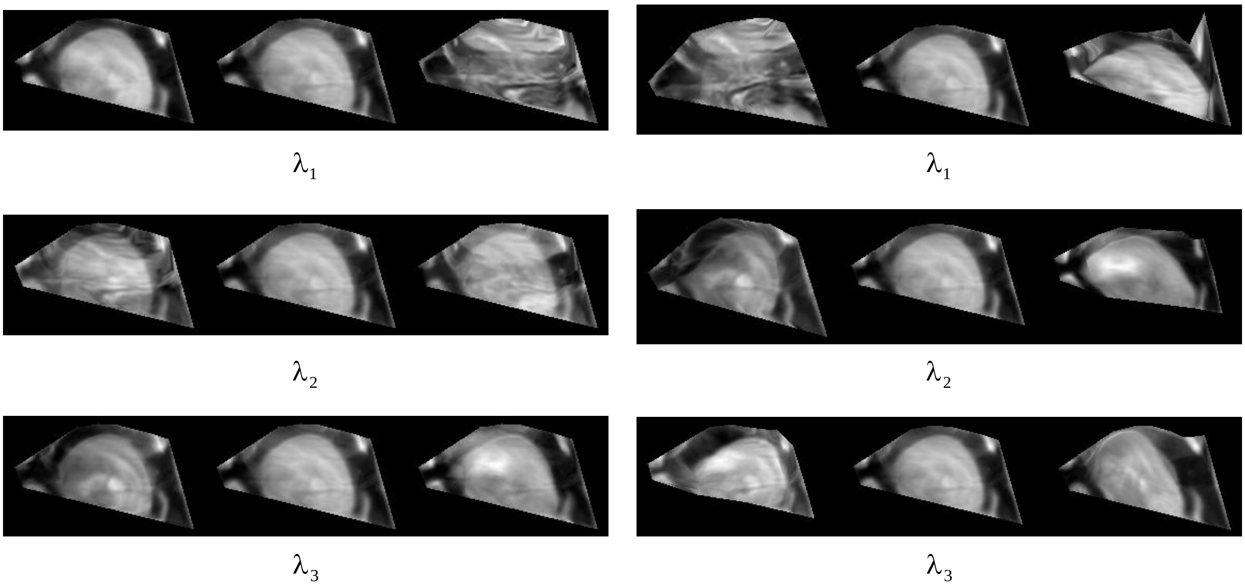


Figure 5

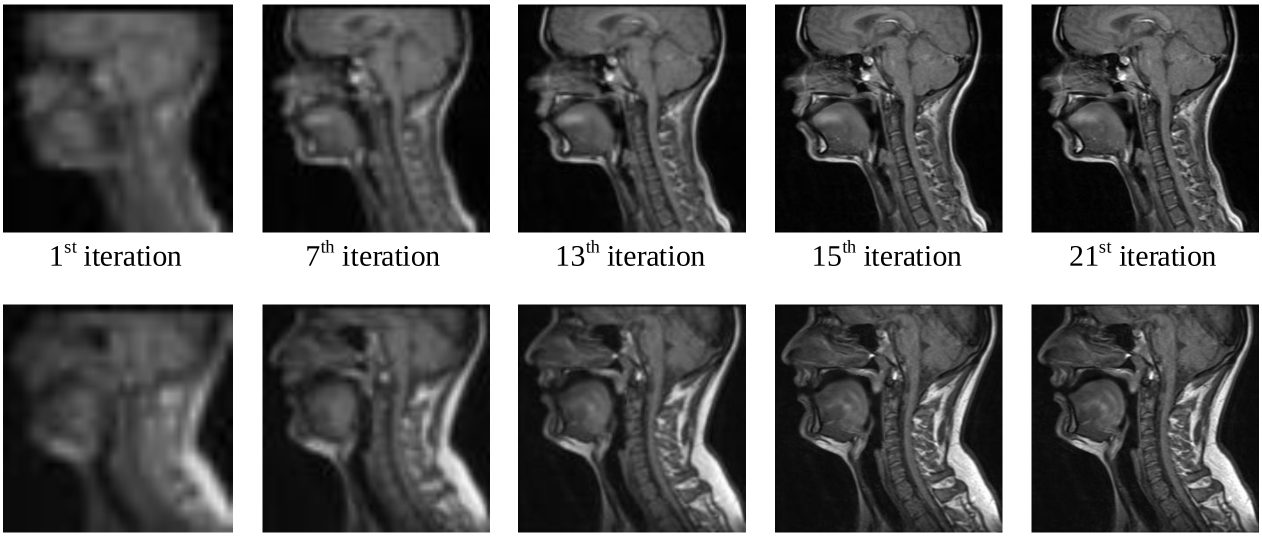


λ_1 λ_1

λ_2 λ_2

λ_3 λ_3

Figure 6



1st iteration 7th iteration 13th iteration 15th iteration 21st iteration

1st iteration 7th iteration 10th iteration 12th iteration 19th iteration

Figure 7

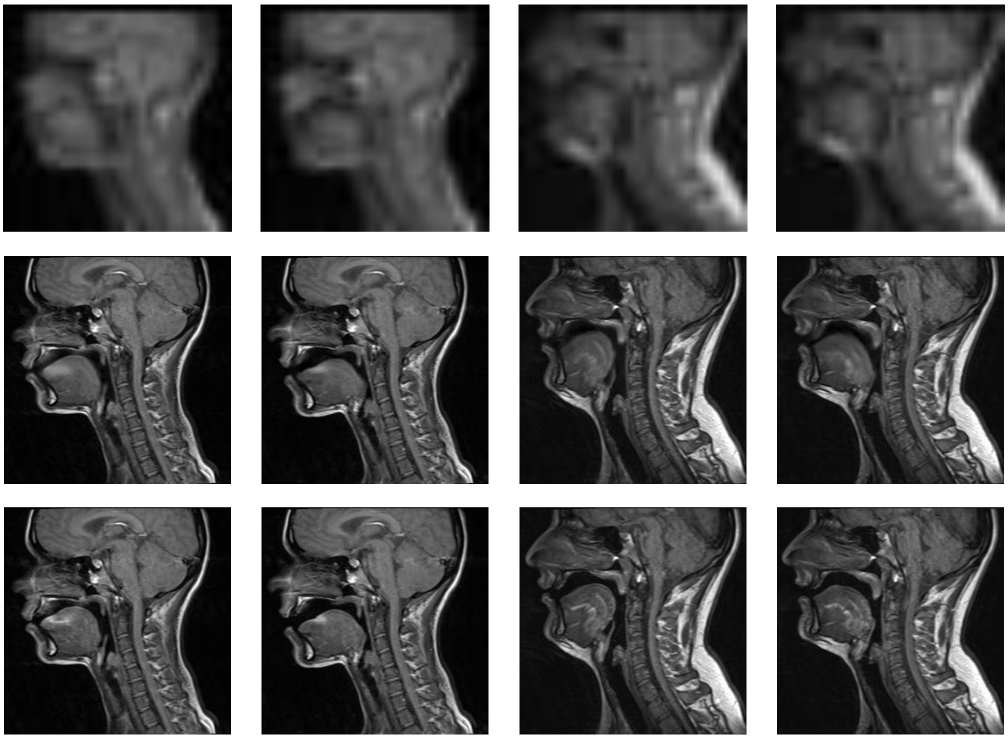


Figure 8

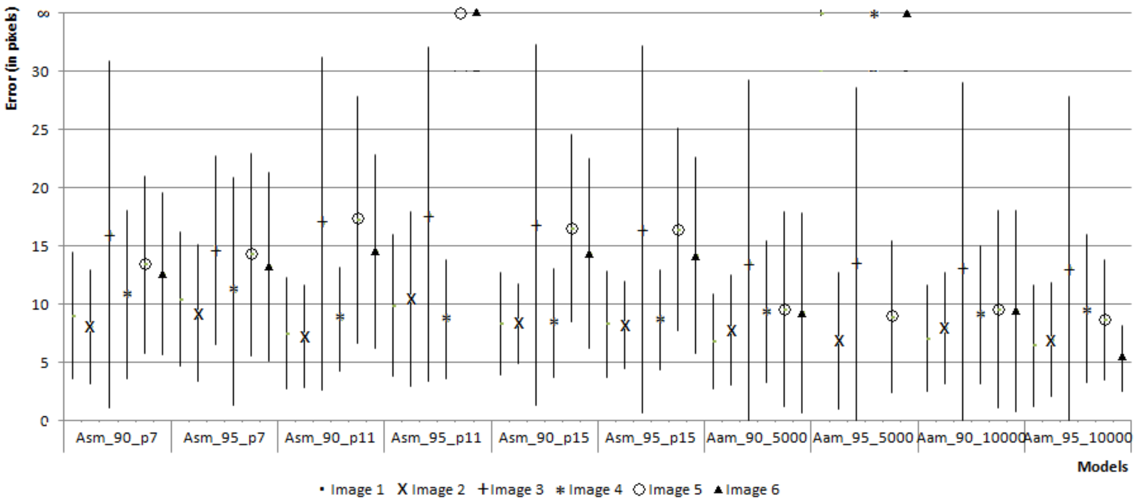


Figure 9

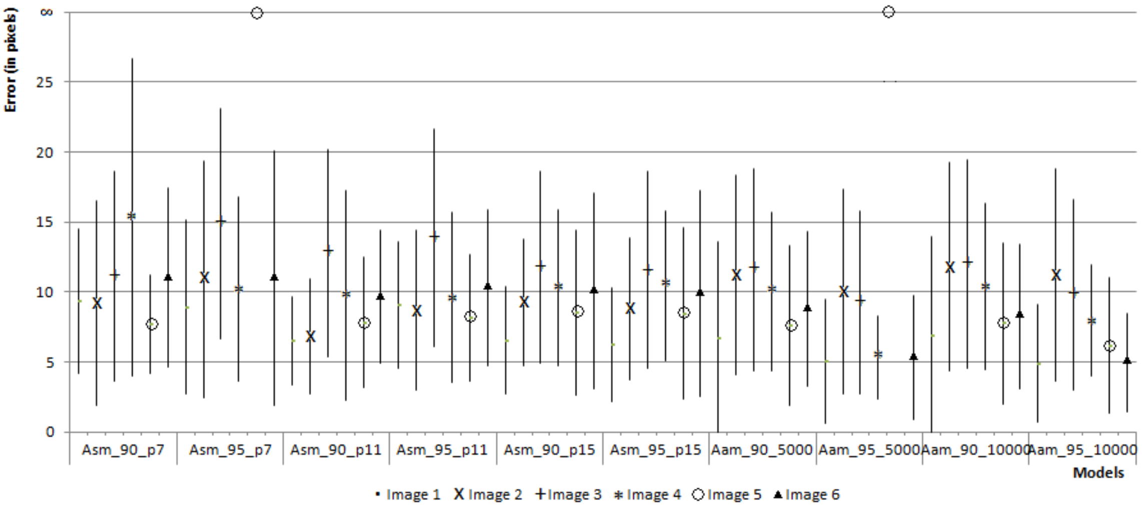


Figure 10

TABLES

Table 1

Mode of variation	Retained %	Cumulative Retained %
λ_1	40.893	40.893
λ_2	16.348	57.241
λ_3	8.065	65.306
λ_4	7.404	72.710
λ_5	4.595	77.305
λ_6	3.920	81.225
λ_7	2.515	83.740
λ_8	2.115	85.855
λ_9	1.703	87.558
λ_{10}	1.397	88.955
λ_{11}	1.296	90.251
λ_{12}	1.108	91.359
λ_{13}	1.021	92.380
λ_{14}	0.787	93.167
λ_{15}	0.677	93.844
λ_{16}	0.632	94.476
λ_{17}	0.562	95.038

Table 2

Female subject						
Model	**Image 1**	**Image 2**	**Image 3**	**Image 4**	**Image 5**	**Image 6**
Asm_90_p7	8.99±5.45	8.05±4.92	16.02±14.93	10.78±7.23	13.39±7.65	12.61±7.01
Asm_95_p7	10.40±5.77	9.23±5.92	14.63±8.16	11.07±9.80	14.29±8.70	13.21±8.13
Asm_90_p11	7.50±4.80	7.25±4.42	16.93±14.29	8.70±4.46	17.29±10.62	14.49±8.33
Asm_95_p11	9.89±6.11	10.42±7.48	17.72±14.40	8.70±5.11	-	-
Asm_90_p15	8.28±4.41	8.29±3.44	16.77±15.50	8.38±4.68	16.54±8.05	14.34±8.17
Asm_95_p15	8.29±4.56	8.19±3.78	16.40±15.79	8.67±4.29	16.40±8.73	14.19±8.41
Aam_90_5000	6.75±4.09	7.81±4.74	13.61±15.67	9.37±6.07	9.54±8.36	9.28±8.59
Aam_95_5000	-	6.87±5.89	13.53±15.06	-	8.89±6.53	-
Aam_90_1000	7.04±4.55	7.93±4.76	13.16±15.84	9.05±5.91	9.54±8.50	9.42±8.67
Aam_95_1000	6.43±5.21	6.92±4.91	13.10±14.72	9.63±6.38	8.66±5.15	5.34±2.82
Male subject						
Model	**Image 1**	**Image 2**	**Image 3**	**Image 4**	**Image 5**	**Image 6**
Asm_90_p7	9.35±5.18	9.23±7.31	11.11±7.48	15.35±11.34	7.68±3.53	11.05±6.39
Asm_95_p7	8.93±6.21	10.90±8.48	14.92±8.23	10.20±6.57	-	11.02±9.10
Asm_90_p11	6.51±3.12	6.83±4.12	12.81±7.41	9.80±7.50	7.83±4.65	9.66±4.73
Asm_95_p11	9.08±4.55	8.71±5.71	13.87±7.77	9.65±6.09	8.13±4.53	10.30±5.55
Asm_90_p15	6.53±3.85	9.25±4.54	11.75±6.86	10.33±5.55	8.56±5.91	10.11±7.01
Asm_95_p15	6.25±4.09	8.84±5.07	11.59±7.05	10.46±5.36	8.47±6.11	9.94±7.36
Aam_90_5000	6.75±6.84	11.22±7.13	11.61±7.23	10.05±5.65	7.62±5.68	8.81±5.51
Aam_95_5000	5.06±4.40	10.05±7.29	9.28±6.52	5.32±2.98	-	5.32±4.45
Aam_90_10000	6.93±7.06	11.82±7.47	11.99±7.46	10.37±5.93	7.78±5.78	8.24±5.19
Aam_95_10000	4.91±4.19	11.20±7.59	9.79±6.81	7.96±3.97	6.19±4.81	4.97±3.55