Towards the Automatic Study of the Vocal Tract from Magnetic Resonance Images

Maria João M. Vasconcelos
Faculty of Engineering, University of Porto
Laboratory of Optics and Experimental Mechanics, 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, School of Allied Health Science – IPP
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
Laboratory of Optics and Experimental Mechanics, 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
Laboratory of Optics and Experimental Mechanics, Institute of Mechanical Engineering and Industrial Management
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

Towards the Automatic Study of the Vocal Tract through Magnetic Resonance Images

Abstract: Over the last few decades, researchers have been investigating the mechanisms involved in speech production. Image analysis can be a valuable aid in the understanding of the morphology of the vocal tract. The application of magnetic resonance imaging to study these mechanisms has been proven to be reliable and safe. We have applied deformable models in magnetic resonance images in order to conduct an automatic study of the vocal tract; mainly, to evaluate the shape of the vocal tract in the articulation of some European Portuguese sounds, and then to successfully automatically segment the vocal tract's shape in new images. Thus, a Point Distribution Model has been built from a set of magnetic resonance images acquired during artificially sustained articulations of twenty one sounds, which successfully extracts the main characteristics of the movements of the vocal tract. The combination of that statistical shape model with the grey levels of its points are subsequently used to build Active Shape Models and Active Appearance Models. Those models have then been used to segment the modelled vocal tract into new images in a successful and automatic manner. The computational models have thus been revealed to be useful for the specific area of speech simulation and rehabilitation, namely to simulate and recognize the compensatory movements of the articulators during speech production.

Keywords*:* Vocal Tract's Shape; Morphological Study; Speech Production; Speech Rehabilitation; Modelling and Simulation; Image Segmentation; Deformable Models.

1. Introduction

Verbal communication is the most common, familiar and frequently used form of human interaction, which is the direct result of the organised and synchronised performance of a set of anatomic organs. The articulation is a result of the activity of a set of organs: the vocal tract being responsible for modifying their position and shape during the air expulsion (expiration) process, thereby producing different sounds and consequently, distinct acoustic representations [1-3].

The main anatomic aspects and the physiology of the vocal tract are common to all individuals. However, the mechanism engaged in human speech production is complex and unique by nature, due to a variety of anatomical structures that compose the vocal tract, which implies that any computational modelling developed needs to be flexible so as to allow for accurate individual characterizations [1-3]. Due to this, the study of the speech production is a multidisciplinary subject which involves areas as diverse as, for instance: Medicine, due to the anatomic and functional study of the vocal tract organs; Engineering, in particular biomedical engineering, more specifically, in the field of an acoustics analysis and speech processing; Phonetics, in terms of the study of the production and perception of speech and sounds; Speech therapy as far as the assessment of anatomic and physiological aspects related to communication disorders, language and speech are concerned; and finally Medical imaging, in terms of the improvement and application of computational image techniques which can be used in the study or the vocal tract during speech production [1-4].

The use of Magnetic Resonance Imaging (MRI) has produced very useful results in assessing speech and vocal tract's shapes [1-3, 5-6]. This image technique allows for morphologic measurements in static [1-2, 7-8] as well as dynamic studies [3, 9-10], in addition to being capable of representing the entire vocal tract along any orientation. Image quality allied with high resolution as far as soft-tissues are concerned, represent another key advantage of using MRI as this allows for the enhanced analyses of the vocal tract through the calculi of several descriptive parameters. Another imaging technique that has been used to acquire shape information about the vocal tract is X-ray Computed Tomography (CT) [11]. However, this imaging modality has the disadvantage of requiring significant ionizing radiation doses.

Since the mid seventies one has witnessed an extensive tradition of adopting statistical methods to analyse data from speech production. For example, in [12] the authors used component analysis to identify an adequate set of articulatory features for the tongue shapes of ten English vowels. One year later, the statistical analysis of real data to describe the position of the articulatory organs was used [13]. Additionally, in another study, [14], a factor analysis of the lateral shapes of the vocal tract was described. Thereafter, a principal component analysis to examine sagittal tongue contours for five English vowels from ultrasound images was put into practice [15]. Yet, despite all the work that has been extensively carried out in the area, the application of statistical deformable models to characterize and reconstruct speech sounds with MR images has only just become a reality. Thus, information concerning this important research subject is still incredibly scarce, particularly in the case of the European Portuguese (EP) language [1-3, 16-17].

The use of deformable models is one of the image analysis approaches which has produced interesting results in innumerable unrelated applications. Active contours, deformable templates, physical models and point distribution models are the most relevant techniques in the extraction of the characteristics of an object from images based on deformable models [18]. Active contours, more commonly known as snakes in the field of computational vision, were the first deformable models to be used in image analysis [19]. They consist of an elastic set of points which may be adapted to the shape of the object being studied through the physical combination of internal and external forces. Recently, enhanced physical modelling approaches have allowed for the integration of the physical behaviour of the actual objects by taking previously acquired knowledge about them into consideration in the models built [20]. On the other hand, image analysis techniques based on deformable templates resort to geometrical shapes (templates) driven by parameterized functions in order to correctly identify the objects modelled in images [21]. Finally, Point Distribution Models (PDM) are built from a set of training images and are capable of extracting the main characteristics of the object through statistical techniques [22-23].

In this work, deformable models have been applied in the automatic study of the vocal tract from magnetic resonance images; mainly, to evaluate the shape of the vocal tract in the articulation of European Portuguese sounds and then to automatically segment the vocal tract into new images. Thus, a Point Distribution Model has been built from a set of magnetic resonance images acquired during the artificially sustained articulation of 21 European Portuguese sounds, in order to determine the main characteristics of the movements involved in the vocal tract. The combination of this particular statistical shape model with the grey levels of its points is then used to construct Active Shape Models (ASMs) and Active Appearance Models (AAMs). Next, these active models are used to segment the modelled vocal tract into new images.

This paper is organized as follows: in the following section, the MRI protocol adopted is presented. Next, PDM and active models, in addition to the data used and the assessment performed in terms of quality segmentation, are then focused on. Following this, the statistical models built and their employment in terms of the segmentation of EP speech sounds from new MR images are presented. In the final section of the paper, conclusions are drawn.

2. Methods

2.1. MRI protocol

Image acquisition was obtained through the use of a Siemens Magneton Symphony 1.5T system and a head array coil, with the subject lying in the supine position. As a result of this experimental setup, the acquisition of T1-weighted 5 mm thick sagittal slices were obtained by using Turbo Spin Echo Sequences, with the acquisition duration of approximately 10 s. Decreasing the slice thickness entails the use of a low signal noise ratio which makes the posterior segmentation process more complex. Subsequently, this protocol has resulted from a compromise required between the signal noise ratio, the number of slices acquired and the time needed for subjects to successfully sustain articulation during the image acquisition process. The following acquisition parameters were used: field of view equal to 150 mm, image matrix of 128x128 and image resolution equal to 0.853 px/mm.

The study which was carried out under the scope of this project was designed to obtain morphologic data about the vast majority of the articulators' range positions with the distinct aim of characterising and reconstructing EP speech sounds. In view of the fact that, the sagittal data considered is particularly useful in the study of the entire vocal tract anatomy, as may be confirmed by Figure 1; it is possible to obtain information as to the main aspects of the shape and positions of some articulators, e.g. tongue, lips and velum. Thus, the speech corpus consisted of a set of 25 MR images obtained during sustained articulations of 25 EP speech sounds; that is, 1 (one) sagittal image was acquired for each sound considered.

2.2. Point Distribution Models and Active Models

As has been previously mentioned, Point Distribution Models have been used in the statistical modelling of objects to analyse their shape configurations from a set of training images. Thus, a PDM describes the mean shape of the object modelled, as well as the admissible variations in relation to that same mean shape [4, 23]. In this work, the vocal tract was statistically modelled by a PDM from a set of 21 MR images partially depicted in Figure 2. These images allow one to observe the vocal tract configurations during the production of distinct EP vowels and consonants, as well as of some oral and nasal sounds. In order to obtain a robust PDM, the sounds used in the training process of the model built should adequately represent the variability of the vocal tract's shape. Additionally, each shape of the vocal tract that is presented in the training set should be described by a group of labelled landmark points conveying important aspects of the vocal tract, Figure 3. (In the current and subsequent images, the landmark points appear connected by fictitious line segments so as to enhance their visualisation). Consequently, the manual identification of the landmark points in the 21 training obtained images requires a comprehensive knowledge of the object in question, as the resultant model behaviour greatly depends on the landmark points selected. Hence, the fact that the manual selection of the landmark points was carried out by one of the authors who has excellent knowledge on medical imaging and the anatomy of the vocal tract, in addition to being cross-checked by another co-author, in accordance with the following criteria:

- 4 points at the lips (front and back of 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 (one) point at the velum (or soft palate);

- 3 points equally spaced at the posterior margin of the oropharynx (behind the oral cavity).

Hence, 25 landmark points were defined in each of the 21 MR images used to obtain the PDM of the shape of the vocal tract during the production of the EP sounds.

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 dynamic programming for instance [4, 23-27]. Hence, given the co-ordinates (x_{ij}, y_{ij}) at each landmark point j of the shape i of the modelled object, the shape vector is:

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

where $i = 1 \ldots N$, with N representing the number of shapes in the 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 object may be found. The modes of variation characterize the manner in which the landmarks of the modelled object tend to move together, the result of which may be obtained by applying Principal Component Analysis (PCA) to the deviations from the mean. Thus, it is possible to rewrite each vector x_i as:

$$x_i = \overline{x} + P_s b_s , \qquad (1)$$

where x_i represents the coordinates of the n landmark points of the new shape of the modelled object, (x_k, y_k) are the coordinates of landmark point k, $\overline{x}$ is the mean position of all landmark points, $P_s = \begin{pmatrix} p_{s1} & p_{s2} & \ldots & p_{st} \end{pmatrix}$ is the matrix of the first t modes of variation, p_{si} corresponds to the most significant eigenvectors in a Principal Component Analysis applied to the coordinates of all landmark points, and $b_s = \begin{pmatrix} b_{s1} & b_{s2} & \ldots & b_{st} \end{pmatrix}^T$ is a vector of weights for each variation mode of the modelled object. Each eigenvector describes the manner in which linearly correlated x_{ij} move together over the training set, and due to this, it is commonly known as a mode of variation. Thus, Equation (1) represents the PDM of an object and may be used to generate its new shapes [4, 23].

The local grey-level environment of each landmark point may also be considered in the statistical modelling of objects from images [23, 28]. Thus, statistical information is obtained in relation to the mean and covariance of the grey values of the pixels around each landmark point. Hence, this information may be used to evaluate the matching between landmark points with Active Shape Models, in addition to, constructing appearance models of objects with the aid of Active Appearance Models, as is explained further on in the paper.

Active Shape Model

The consideration of a PDM and the grey level profiles of each landmark point used in the construction of the PDM may be used to segment the object modelled in new images through Active Shape Models, which are based on an iterative technique for fitting flexible models with objects represented in images [23, 29]. This technique is an iterative optimisation scheme which is combined with PDMs that then refine an initial estimated shape, that is, the mean PDM's shape, $\overline{x}$, of a modelled object according to the PDM's modes of variation into a new image, i.e. this process segments the modelled object in new images. 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 which 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 shape desired.

The image segmentation process with the aid of Active Shape Models was improved in [30] due to the adoption of a multiresolution approach which may be summarised as follows: To begin with, 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 is capable of segmenting the input images in a more rapid and reliable manner.

Active Appearance Model

An image segmentation approach based on Active Appearance Models was first proposed in [3 1]. It allows for the building of texture and appearance models of objects from images. These models are generated by combining a shape variation model (a geometric model) with an appearance variations model in a shape-normalised frame [23]. The geometric model of the object modelled is a PDM which may also described by Equation **Error! Reference source not found.**.

To build a statistical model of the grey level appearance of an object represented in images, one needs to deform each training image in order for the landmark points to match the mean shape of the object. This is done by using a triangulation algorithm. Next, the grey level information or 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 the global light variation in the images, the average vector of the grey levels (g_{im}) is once again normalized, thereby resulting in vector g. Following the application of a Principal Component Analysis 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 object modelled can be defined by vectors b_s and b_g.

Given that a correlation may exist between the variations of the shape and grey levels, a further Principal Component Analysis is applied to the data of the object. 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 Principal Component Analysis is applied to these vectors, which results in a further model:

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

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

2.3. Data used and assessment on segmentation quality

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

According to the International Phonetic Alphabet (IPA), the EP speech language consists of a total of 30 sounds. In this work, 21 of these sounds have been considered in the construction of the statistical models of the vocal tract's shape by using a MR image for each one, Figure 2. The sounds considered include the most representative sounds of the EP speech language. Additionally, 4 distinct MR images, related to 4 other EP speech sounds, were later used to evaluate the quality of the segmentation obtained by the Active Models built.

Due to the inevitable variability of the speech subject, in addition to the considerable amount of sounds studied (speech corpus), a set of 25 MR images was acquired from one young male subject in a similar manner to that which has been used in other works which resort to MRI to study the vocal tract during speech production [1-3, 7-8, 34]. The subject was trained so as to ensure the correct production of the intended EP speech sounds and to reduce speech subject variability. Moreover, it should be stressed that the subject in question had a vast knowledge of EP speech therapy.

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 [23], ASMs were built with values ranging from 95% to 99% of retained variance and with profiles of 7, 11 and 19 pixels for the grey levels. Similarly, Active Appearance Models were built with identical values ranging from 95% to 99% of retained variance and the following values of 50000 and 10000 pixels were considered for the texture model.

Following the construction of the Active Shape Models and the Active the Appearance Models from the training set of 21 MR images, these were then used to segment the vocal tract's shape in 4 MR images which were not included in the training set used. As a stopping criteria for the segmentation process, a maximum of 5 iterations for each resolution level was taken into consideration. Due to the fact that 4 resolution levels were defined based on the dimensions of the images in question, this criteria meant that from the beginning of the segmentation process to its end, a maximum of 20 iterations could take place [23]. This maximum number of iterations was chosen as a result of the fact that in the concrete case of the images considered it lead to excellent segmentation results. Additionally, it was proven that an inferior value was not always sufficient to obtain satisfactory segmentations and a superior value constantly resulted in the same segmentation results.

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

3. Results and Discussion

As has been previously mentioned, the initial task of the construction process of the active models under consideration is related to the manual labelling of the MR images of the vocal tract. As expected, this task was revealed to be difficult and extremely time consuming. This because, the considerable noise present in the MR images and the significant variability of the sounds under study cause the use of automatic approaches to be a complex process.

In Table 1, the initial 15 modes of variation of the active shape model built and their retained percentages 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 a total value of 95% of all variance in addition to the fact that the initial 15 modes which correspond to 30% of the total modes of variation, provide an explanation for 99% of all variance. This indicates that the ASM built is capable of considerably reducing the data required to represent all of the shapes that the vocal tract assumes in the set of training images acquired.

The effects of varying the first 6 modes of variation are depicted in Figure 4. This figure allows one to conclude that the first mode is associated with the movements of the tongue from the high front to the back positions of the oral cavity. In respect to the second mode of variation, it is possible to observe the vertical movement of the body of the tongue towards the palate. On the one hand, the variations of the third mode are related to the opening of the lips and tongue's movement to a backward position. On the other hand, the fourth mode of variation reflects the movement of the tongue tip from the central position of the tongue to the alveolar ridge of the palate. Additionally, the fifth mode of variation represents the opening of the lips and the overall lateral enlargement of the vocal tract. Finally, the sixth mode is related to the movement of the body of the tongue from back to front and downward positions.

Next, 4 MR images of 4 distinct EP speech sounds, which had not been previously considered in the set of training images used, were segmented by the active shape models built. In Figure 5, one of the segmentations which was obtained is depicted. In this Figure one may observe the evolution of the segmentation process through the actual active shape model which is built: the process begins with a rough estimate as to the localization of the vocal tract in the image (1st iteration), then moving downwards to each multiresolution level (4th and 9th iteration) until it converges into the vocal tract's shape that is to be segmented after 14 iterations. This segmentation was obtained by considering an active shape model capable of explaining 95% of all variance of the vocal tract's shape under study and adopting a grey level profile length of 7 pixels, that is by considering 3 pixels from each side of the landmark points [23]. Similarly, the segmentation results obtained by using this model on the 4 MR images tested are presented in Figure 6.

In Table 2, the values of the mean and standard deviation, which reflect the quality of the segmentation obtained in each testing MR image through the active shape models built, are put forward. (For a more comprehensive understanding of the data presented in this table, the models are named as: *Asm_varianceretained_profiledimension*.). As has been previously said, active shape models with a grey level profile of dimensions equal to 11 and 19 pixels were also built. However, these active shape models were not able to successfully segment the modelled organ in the testing images. This failure is due to the relatively small size of the images considered: during the segmentation process, at each landmark point a segment of 22 (or 38) pixels long has been considered in the active search process. This result in the fact that the model built can easily diverge [23].

As has been formerly indicated, active appearance models are also capable of segmenting the objects modelled in new images. By considering 95% of all the shape's variance and 10000 pixels in the construction of the texture model, a total of 9 modes of shape variation were extracted corresponding to 18% of the modes of variation, 17 texture modes, i.e. 34% of the modes of variation and 13 appearance modes, which sums up to 26% of the modes of variation. Additionally, if an active appearance model is built using 99% of all the shape variance in addition to 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 in terms of the texture and appearance of one of the active appearance models built are depicted in Figure 7. This figure clearly demonstrates that the first mode is associated with tongue's movements from the high front to backward 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 corresponds to the lips' and the tongue's movement to a backward position. It should be stressed that these modes of variation also provide information as to the appearance. As such, the intensity profiles associated with each structure of the vocal tract have also been considered.

Figure 8 presents the segmentation result obtained using one of the active appearance models built on a MR testing image. In this figure, it is possible to observe the evolution of the active search required to correctly segment the organ modelled: the process is begun with a rough estimation as to the localization of the vocal tract in the image (1st iteration), moving downwards to each multiresolution level (7th and 12th iteration) until it converges into the desired vocal tract' shape after a total of 20 iterations. Similarly, the segmentation results using the same model on the all testing MR images are depicted in Figure 9. Additionally, the values obtained for the mean and standard deviation, which translate the quality of the segmentation obtained in each testing MR image by the active appearance models built, are presented in Table 2. (For a clearer understand of the data indicated, the models have been named: *Aam_varianceretained_npixelsused*).

As a result of the analysis of the data presented in the Table 2, one may conclude that active appearance models lead to superior results to those obtained by active shape models. Furthermore, the use of a larger number of modes of variation leads to superior results when the active appearance models are used. Once again this contrasts with the segmentation results obtained by using active shape models. In the latter case, the use of more modes of variation (retained percent) does not necessarily correspond to improved results.

4. Conclusions

In this work, statistical deformable models were applied to Magnetic Resonance Images so as to study the shape of the vocal tract in the articulation of a number of European Portuguese sounds, Furthermore, the models built were used to segment the shape of the vocal tract's shape in new MR images.

The use of MR images permits the non-invasive study of the entire vocal tract with the required safety and high quality imaging of the soft tissues. However, the main drawback of this procedure is the amount of noise that is usually presented in the images acquired. Nevertheless, this drawback was overcome by the models built and satisfactory segmentation results have been obtained.

From the experimental results obtained, one may conclude that the point distribution model built can successfully extract the main characteristics of the movements of vocal tract from magnetic resonance images. Furthermore, one may verify that the active shape models and the active appearance models can be used to segment the modelled vocal tract in new MR images in an efficacious and automatic manner. Therefore, the models built are accurate and efficient tools to be used in terms of the automatic study of the vocal tract from magnetic resonance images during speech production.

The knowledge obtained in this work has resulted in a better understanding of the articulatory movements during the phonation process, which is extremely useful for the assessment of speech disorders in addition to speech simulation. Additionally, these findings may also be used as a supplementary tool for therapeutic planning and follow-up by both physicians and speech therapists.

In the near future, MR volumetric data will be used by the generation of 3D models in order to obtain more accurate representations of the vocal tract during speech production. Moreover, the statistical models adopted in this work will be employed in the analysis and segmentation of that same data.

5. 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 following corresponding references PTDC/EEA-CRO/103320/2008, UTAustin/CA/0047/2008 and UTAustin/MAT/0009/2008, respectively, and were financially supported by FCT - Fundação para a Ciência e a Tecnologia in Portugal.

The first and second authors would like to hereby acknowledge the support of their PhD grants with the following references SFRH/BD/28817/2006 and SFRH/PROTEC/49517/2009, respectively, from FCT.

The images considered in this work were acquired at the Radiology Department of the Hospital S. João, Porto, with the collaboration of Isabel Ramos (Professor at the Faculty of Medicine, University of Porto, and Department Director) and of the technical staff, which is gratefully acknowledged.

Finally, the authors would like to express their deep gratitude to Dr. Vanessa Esteves from the University of Porto – Portugal for her valuable help with the English version of this paper.

6. References

1. Ventura, S., Freitas, D., Tavares, J.M.R.S., *Imaging of the Vocal Tract based on Magnetic Resonance Techniques*, in *Computer Vision, Imaging and Computer Graphics: Theory and Applications*. 2010, Springer. p. 146-157.
2. Ventura, S.R., Freitas, D.R., 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): p. 671-681.
3. Ventura, S.R., Freitas, D.R., Tavares, J.M.R.S., *Towards Dynamic Magnetic Resonance Imaging of the Vocal Tract during Speech Production.* Journal of Voice, 2010. **(in press)**.
4. Vasconcelos, M.J.M., Ventura, S.M.R., Freitas, D.R.S., Tavares, J.M.R.S., *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. **(in press)**.

5. Engwall, O. *A revisit to the Application of MRI to the Analysis of Speech Production - Testing our assumptions*. in *6th International Seminar on Speech Production*. 2003. Sydney, Australia.
6. Crary, M.A., Kotzur, I.M., Gauger, J., Gorham, M., Burton, S., *Dynamic Magnetic Resonance Imaging in the Study of Vocal Tract Configuration.* Journal of Voice, 1996. **10**(4): p. 378-88.
7. Badin, P., Serrurier, A. *Three-dimensional Modeling of Speech Organs: Articulatory Data and Models*. in *IEICE Technical Committee on Speech*. 2006. Kanazawa, Japan.
8. Serrurier, A., Badin, P., *Towards a 3D articulatory model of velum based on MRI and CT images.* ZAS Papers in Linguistics, 2005. **40**: p. 195-211.
9. Avila-Garcia, M.S., Carter, J.N., Damper, R.I., *Extracting Tongue Shape Dynamics from Magnetic Resonance Image Sequences*, in *International Conference on Signal Processing* 2004: Istanbul, Turkey. p. 288-291.
10. Mády, K., Sader, R., Zimmermann, A., Hoole, P., Beer, A., Zeilhofe, H., Hannig, C. *Use of real-time MRI in assessment of consonant articulation before and after tongue surgery and tongue reconstruction*. in *4th International Speech Motor Conference*. 2001. Nijmegen, Netherlands.
11. Inohara, K., Sumita, Y.I., N.Ohbayashi, Ino, S., Kurabayashi, T., Ifukube, T., Taniguchi, H., *Standardization of Thresholding for Binary Conversion of Vocal Tract Modeling in Computed Tomography.* Journal of Voice, 2009. **(in press)**.
12. Harshman, R.A., Ladefoged, P., Golstein, L., *Factor analysis of tongue shapes.* Journal of the Acoustical Society of America, 1977. **62**: p. 693-707.
13. Shirai, K., Honda, M., *Estimation of articulatory motion by a model matching method.* Journal of the Acoustical Society of America, 1978. **64**(S1): p. S42-S42.
14. Maeda, S., *Improved articulatory models.* Journal of the Acoustical Society of America, 1988. **84**(S1): p. S146-S146.
15. Stone, M., Cheng, Y., Lundberg, A., *Using principal component analysis of tongue surface shapes to distinguish among vowels and speakers.* Journal of the Acoustical Society of America, 1997. **101**(5): p. 3176-3177.
16. Teixeira, A., Vaz, F., Martinho, L., Coimbra, R.L. *Articulatory synthesis of Portuguese*. in *III Encontro do Forum Internacional de Investigadores Portugueses*. 2001. IEETA, Aveiro, Portugal.
17. Martins, P., Carbone, I., Pinto, A., Silva, A., Teixeira, A., *European Portuguese MRI based speech production studies.* Speech Communication, 2008. **50**(11-12): p. 925-952.
18. Ma, Z., Tavares, J.M.R.S., Jorge, R.N., Mascarenhas, T., *A Review of Algorithms for Medical Image Segmentation and their Applications to the Female Pelvic Cavity.* Computer Methods in Biomechanics and Biomedical Engineering, 2010. **13**(2): p. 235-246.
19. Kass, M., Witkin, A., Terzopoulos, D., *Snakes: Active Contour Models.* International Journal of Computer Vision, 1987. **1**(4): p. 321-331.
20. Gonçalves, P.C.T., Tavares, J.M.R.S., Jorge, R.M.N., *Segmentation and Simulation of Objects Represented in Images using Physical Principles.* Computer Modeling in Engineering & Sciences, 2008. **32**(1): p. 45-55.
21. Yuille, A.L., Cohen, D., Hallinan, P., *Feature extraction from faces using deformable templates.* International Journal of Computer Vision, 1992. **8**(2): p. 104-109.
22. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J. *Training Models of Shape from Sets of Examples*. in *British Machine Vision Conference*. 1992. Leeds, UK.
23. Vasconcelos, M.J.M., 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): p. 213-241.
24. Oliveira, F.P.M., 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, No. 1**: p. 1-11.

25. Oliveira, F.P.M., 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): p. 1-11.

26. Oliveira, F.P.M., Tavares, J.M.R.S., *Matching Contours in Images through the use of Curvature, Distance to Centroid and Global Optimization with Order-Preserving Constraint.* Computer Modeling in Engineering & Sciences, 2009. **43**(1): p. 91-110.

27. Oliveira, F.P.M., Tavares, J.M.R.S., Pataky, T.C., *Rapid pedobarographic image registration based on contour curvature and optimization.* Journal of Biomechanics, 2009. **42**(15): p. 2620-2623.

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

29. Cootes, T.F., Taylor, C.J. *Active Shape Models - 'Smart Snakes'*. in *British Machine Vision Conference*. 1992. Leeds, UK.

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

31. Cootes, T.F., Edwards, G. *Active Appearance Models*. in *European Conference on Computer Vision*. 1998. Freiburg, Germany.

32. Hamarneh, G. *Active Shape Models (MATLAB)*. 1999; Available from: http://www.cs.sfu.ca/~hamarneh/software/code/asm.zip.

33. Cootes, T.F. *For building active appearance models (AAMs - am_build_aam)*. 2004; Available from: http://www.wiau.man.ac.uk/~bim/software/am_tools_doc/download_win.html.

34. Story, B.H., *Comparison of Magnetic Resonance Imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002.* Journal of the Acoustical Society of America, 2008. **123**(1): p. 327-335.

FIGURE CAPTIONS

Figure 1: MR sagittal slice with the indication of the organs of the vocal tract.

Figure 2: Examples of the training images used to build the models of the vocal tract's shape.

Figure 3: Training image a), landmark points used b) and training image with the overlapped landmark points used c).

Figure 4: Effects of varying each of the first 6 variation modes of the model built for the vocal tract's shape $(\pm 2sd)$.

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

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 (e-h).

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

Figure 8: Results after the 1st, 7th, 12th and 20th iterations of the segmentation process using one 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).

FIGURES

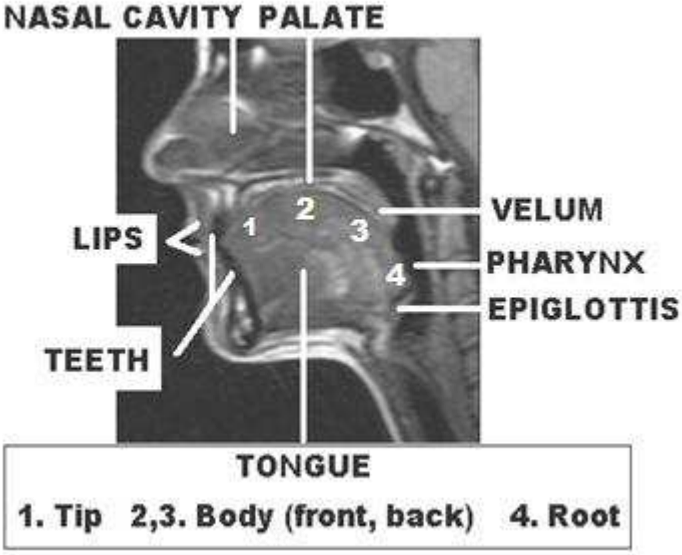


Figure 1

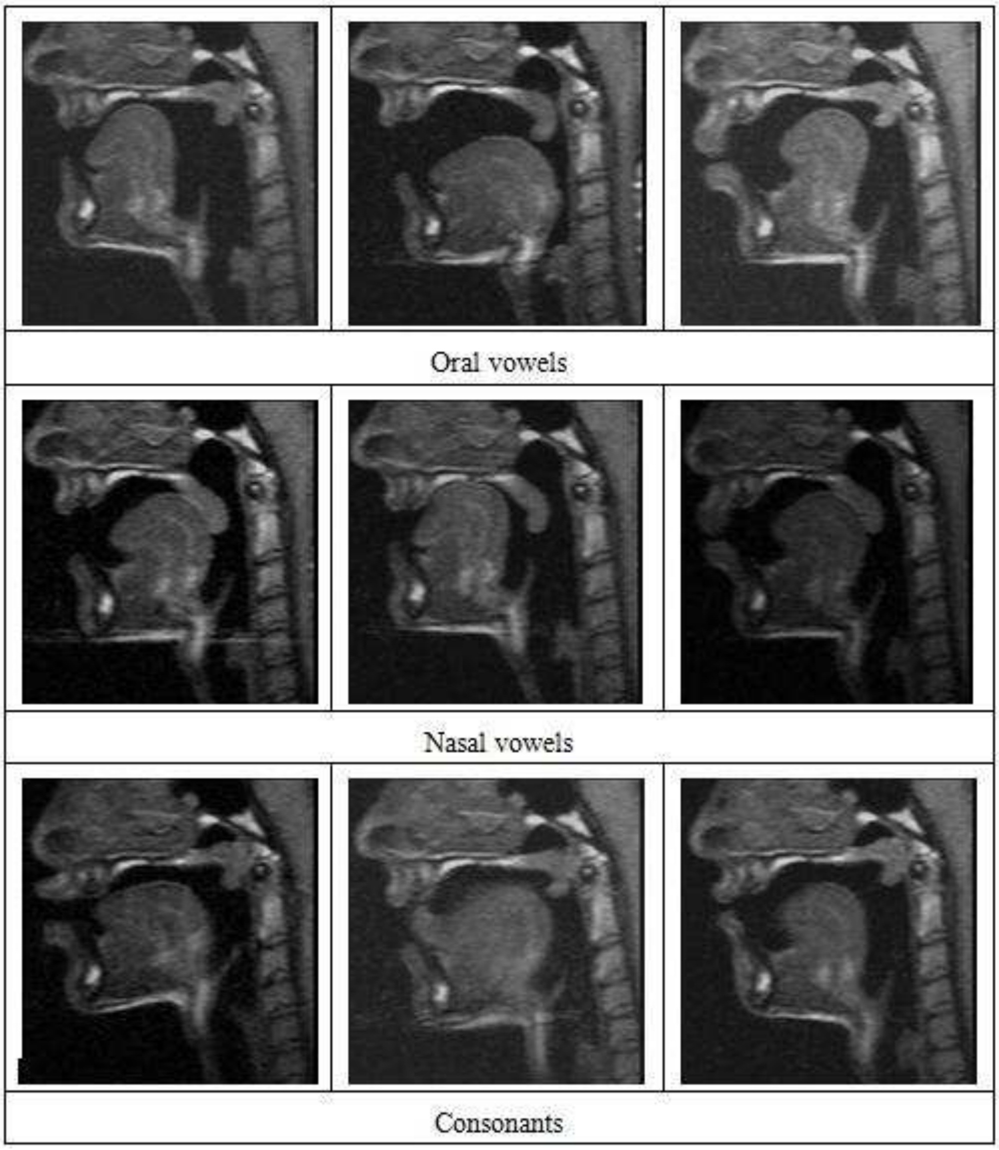


Figure 2

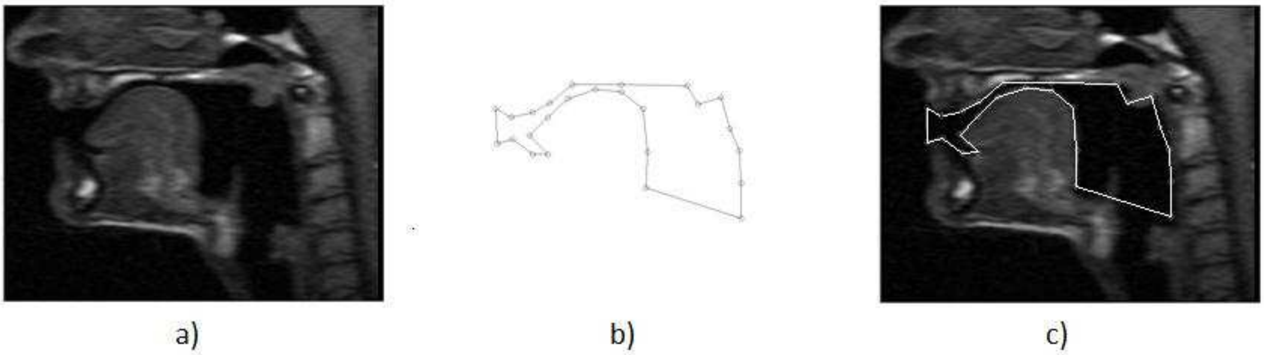
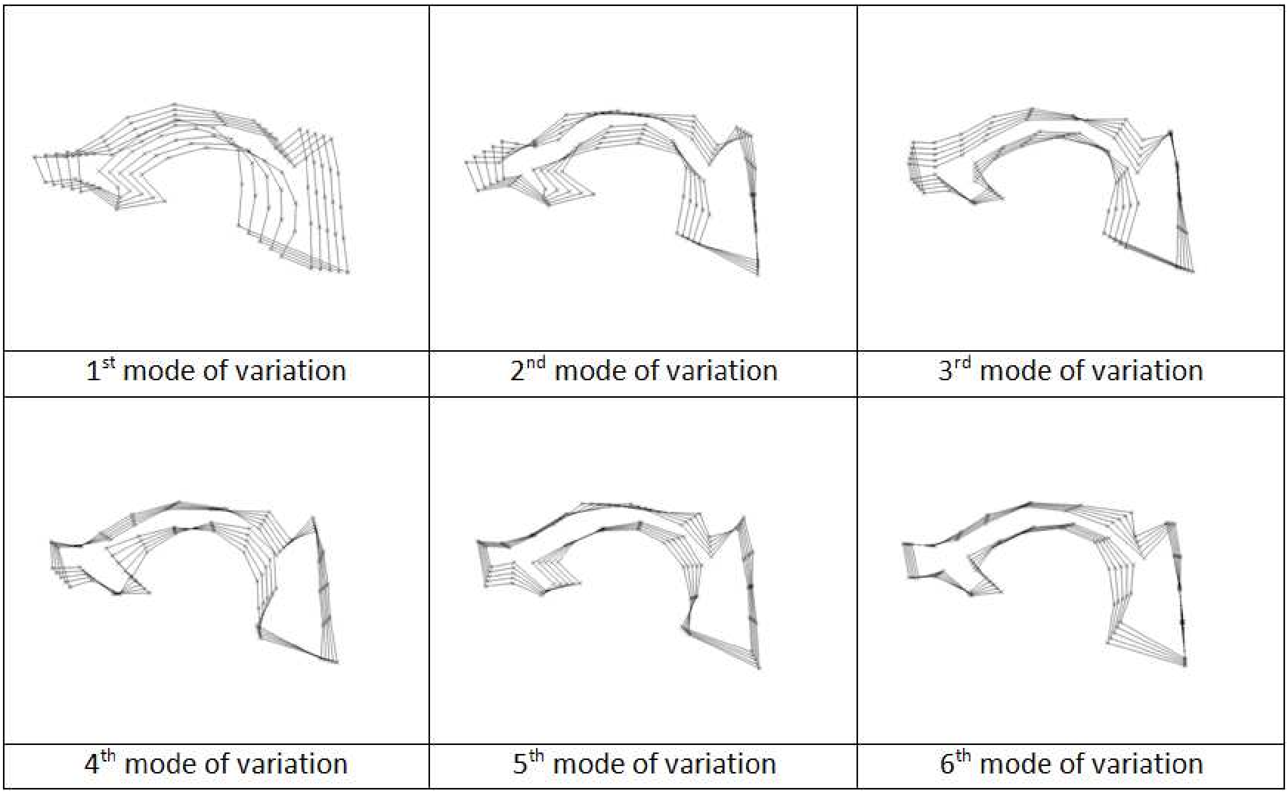

a)
b)
c)

Figure 3



1st mode of variation

2nd mode of variation

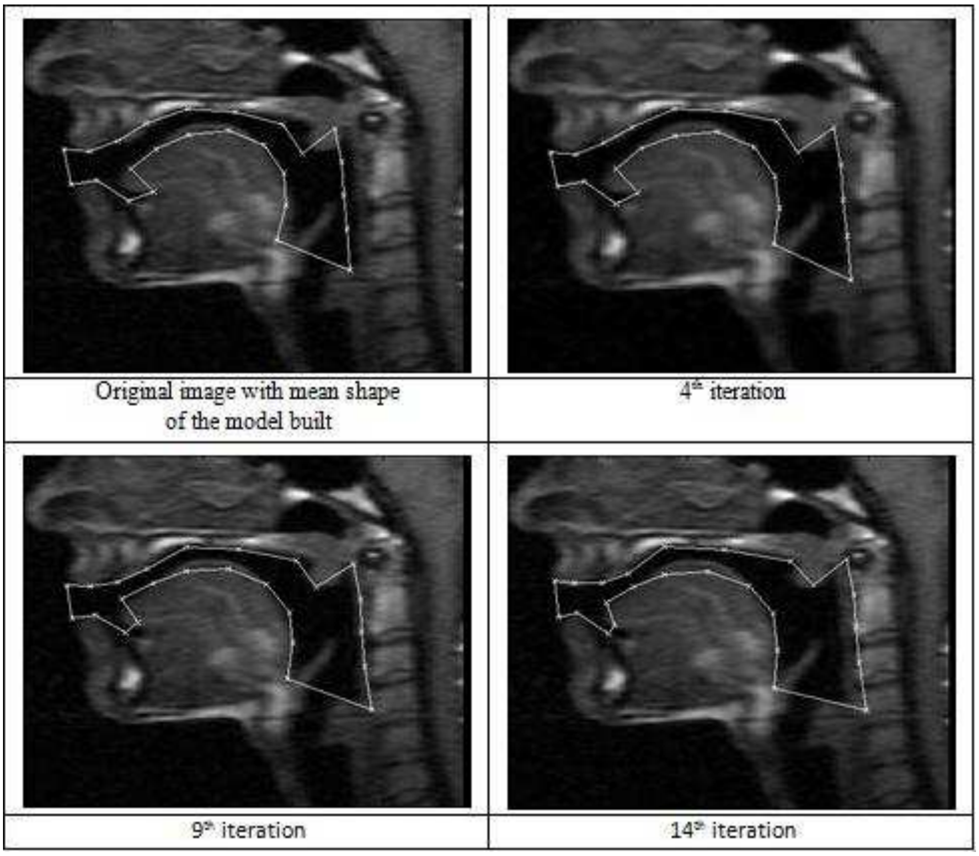
3rd mode of variation

4th mode of variation

5th mode of variation

6th mode of variation

Figure 4



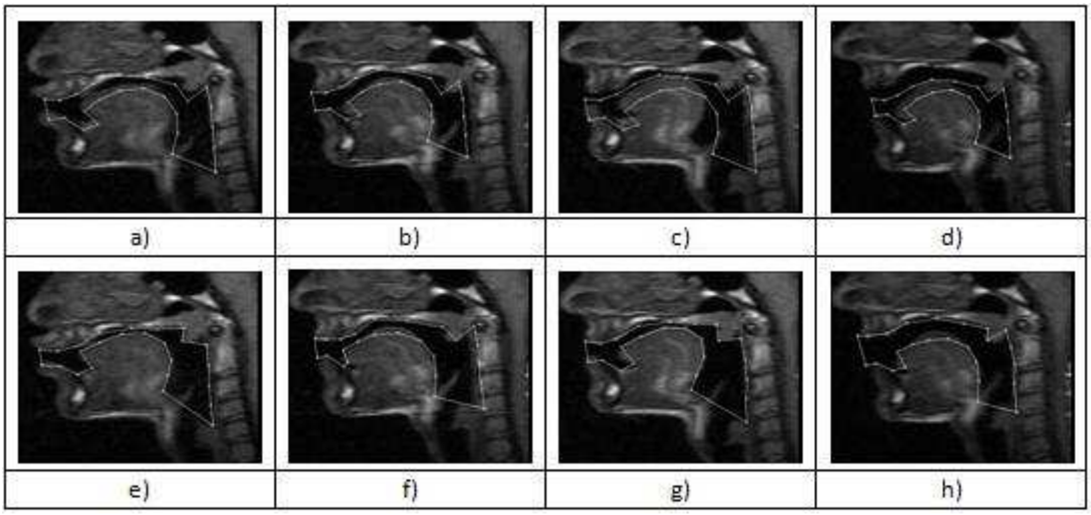
Original image with mean shape of the model built

4th iteration

9th iteration

14th iteration

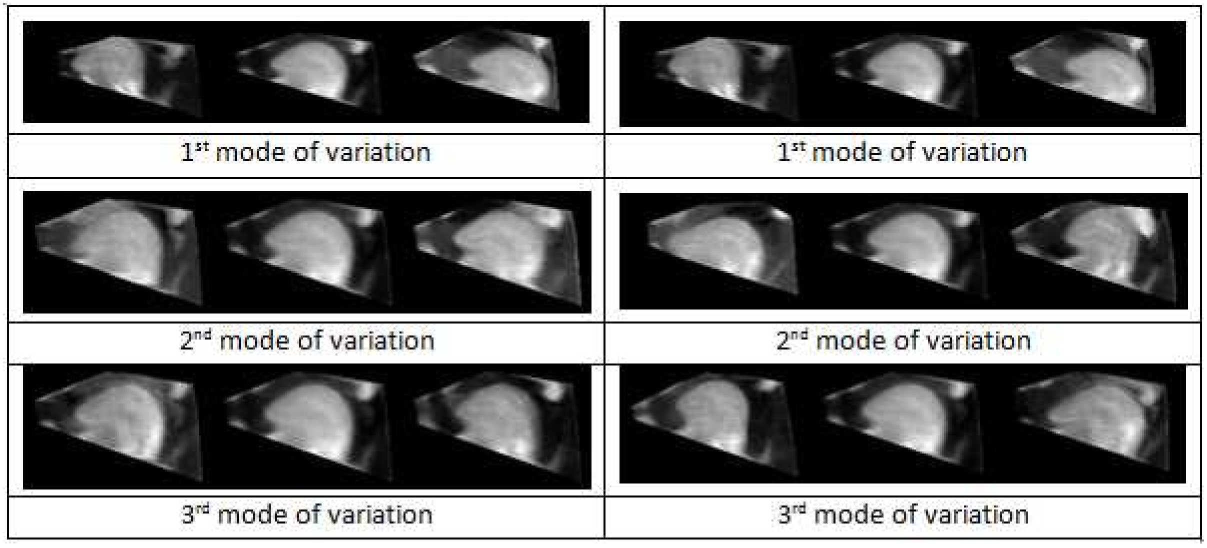
Figure 5



a) b) c) d)

e) f) g) h)

Figure 6

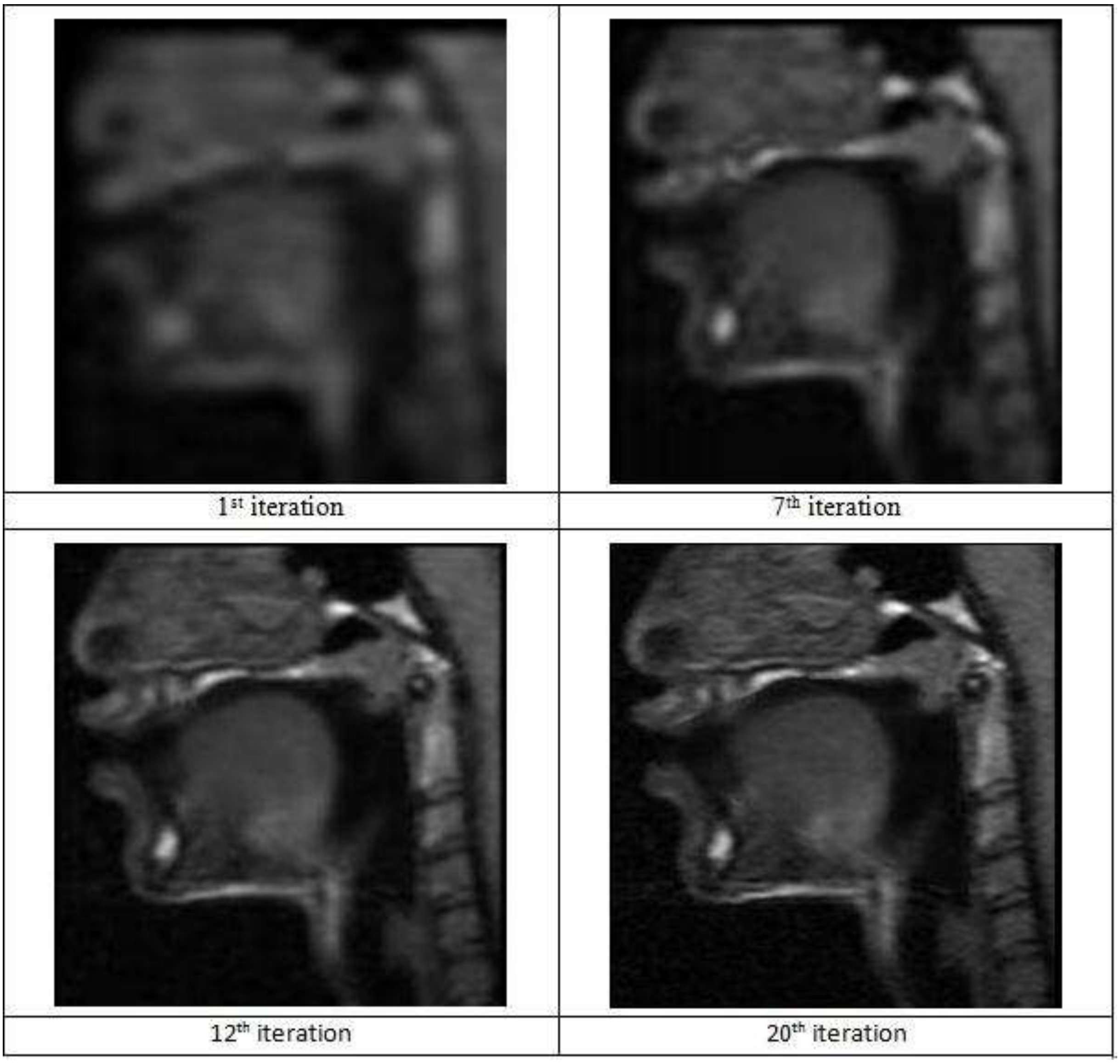


1st mode of variation | 1st mode of variation

2nd mode of variation | 2nd mode of variation

3rd mode of variation | 3rd mode of variation

Figure 7



1st iteration

7th iteration

12th iteration

20th iteration

Figure 8

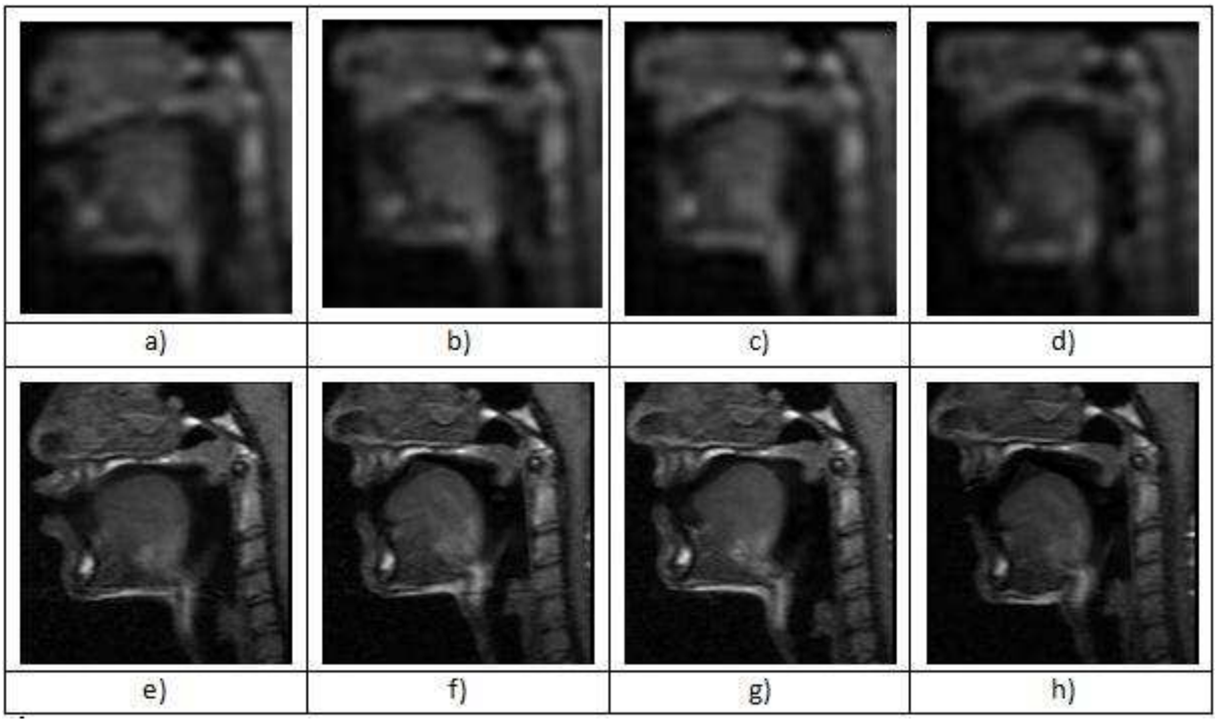


Figure 9

TABLE CAPTIONS

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

Table 2: Average and standard deviation errors of the segmentations obtained from testing images using the statistical models built.

TABLES

Table 1

Mode of variation	Retained Percentage	Cumulative Retained Percentage
λ_1	45.349%	43.349%
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}	0.979%	97.019%
λ_{12}	0.701%	97.720%
λ_{13}	0.507%	98.227%
λ_{14}	0.494%	98.721%
λ_{15}	0.396%	99.227%

Table 2

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