Segmentation of tongue shapes during vowel production in magnetic resonance images based on statistical modelling

Jessica C. Delmoral, MSc[1], Sandra M. Rua Ventura, PhD[2], João Manuel R.S. Tavares, PhD[3]

[1] Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal

e-mail: jessica.delmoral@fe.up.pt

[2] Centro de Estudos de Movimento e Atividade Humana, Escola Superior da Tecnologia de Saúde, Instituto Politécnico do Porto, Porto, Portugal

e-mail: sandra.rua@eu.ipp.pt

[3] Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal

e-mail: tavares@fe.up.pt

Corresponding author:

Prof. João Manuel R. S. Tavares

Faculdade de Engenharia da Universidade do Porto

Departamento de Engenharia Mecânica

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

Segmentation of Tongue shapes during Vowel Production in MR Images based on Statistical Modelling

Abstract

Quantification of the anatomic and functional aspects of the tongue is pertinent to analyse the mechanisms involved in speech production. Speech requires dynamic and complex articulation of the vocal tract organs, and the tongue is one of the main articulators during speech production. Magnetic Resonance (MR) imaging has been widely used in speech related studies. Moreover, the segmentation of such images of speech organs is required to extract reliable statistical data. However, standards solutions to analyse a large set of articulatory images have not yet been established. Therefore, this article presents an approach to segment the tongue in 2D MR images and statistically model the segmented tongue shapes. The proposed approach assesses the articulator morphology based on an Active Shape Model, which captures the shape variability of the tongue during speech production. To validate this new approach, a dataset of mid-sagittal MR images acquired from four subjects was used, and key aspects of the shape of the tongue during the vocal production of relevant European Portuguese (EP) vowels were evaluated.

Keywords: Medical Imaging; speech imaging; image analysis; image segmentation; image-based modelling.

1. Introduction

Voice and speech production are one of the most complex neuromuscular physiological functions of the human body. Speech is a dynamic process, which comprises air phonation through the glottis to generate sounds. These sounds are then modified by changes in the configuration of the vocal tract, and consequently different vowels and consonants are produced. The phenomenon entails that the shape of the vocal tract is altered by the dynamic shape variations of the structures that delimit it [1]. Among these structures, a key articulator is the tongue.

The tongue is an organ that is primarily composed of skeletal muscle tissue and it occupies the greater part of the oral cavity and oropharynx. The tongue plays a critical role in breathing, feeding and speech. It is posteriorly attached to the floor of the oral cavity, namely via tendons and other neighbouring muscles. Moreover, the tongue is a muscular hydrostat, i.e. an arrangement of incompressible agonist and antagonist muscles without any rigid structure for the muscles to act upon, making the mechanisms of its deformation even more challenging to understand [2].

To analyse the shape of the tongue and its articulatory movements during the production of different sounds is pertinent to extract speech information and thus be able to analyse the anatomic origin of speech disturbances. Speech therapists, require the analysis of speech-related anatomies through medical images in order to analyse speech articulation of vocal organs, such as the tongue. Furthermore, the quantification of tongue movements may also be used to provide information on how humans acquire new strategies for speaking tasks to compensate for losses in function caused by disease, surgical interventions and/or aging. Figure 1 shows the relevant anatomies related to the vocal tract during speech production on Magnetic Resonance (MR) image.

< Figure 1 should be around here >

The segmentation of vocal tract structures in medical images is therefore, highly important for quantitative analysis of speech dynamics. Quantitative studies require the processing and analysis of large datasets to retrieve meaningful information. However, many such segmentations

are carried out manually making therefore, the results susceptible to human reproducibility error. Furthermore, such manual segmentations are extremely time consuming, especially when tomographic or dynamic imaging modalities, such as MR or Ultrasound (US) imaging, are used as they generate huge amounts of image data. Thus, semi- or fully-automatic approaches suitable for the segmentation of images acquired during speech production are required in order to facilitate the tasks of professionals in these areas. The segmentation of the different shapes that the tongue assumes in MR images is required for the extraction of the articulatory anatomic configurations that characterize distinct speech sounds.

From a Computational Vision perspective, shape configuration is the key aspect in the analysis of the shape of speech structures. Therefore, the integration of *a priori* knowledge into the segmentation framework is appropriate. Statistical Shape Models (SSMs) have the ability to capture prior information about the shape of the object under study that can be used in the segmentation of the object. One of the most prominent approaches among SSMs is the Active Shape Model (ASM) proposed by Cootes et al. (1995).

In the present study, the potential of the ASM to segment the different tongue shapes in a set of MR images depicting speech articulations of European Portuguese (EP) sounds acquired under sustained phonation is evaluated. In addition, the viability of the statistical model built to capture the variability of the tongue shape in the same MR image dataset is analysed. The statistical data retrieved can be used to complement speech studies. Therefore, in this work, 18 MR images of 2 subjects and 9 EP sounds were used to build an ASM, and 11 MR images of 2 different subjects producing 6 of the 9 sounds used in the ASM building process were used to evaluate the segmentation results. These results were compared against manual annotations made by an expert, and the results confirmed that the ASM is a promising model to segment the human tongue in MR images during speech production, particularly if the original MR images are smoothed by applying a denoising filter. To the best of the authors' knowledge, this is the first study that explores the use of a denoising filter in order to improve the segmentation of the human tongue during speech production in MR images by an ASM.

2. Related Work

Extracting information related to the shape of the vocal tract and associated structures during speech production from MR images is a relatively new field of research. Image-based studies aiming at characterizing several languages phrases, and specific sounds have been reported in the literature [3–5]. The extensive research presented in the literature associated to the segmentation and modelling of the vocal tract, is mainly due to the relatively easy segmentation of the air/tissue boundaries of the vocal tract in MR images. For example, Ventura et al. (2012) proposed a morphological modelling method of the vocal tract to analyse speaker-specific movement patterns. Miller et al. (2014) presented a study on the morphological differences of pitch, related to the shape of vocal structures based on an ASM. The results assessed the mean behaviour and variability that characterises the conjoint movements occurring in the vocal tract structures.

The first tongue image-based reports used US imaging [6], and later on, X-Ray image-based analyses were presented [7]. Despite the advances in MR imaging technology regarding soft tissue contrast, which are now considered the state-of-the-art for studies regarding the vocal tract and related structures, the segmentation of the tongue is still a highly challenging task because of its location in close vicinity to other soft tissues, and therefore requires a higher soft tissue contrast and/or more competent segmentation algorithms.

Voice and speech related tongue studies are commonly focused on investigating the description of the articulatory dynamics and corresponding acoustic production [8] and on clinical research related to the complete physiology, neurophysiology, and structural interplay of the muscular hydrostat complex that is the tongue. The related literature includes studies on intensity-based segmentation [9–12] and statistical model-based segmentation [4,13].

The analysis of tongue shapes from MR images has been studied using manual annotations that were analysed based on principal component analysis (PCA) and mesh representations associated to specific sounds [14]. However, efforts towards the segmentation of the shape of the tongue through less user-dependent methods have also been proposed. Peng et al. (2010) presented a method based on active contours, using a PCA model built from a single patient, which was only able to partially segment the tongue contours, mainly the tongue dorsum. Later on,

Eryildirim and Berger complemented the previous method by adding coverage of the tongue root and frenulum to the segmentation result [15].

Also, statistical models have been used to investigate the shape of the tongue, modelling the shape variability via a parametric set of equations, and describing the statistical information of the deformations suffered by image-derived shape contours [16]. Moreover, ASMs have also proven useful in applications for other bio-structures, such as the brain [17] and heart [18].

The ASM is based on a Point Distribution Model (PDM) that compactly learns the space of plausible shapes of an object from a set of known shape contours, and a Profile Appearance Model (PAM) that captures the boundary appearance information variability in the corresponding training images. To the best of the authors' knowledge, the first study that applies a PDM to characterize the shape of the tongue and its movements in MR images acquired during speech production was presented by Vasconcelos et al. (2009). However, MR images are characterized by Additive White Gaussian Noise (AWGN) that makes visualization and segmentation difficult. In fact, intensity inconsistencies may be presented in MR images due to noise, which potentially compromises the adequate identification of the tongue boundaries, particularly when using computational segmentation algorithms. In the case of the ASM, these intensity inconsistencies can affect the ability of the PAM to search for the true boundaries of the object and therefore affect the efficient segmentation results. Consequently, the approach in this article includes a denoising step of the original MR images, which is followed by the segmentation task based on an ASM built from the training dataset [19]. Then, the ASM is guided towards the true boundaries of the tongue during the segmentation process based on more reliable image intensity and gradient information.

Compared to the studies found in the literature, the approach presented in this article has some advantages: (i) it allows successful 2D segmentations of tongue shapes in MR images independently of tongue size, which is feasible due to the analysis performed in the model coordinate space; (ii) it is based on a minimal user initialisation of the segmentation process that does not require advanced knowledge of the morphology of the vocal tract, mainly in relation to the tongue; (iii) the model built can be used in statistical studies of inter-speaker variability and to assess statistically speech impairment differences in tongue articulation.

3. Methods

3.1. Image Dataset

The images used here were acquired from a Siemens MAGNETOM Trio 3.0 Tesla (3.0T) system and head and neck array coils: a 32-channel head coil and a 4-channel neck matrix coil, respectively. The T1-weighted sagittal slices were obtained using Turbo Spin Echo Sequences with an acquisition duration of approximately 10.6 s and a thickness of 3 mm, according to the following parameters: a repetition time of 7.6 ms, an echo time of 2.87 ms, a flip angle = 5º, a square field of view (FOV) of 240 mm, a matrix size of 256x256 pixels, a resolution of 1.067 pixels per mm and a pixel spacing of 0.94 mm x 0.94 mm. Two male (denoted here as OM and AA) and two female (denoted in here as LF and IF) volunteers, aged between 30 and 47 years old (36 ± 6.59 years), with no record of speech disorders, were placed in a supine position. To allow intercommunication during image acquisition and to reduce MR acoustic noise, headphones were used. The speech corpus consisted of a set of 9 images per subject, during sustained articulations of 9 European Portuguese sounds, which consisted of 3 oral vowels in vowel sustention in two tones: normal and high pitch phonation, and consonant-vowel (CV) contexts: */pi/*, */pa/* and */pu/*.

The image acquisition process was designed and performed to obtain morphologic data covering the maximum range possible of the positions of the articulators in order to characterize and reconstruct EP speech sounds. Thus, the MR sagittal data was able to capture the main aspects of the shape and position of the different articulators involved, such as the tongue, lips and velum. The acquired images of the sound set represent the configurations in which the tongue assumes 9 stable and distinct positions in the oral cavity. Three examples of such positions are depicted in Figure 1.

To obtain a reliable ASM, the sounds to be used in the training process of the statistical model should adequately represent the variability of the shapes taken by the tongue. Additionally, each shape of the tongue presented in the training set should be described by a group of labelled landmark points that convey important features of the structure. Thus, the key articulation points that need to be identified in each training MR image are: the tip, which usually rests against the

incisors; the margin; the body; the dorsum, which has a convex shape that contacts with the palate; the inferior surface and the root.

3.2. Image pre-processing

MR images are characterized by noise that is highly dependent on the acquisition time. A suitable compromise between image quality and image acquisition time was considered in the present study. Additionally, in order to eliminate noise and enhance the boundaries, a Non-Local Means (NLM) denoising algorithm was applied to each MR image used in this study.

NLM is a nonlinear filter based on a weighted average of the image pixels inside a search window. The NLM method used in this study is an enhanced version of the original NLM algorithm as suggested by Tristán-Vega et al. (2012), which was tested with the dataset under study in order to establish a proper trade-off between noise removal and boundary preservation. Considering an empirical noise power (σ) equal to 0.1 and an attenuation correction parameter proportional to the differences between local neighbourhoods, the NLM method used here was applied to each original MR image of the dataset under study and presented an adequate computational processing time.

3.3. Point Distribution Model

The main objective of a statistical shape model is to describe statistically the shape variations of a non-rigid object in a representative training dataset, where each shape, which is a contour here, is defined by a set of points, which are commonly designated as landmarks. Hence, a PDM conveys the different shapes of an object from collections of landmarks that define the contours of the different shapes presented in the training dataset. Thus, given a set of K pairs $D = \{(I, c)^k\}_{k=1}^{|K}$, with each pair containing the training image I^k and the corresponding set of contour landmarks c^k, a PDM learns the mean shape of the object under study and the acceptable shape variations in relation to that same mean shape [21].

In this study, the shape of the tongue was statistically modelled by a PDM using a set of 18 MR images ($K = 18$). The manual annotation of the points in the training images requires a comprehensive knowledge of the object in question, as the behaviour of the resultant model depends on the suitability of these points. Hence, the manual selection of the points was carried out by one of the authors who is highly knowledgeable in medical imaging and in the anatomy of the oral cavity, thus providing confidence in the model under construction.

The methodology used to build the PDM is shown in Figure 2. The set of points was chosen to include a set of 16 relevant morphologic points of the tongue: two points on the lingual frenulum (anterior and posterior), one point on the tip, one point on the root, six points along the body, and six points along the inferior surface of the tongue. Figure 3 shows the aforementioned set of points on a MR image with a fictitious line connecting these points in order to visualize the anatomy of the object in question more easily.

< Figures 2 and 3 should be around here >

The manually annotated contours were converted into N evenly distributed points, defining each a k-th contour vector: $c^k = (x_1, x_2, \dots, x_N, y_1, y_2, \dots, y_N)$, where $c_n^k = \{x_n, y_n\}_{n=1}^{|N}$ are the coordinates of the n-th contour point. The set of all K contours is centred, using translation, rotation and scaling transformations, in order to minimize the sum of the squared distances between the points to the origin using Procrustes Analysis, followed by PCA applied to the matrix of the shape vectors:

$$M = ([x_{1:N}^1; y_{1:N}^1], \dots, [x_{1:N}^K; y_{1:N}^K])^T. \quad (1)$$

The PCA is performed in order to analyse the contours in the space of the shape variations. Thus, using Singular Value Decomposition (SVD), the covariance matrix, defined by $\frac{1}{2N} MM^T$, is obtained as:

$$\frac{1}{2N} MM^T = P \Sigma P^T, \quad (2)$$

where P is a matrix whose column vectors represent the set of orthogonal modes of shape variation, i.e. the eigenvectors, and Σ is a diagonal matrix containing the corresponding singular values, i.e. the eigenvalues. After the computation of the mean shape and covariance matrix of the model, equation (2) can be rearranged in order to compute the set of eigenvalues ($diag(\Sigma)$) and eigenvectors (P) that characterize each shape assumed by the object in the training set. This notation is transposed into the PDM notation, whereas any training and/or intermediate shape configuration x can be approximated as:

$$x \approx \bar{x} + P_s b_s, \tag{3}$$

where $\bar{x}$ represents the mean shape of the model, P_s the matrix of the corresponding eigenvectors, and b_s the eigenvalues vector. The basis of the PCA lies in the principle that the magnitude of each eigenvalue and the corresponding eigenvector has a proportional magnitude that explains the shape variability observed in the training set. Hence, the set of eigenvalues ($\mathrm{diag}(\Sigma)$) is organized in a descendant order of magnitude, and the model retains the most significant *i* eigenvalues according to:

$$b_s = \mathrm{diag}(\Sigma_{1:i}) : \frac{\sum_{k=1}^{i} \mathrm{diag}(\Sigma_k)}{\sum_{k=1}^{K} \mathrm{diag}(\Sigma_k)} \geq v\ \%. \tag{4}$$

Then, by ordinal correspondence, the corresponding eigenvectors P_s are selected, and the PDM is therefore, able to explain a known shape variance v, which is usually retained from 90 to 99.5%.

3.4. Profile Appearance Model

The Profile Appearance Model is defined based on the annotated contours using the corresponding intensity information. Hence, for each set of contours, the intensity information that characterizes the local context of the object's boundary is obtained from grey-level intensity profiles. Assuming that there is connectivity between the contour points, the perpendicular direction along the contour can be explored to find the boundary of the object. Then, along the perpendicular direction at each contour point, the image pixels are sampled using a fixed step size l, originating profiles of length $2l + 1$. Similar to the PDM building process, for each contour point c_n^k, of point n

and k-th training image I^k, an intensity profile g_n^k is extracted from each k-th image and their corresponding gradient information is organized into a matrix to which PCA is applied in order to calculate the PAM according to:

$$g \approx \bar{g} + P_g b_g, \tag{5}$$

where $\bar{g}$ denotes the mean intensity profile, b_g the most significant i eigenvalues (Eq. 4), and P_g the matrix of the corresponding i eigenvectors. Hence, the model obtained explains the intensities and texture variation of the points of the training contours.

3.5. Segmentation based on an Active Shape Model

A total of N = 64 interpolated points represented each training shape used to build the ASM in this study. The transformation that each of the ASM modelled contours underwent consisted in a translation, scaling and rotation. These pose transformations were applied in the segmentation of the modelled object in test images.

In order to use the developed semi-automated tongue segmentation approach, the user defines initially four points in the MR image to be segmented: the lowest point of the anterior wall of the tongue, the tip, the highest point of the dorsum and the root of the tongue, as shown in Figure 4. These points were chosen according to two criteria: (i) minimal user interaction and (ii) they are associated to anatomical points that can be easily identified. The defined points are then used in the initialization of the ASM that is built through the minimization of a Weighted Least Squares (WLS) fitting towards their equivalents in the shape model, modelling the key points in the model with weights equal to 1 (one). The following steps to segment the shape of the tongue in a test MR image are fully automatic.

< Figure 4 should be around here >

The segmentation technique used here is an iterative optimization technique for ASMs which allows initial estimates of pose, scale and shape of the modelled object to be adjusted in a test image. The stages of this iterative segmentation approach can be summarized into the following

steps: i) at each point of the model, the necessary movement to displace that point to a better position is calculated; ii) changes in the overall position, orientation and scale of the model that best satisfy the displacements are found; iii) finally, by calculating the required adjustments for the shape parameters, residual differences are used to deform the shape of the model towards the desired shape. The iterative segmentation methodology was implemented using a multi-resolution search approach. For the intensity profile model fitting process performed in step i), first it is necessary to define the best match regarding the image intensity pattern that together with all other profiles in the profiles matrix yields maximum matrix fit in each iteration of the process. The best profile match generates a new set of contour points x_{ip} that are then transformed by the PDM into a plausible tongue shape. Following the classical ASM procedure, starting from the mean shape of the model built, each point is moved along the direction perpendicular to the contour in order to minimize the residual distances between the new x_{ip} and previous x point locations:

$$\delta x = (x - x_{ip}). \tag{6}$$

The objective culminates in finding the new shape parameters b that minimize the residuals of the new intensity profile point positions x_{ip}, as $E(\delta x^T \delta x) = E(b_s + \delta b_s)$. Taking into account that the eigenvector matrix P is orthonormal, and considering the inverse PDM equation, this problem is simplified as:

$$\delta x \approx \bar{x} + P_s \delta b_s \;\Leftrightarrow\; \delta b_s = \left(P_s{}^T P_s\right)^{-1} P_s{}^T \delta x \;\Leftrightarrow\; \delta b_s = P^T \delta x. \tag{7}$$

The update of the shape parameters b_s is performed iteratively, until convergence.

4. Results

The proposed approach automatically builds Active Shape and Profile Appearance Models for the segmentation of the shape of the tongue in new MR images, i.e. in MR images not included in the training image dataset.

As already mentioned, the four initialization points in the MR image to be segmented were chosen to facilitate their identification by non-expert users, and their main use is to define the

geometrical limits of the tongue presented in the input MR image. A maximum of 25 iterations was allowed in the segmentation process, and 11 test images were segmented and analysed.

4.1. Pre-Processing

The denoised images presented clearer and homogeneous depictions of the oral cavity and tongue pixels with improved intensity distributions than those of the original images. The result of the denoising algorithm on an example image can be seen in Figure 5, which also shows the noise present in the original images. The noise content image, which was obtained by subtracting the denoised images from the originals, shows the eliminated intensity inhomogeneities present in different regions of the original image. The main purpose of the denoising algorithm was to smooth this noise effect. This smoothing led to a more efficient optimization of the profile intensity boundary search during the segmentation process, as is confirmed by the results presented in the next section.

< Figure 5 should be around here >

To assess the efficacy of the denoising algorithm used, the Mean Squared Error (MSE) and Peak Signal to Noise Ratio (PSNR) quality metrics were used. The results from these metrics evaluating the efficiency of the NLM algorithm are presented in Table I.

< Table I should be around here >

4.2. Active shape model-based segmentation

A total of 11 test MR images of 3 distinct EP speech sounds acquired from 2 different subjects were segmented by the proposed statistical based approach.

The suitability of key aspects of the ASM method in segmenting the tongue in MR images during speech production was analysed in terms of retained variance percentage, type of search and number of search profile pixels. Thus, two Active Shape Models were built, one with $v = 95\%$

and another one with $v = 99\%$ of retained variance, and search profiles of 7 and 17 pixels long were tested. The optimal intensity profile was 7 pixels long combined with the ASM of 99% of variance retention. The effect of variability captured by each of the eight modes of variation is illustrated in Figure 6.

< Figure 6 should be around here >

As aforementioned, the ASM built here was applied to segment 11 MR images not included in the training image dataset. Figure 7 presents the segmentations obtained for two of these MR images, which are associated to the simplest and the most complex shapes of the tongue under study (on the top and bottom rows, respectively). Hence, in each case of Figure 7, the results of the segmentation at the initial step, at the two intermediate steps and the final result, i.e. at the final step, are seen overlapped with the corresponding manual annotation. The accuracy of the computational segmentations achieved by the proposed approach was also quantitatively assessed based on pixel mean MSE and standard deviation relatively to the manual annotations made by an expert. The results found are shown in Table II.

< Figure 7 and Table II should be around here >

5. Discussion

Imaging technology available nowadays can meaningfully support the interpretation of muscle interactions of the tongue during both normal and disordered speech production. Most medical studies that involve ASMs in MR images are usually related to the localization and characterization of bones and organs; however, the current study concerns the human tongue. The proposed segmentation approach successfully segments the shape of the tongue in MR images acquired during speech production by combining a semi-automatic initialization approach with an Active Shape Model, using combined shape and appearance intensity and texture learning.

The quality of an ASM can be assessed by analysing the shape variance attained by varying the individual eigenvalues, also referred to as modes of variability, usually between -3σ to +3σ. A correctly built ASM comprises major shape variations in higher magnitude modes, i.e. higher eigenvalues. Hence, this analysis was performed by analysing the first most significant modes of variability.

The effects on varying the first 8 modes of variation of the model built are shown in Figure 6. The first mode comprises movements of the whole body of the tongue, mainly associated with the rotation of its shape, specifically with the forward and backward movements of the frenulum and tip presented in the training images, that have the greatest shape variations in these directions and influence the whole set of contour points. These changes are associated with the horizontal spreading in the production of the open front unrounded vowel [a] in Portuguese words like /casa/ (home).

The second variation mode is particularly associated with movements of the whole body of the tongue simultaneously along the vertical and horizontal axes. This captures the spreading and narrowing combined with upward and downward movements of the upper and posterior boundaries of the tongue presented in the training set. In the third variation mode, it is possible to observe changes in the movements of the upper section of the tongue. The former changes convey the narrowing of the upper posterior wall for the pronunciation of the close front unrounded vowel [i] in Portuguese words such as /riso/ (laughter). These varying shapes are complemented by a compensative inferoanterior movement of the frenulum section of the tongue. In the fourth variation mode, the movements of the curvature of the upper walls are captured and are complemented by a compensative inferoposterior movement. This shape conformation comes, for instance, from the production of the close back rounded vowel [u] from the Portuguese word /tu/ (you), which implies the upper posterior movement of the tongue dorsum and lower anterior movement of the base. Finally, the fifth and sixth variation modes captured subtle movements of the horizontal width of the lower and posterior portion of the tongue specifically, which complements the movements described by the third and fourth variation mode in minor pose details. The following modes of variance describe more subtle changes in the horizontal width.

From the results reported in Table II, it is possible to conclude that the NLM algorithm aided the model in finding the correct boundaries over the whole extent of the tongue. The intensity inconsistencies often hindered the model from finding the true boundaries of the tongue by moving towards high gradient magnitudes instead that were not part of the tongue contour; however, this can be avoided by denoising the original MR images. Table II also shows that the model with 99% of retained variability, obtained the best segmentation results.

Figure 7 shows that the rotation of the tongue tip and consequently the tongue body were successfully captured in the segmentation as well as the finer details of the different curvatures of the frenulum and dorsum. The intensity profile directions in the classical ASM are determined by the order in which the points are numbered; i.e., by the position of the previous and the following points relative to the current point. Nevertheless, the initialization process of the proposed segmentation approach, when done correctly, allows the model to adapt towards the correct boundaries.

ASMs for segmentation are very susceptible to noise and lack of boundary definition, which was successfully overcome by the pre-processing step included in the developed approach. The segmentation was also improved by the up-sampling of the contour points used in the definition of the shapes assumed by the tongue during speech production, along with the learning of the intensity and gradient information of the boundaries to be detected. These factors allowed the model to learn the edge context information located within the real boundaries, in order to converge towards the boundaries of interest in fewer iterations. Finally, adequate adaptation of the model built using new images was achieved by the shape prior knowledge imposed by this type of statistical model.

A similar analysis concerning the shape of the human tongue was made by Vasconcelos et al. (2009), who developed statistical shape models from a MR image dataset representative of EP oral vowels. In their study, the authors analysed the shape variability of the tongue of one male subject based on an ASM with 7 modes of variability, and the distribution of the first four modes of variation were similar to the ones obtained in this study [22]. In addition, Vasconcelos et al. (2012) presented an ASM to segment the vocal tract, and the contours obtained partly described the

upper and posterior sections of the surface of the tongue. Hence, the study conducted by these authors resulted in a shape statistical model that captured the shape of the tongue variability observed in the dataset used in the first mode of variation of the ASM built. The findings of this work suggest that the tongue is a central articulator in speech production, which indicates that the shape behaviour analysis of this articulator, as presented in this work, is important.

A quantitative comparison between the proposed method and other methods in the literature shows that our method is comparable to the one by Zhang et al. (2016). These authors indicated an average root-mean-square error (RMSE) of 0.74, which is comparable to the mean RMSE of 1.52 obtained using the proposed segmentation method. Peng et al. (2010) addressed the 2D segmentation of part of the tongue contour and their results are in accordance with ours. To the best of our knowledge, there are no other works in the literature that segment the complete 2D tongue contour. Therefore, our proposed method complements the existing studies by facilitating the successful 2D segmentation of the complete tongue boundary, including the inferior tongue walls. Moreover, the accuracy of the proposed method can be enhanced by enlarging the speech corpus used when building the ASM.

Nevertheless, the ASM used in this work was able to convey the statistical information of the shape of the human tongue from the MR images used here successfully, despite the large range of tongue shapes and anatomic sizes produced by the different subjects during EP speech production.

6. Conclusions

The ability of an Active Shape Model to properly convey the statistical information regarding the shape of the human tongue in MR images during speech production and to segment it in new images was assessed in this study. Hence, this work analysed the ability of a Point Distribution Model to capture the statistical variation of the shape of the tongue that characterizes the articulation of vocalic European Portuguese sounds. After which, an evaluation of the results of the segmentation of the tongue in new MR images acquired during EP speech production was performed. The results obtained confirm that the ASM used here is promising to achieve both

goals, i.e. convey statistical information about the shape of the tongue and its movements and to segment the tongue in new MR images acquired during speech production. However, for enhanced segmentation results, the input MR images should be properly denoised.

The approach in this work is useful in speech rehabilitation, namely, to recognize compensatory tongue movements during speech production in MR images. This knowledge is useful to understand speech production disorders in children, acquired speech impairments and speech impairment of oral cancer patient.

In future works, the proposed segmentation approach can be adapted to segment 2D dynamic MR image sequences and a larger image dataset should undoubtedly bring improvements. Statistical models, such as ASMs, have been shown to achieve good segmentation results when combined with robust initialization approaches, which can be achieved, for example, by using machine learning algorithms. Specifically, Marginal Space Deep Learning (MSDL) is an emerging technique used to align the mean shape of a model based on deep learning neural networks for object localization, and sequential estimation of the pose and scale parameters to be used in the ASM fitting [23]. Nevertheless, the proposed segmentation approach has proven to have potential as a tool for speech shape analysis; namely, for the evaluation of the shape of the tongue and movement patterns during speech production, as well as to improve the knowledge about the physiology of this organ that still needs to be further explored.

7. Acknowledgements

The images used in this work were acquired at the Radiology Department of the Hospital S. João, in Porto, Portugal, with the kind collaboration of Isabel Ramos (Professor at Faculdade de Medicina da Universidade do Porto and Department Director) and the technical staff.

The authors gratefully acknowledge the funding from Project NORTE-01-0145-FEDER-000022 - SciTech - Science and Technology for Competitive and Sustainable Industries, co-financed by “Programa Operacional Regional do Norte” (NORTE2020), through “Fundo Europeu de Desenvolvimento Regional” (FEDER).

8. References

1 Munhall KG. Functional imaging during speech production. *Acta Psychol (Amst)* 2001; **107**: 95–117.

2 Levine W, Torcaso C, Stone M. Controlling the Shape of a Muscular Hydrostat: A Tongue or Tentacle. In: P. Dayawansa W, Lindquist A, Zhou Y (eds). *New Directions and Applications in Control Theory. Lect. Notes Control, vol 321*. Springer-Berlin-Heidelberg, 2005, pp 207–222.

3 Ventura SMR, Vasconcelos MJM, Freitas DRS, Ramos IMAP, Tavares JMRS. Speaker-specific articulatory assessment and measurements during Portuguese speech production based on magnetic resonance images. In: LM. W (ed). *Language Acquisition*. Gazelle, 2012, pp 117–138.

4 Miller NA, Gregory JS, Aspden RM, Stollery PJ, Gilbert FJ. Using active shape modeling based on MRI to study morphologic and pitch-related functional changes affecting vocal structures and the airway. *J Voice* 2014; **28**: 554–564.

5 Martins P, Oliveira C, Silva S, Silva A, Teixeira A. Tongue segmentation from MRI images using ITK-SNAP: Preliminary evaluation. In: *International Conference Computer Graphics, Visualization, Computer Vision and Image Processing*. July, Rome, Italy, 2011, pp 3–10.

6 Sonies BC. Ultrasonic visualization of tongue motion during speech. *J Acoust Soc Am* 1981; **70**: 683.

7 Thimm G. Tracking articulators in X-ray movies of the vocal tract. In: *8th International Conference on Computer Analysis of Images and Patterns*. Springer, Berlin, Heidelberg, 1999, pp 126–133.

8 Havy M, Bouchon C, Nazzi T. Phonetic processing when learning words: The case of bilingual infants. *Int J Behav Dev* 2016; **40**: 41–52.

9 Takemoto H, Honda K, Masaki S, Shimada Y, Fujimoto I. Measurement of temporal changes in vocal tract area function from 3D cine-MRI data. *J Acoust Soc Am* 2006; **119**: 1037.

10 Peng T, Kerrien E, Berger MO. A shape-based framework to segmentation of tongue

contours from MRI data. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Dallas, TX, USA, March, 2010, pp 662–665.

11 Rua Ventura SM, Freitas DRS, Ramos IMAP, Tavares JMRS. Morphologic differences in the vocal tract resonance cavities of voice professionals: An MRI-based study. *J Voice* 2013; **27**: 132–140.

12 Zhang D, Yang M, Tao J, Wang Y, Liu B, Bukhari D. Extraction of tongue contour in real-time magnetic resonance imaging sequences. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China, March, 2016, pp 937–941.

13 Vasconcelos MJMMJ, Ventura SMRSM, Freitas DRSDR, Tavares JMRS. Inter-speaker speech variability assessment using statistical deformable models from 3.0 tesla magnetic resonance images. *Proc Inst Mech Eng Part H J Eng Med* 2012; **226**: 185–196.

14 Badin P, Serrurier A. Three-dimensional linear modeling of tongue: Articulatory data and models. In: *Seventh International Seminar on Speech Production, ISSP7*. Ubatuba SP, Brazil, UFMG, Belo Horizonte, Brazil, December, 2006, pp 395–402.

15 Eryildirim A, Berger MO. A guided approach for automatic segmentation and modeling of the vocal tract in MRI images. In: *19th European Signal Processing Conference*. Barcelona, Spain, September, 2011, pp 61–65.

16 Song C, Wei J, Fang Q, Liu S, Wang Y, Dang J. Tongue shape synthesis based on Active Shape Model. In: *International Symposium on Chinese Spoken Language Processing*. Kowloon, China, December, 2012, pp 383–386.

17 Ettaïeb S, Hamrouni K, Ruan S. Statistical models of shape and spatial relation-application to hippocampus segmentation. In: *International Conference on Computer Vision Theory and Applications, VISAPP*. Lisbon, Portugal, January, 2014, pp 448–455.

18 ElBaz MS, Fahmy AS. Active Shape Model with Inter-profile Modeling Paradigm for Cardiac Right Ventricle Segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*. Springer, Berlin, Heidelberg: Nice, France, October, 2012, pp 691–698.

19 Cootes TF, Taylor CJ, Cooper DH, Graham J. Active Shape Models-Their Training and Application. *Comput Vis Image Underst* 1995; **61**: 38–59.

20 Tristán-Vega A, García-Pérez V, Aja-Fernández S, Westin C-F. Efficient and robust nonlocal means denoising of MR data based on salient features matching. *Comput Methods Programs Biomed* 2012; **105**: 131–144.

21 Cootes TF, Taylor CJ. *Active Shape Models — 'Smart Snakes'*. Springer, London, 1992.

22 Vasconcelos MJM, Ventura SMR, Tavares JMRS, Freitas DRS. Analysis of tongue shape and motion in speech production using statistical modeling. In: Papadrakakis, Mojic M, Papadopoulos V (eds). *2nd South-East European Conference on Computational Mechanics*. Rhodes, Greece, June, 2009.

23 Ghesu FC, Krubasik E, Georgescu B, Singh V, Zheng Y, Hornegger J *et al.* Marginal Space Deep Learning: Efficient Architecture for Volumetric Image Parsing. *IEEE Trans Med Imaging* 2016; **35**: 1217–1228.

TABLE CAPTIONS

Table 1. The results of the quality metrics PSNR and MSE for noise reduction by the NLM algorithm on each training and test image.

Table 2. The mean squared (MSE), standard deviation (mean ± std) errors and Jaccard Similarity Index (JC) of tongue shapes segmented by the ASM model retaining 95% variability (ASM _95%_NLM), and the ASM model retaining 99% variability in each original test image (ASM_99%_original), and in each denoised test image (ASM _99%_NLM), relatively to the manual annotations.

FIGURE CAPTIONS

Figure 1. On the left, a MR mid-sagittal image (slice) indicating the vocal tract organs during a vowel production; On the right, MR images of a female under sustained phonation of the vowel utterances: /pa/, /pi/ and /pu/.

Figure 2. Methodology used to build a statistical shape model.

Figure 3. Initial set of landmark points manually defined on a MR image connected by line segments to facilitate their visualization: two lingual frenulum points (1-2), tongue tip (3), six points along the tongue body (4-9), tongue root (10), and six points along the inferior surface of the tongue (11-16).

Figure 4. Example of landmark reference points used in the initialization of the ASM model.

Figure 5. Denoising of an example image: on the left, original MR image; in the centre, the resultant image after the application of the used NLM denoising algorithm; on the right, the noise residuals present in the original image.

Figure 6. Effect on the tongue shape by varying ($\pm$ 3 std) each of the first 8 modes of variation (λi) of the PDM built with 99% of retained variability.

Figure 7. Segmentation results obtained in two MR test images: one from a male (top row) and the other from a female (bottom row). In each case, the initial shape of the model built, the model after the 5^{th} and 15^{th} segmentation iterations, and the final segmentation obtained (in blue) with the corresponding manual annotations (in red) are shown.

Table 1.

Training dataset		
Subject	PSNR	MSE
pu-LF	37.673	1.71E-04
a2-LF	37.285	1.87E-04
a-LF	37.479	1.79E-04
i-LF	37.585	1.74E-04
i2-LF	37.674	1.71E-04
u-LF	37.863	1.64E-04
pi-LF	35.451	2.85E-04
pa-LF	35.412	2.88E-04
a-OM	35.214	3.01E-04
a2-OM	35.096	3.09E-04
pa-OM	35.060	3.12E-04
pi-OM	35.072	3.11E-04
pu-OM	35.155	3.05E-04
i-OM	35.104	3.09E-04
i2-OM	35.089	3.10E-04
u-OM	35.125	3.07E-04
u2-OM	35.153	3.05E-04
Test dataset		
a-AA	40.273	9.39E-05
i-AA	39.964	1.01E-04
u-AA	39.985	1.00E-04
pa_AA	34.623	3.45E-04
pi_AA	34.652	3.43E-04
pu_AA	34.691	3.40E-04
pa_IF	35.299	2.95E-04
pi_IF	35.233	3.00E-04
pu_IF	35.243	2.99E-04
u_IF	37.974	1.59E-04
i_IF	35.543	2.79E-04

Table 2.

Image	ASM _95%_NLM		ASM_99%_original		ASM_99%_NLM	
	MSD	JC	MSD	JC	MSD	JC
AA_a	3.82±2.21	0.85	2.57 ±2.04	0.89	2.71±1.63	0.93
AA_i	6.58±8.09	0.75	4.70±2.49	0.83	4.81±1.32	0.81
AA_u	3.21±2.62	0.87	3.44±1.83	0.87	1.21±3.32	0.96
AA_/pa/	4.23±3.36	0.78	3.56±2.06	0.86	3.55±1.02	0.89
AA_/pu/	5.52±4.85	0.76	2.31±2.3	0.88	1.54±2.78	0.96
AA_/pi/	1.98±1.34	0.90	4.34±2.61	0.84	1.55±2.78	0.95
IF_/pa/	3.65±4.20	0.84	5.11±6.50	0.78	2.94±3.24	0.92
IF_/pi/	7.98±1.34	0.72	2.51±3.21	0.87	1.54±2.00	0.97
IF_/pu/	4.87±5.40	0.80	2.41±1.14	0.89	2.32±3.45	0.92
IF _u	5.02±1.03	0.78	1.9±0.11	0.93	2.02±1.43	0.95
IF _i	2.04±3.3	0.87	7.02±4.09	0.77	4.27±3.12	0.81

FIGURES

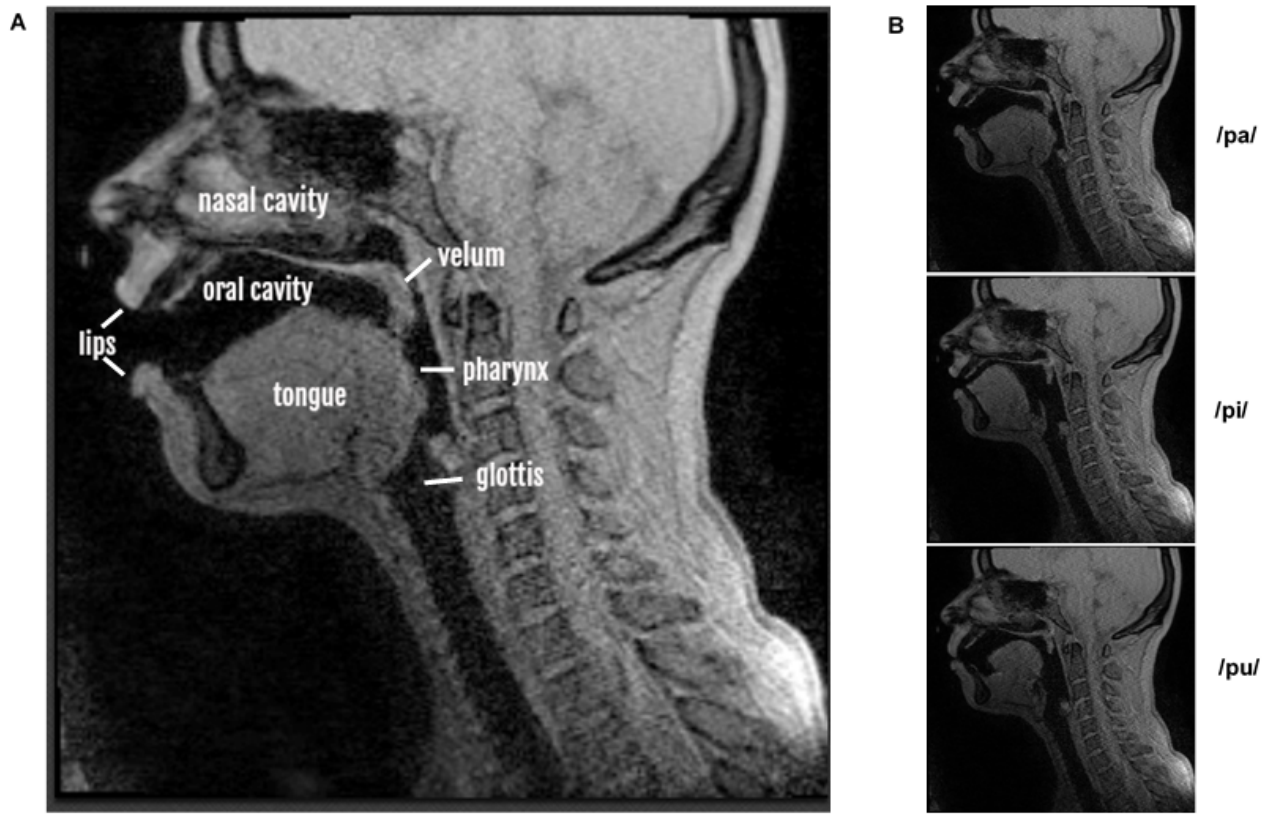

Figure 1

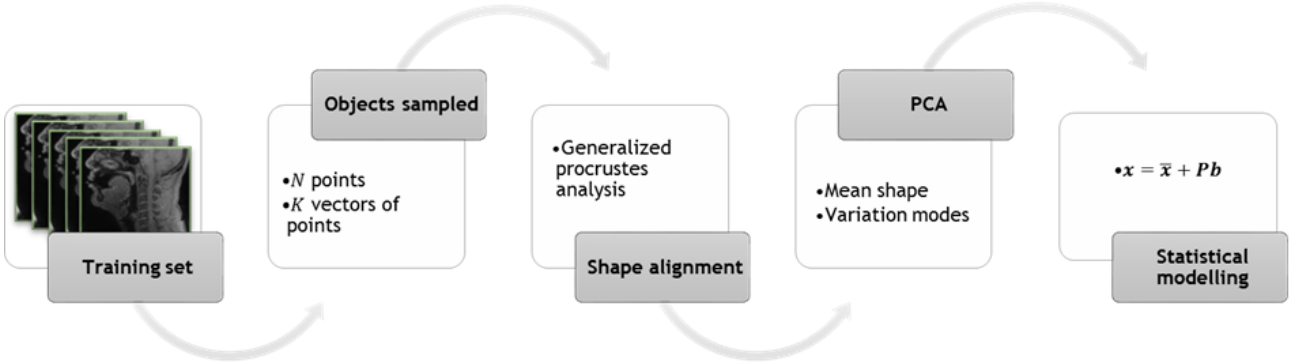

Figure 2

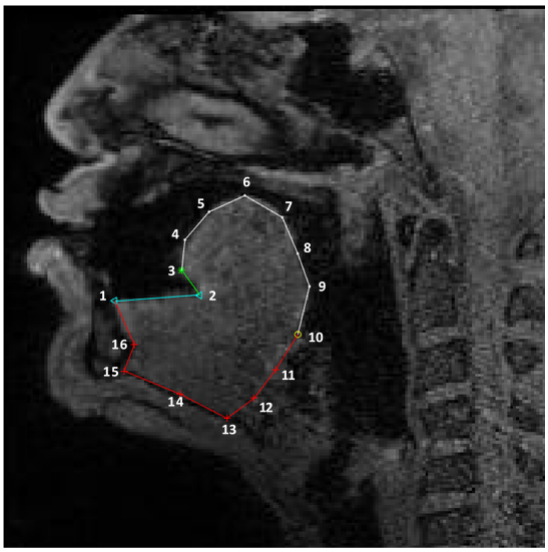

Figure 3

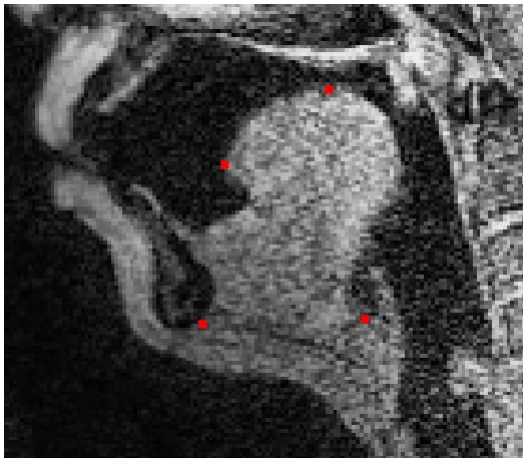

Figure 4

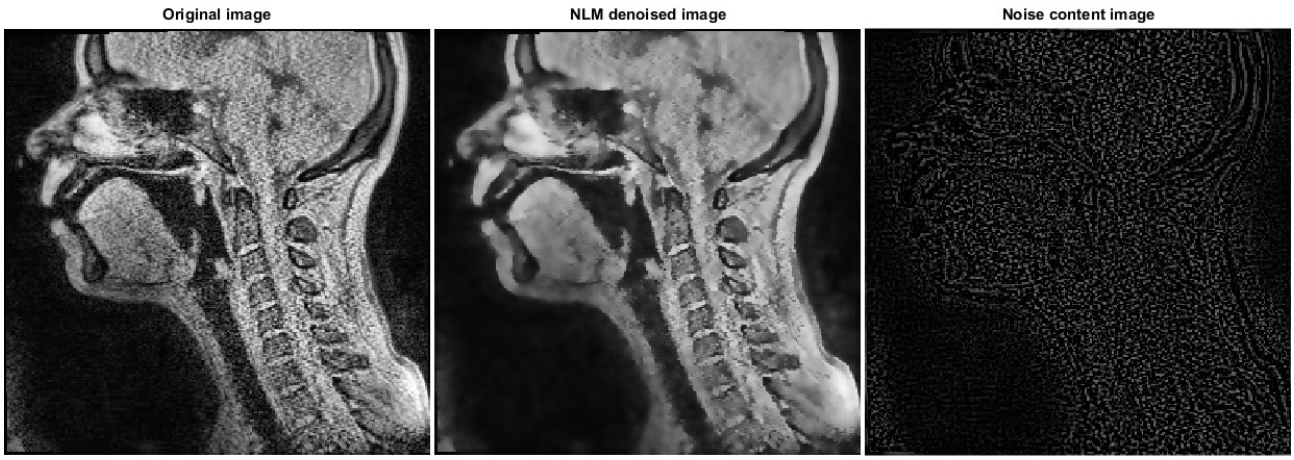

Figure 5

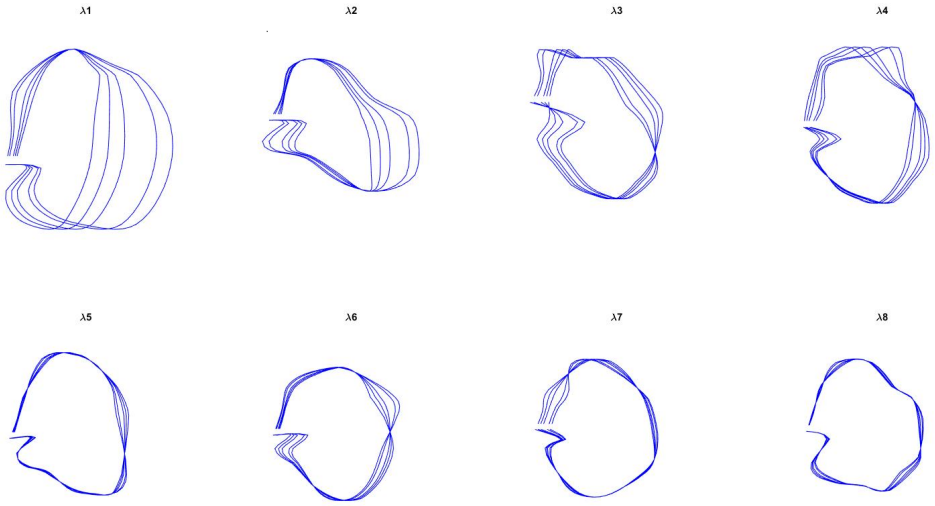

Figure 6

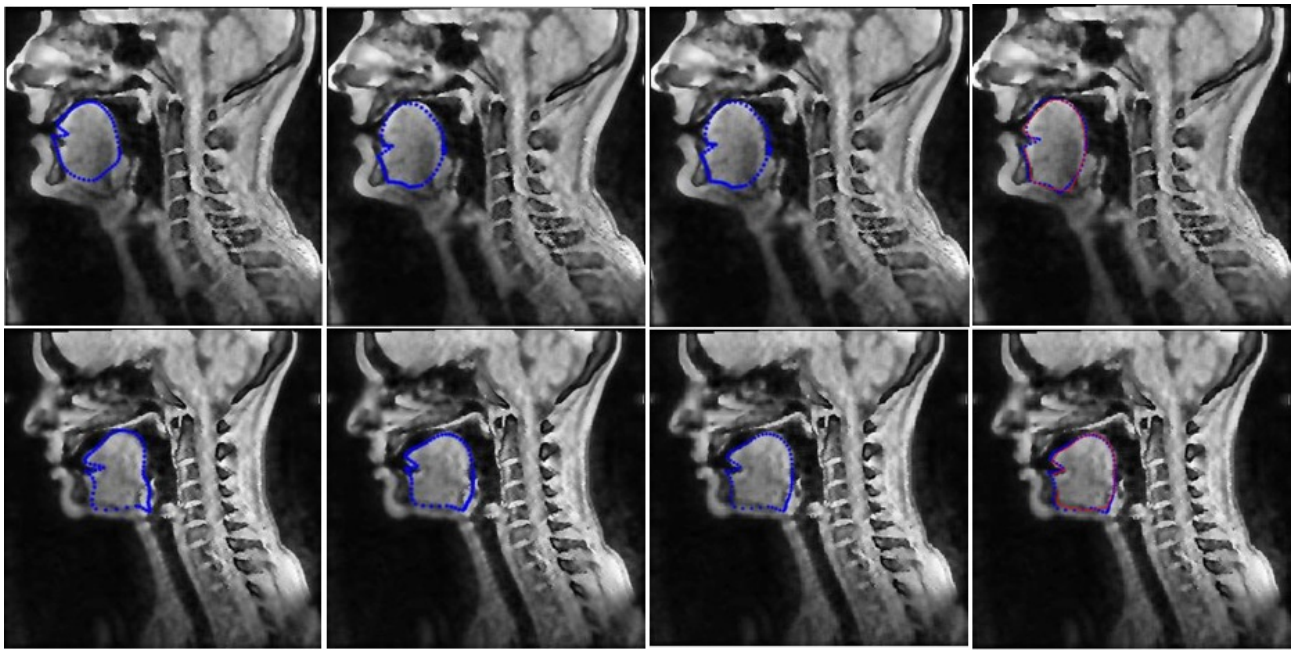

Figure 7