3D segmentation of the tongue in MRI: a minimally interactive model-based approach

Negar M. Harandi*, Rafeef Abugharbieh and Sidney Fels

Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada

(Received 7 June 2013; accepted 6 November 2013)

Static magnetic resonance imaging partially resolves soft tissue details of the oropharynx, which are crucial in swallowing and speech studies. However, delineation of tongue tissue remains a challenge due to the lack of definitive boundary features. In this article, we propose a minimally interactive inter-subject mesh-to-image registration scheme to tackle 3D segmentation of the human tongue from MRI volumes. A tongue surface-mesh is first initialised using an exemplar expert-delineated template, which is then refined based on local intensity similarities between the source and target volumes. A shape-matching technique [Gilles B, Pai D. 2008. Fast musculoskeletal registration based on shape matching. Paper presented at: MICCAI 2008. Proceedings of the 11th International Conference on Medical Image Computing and Computer Assisted Intervention; New York, NY, USA] is applied for regularising the deformation. We enable effective minimal user interaction by incorporating additional boundary labels in areas where automatic segmentation is deemed inadequate. We validate our method on 18 normal subjects using expert manual delineation as the ground truth. Results indicate an average segmentation accuracy of overlap of 90.4 ± 0.4% and distance of 2 ± 0.2 mm, achieved within an expert interaction time of 2 ± 1 min per volume.

Keywords: 3D interactive segmentation; model-based registration; MRI; tongue; shape matching

1. Introduction

Speech, chewing and swallowing are critical and complex neuromuscular functions. Various associated disorders result in medical complications that, if not properly treated, may significantly degrade the quality of life of those afflicted (Logemann 1994). The tongue is the primary organ in the oropharynx and plays an essential role in oropharyngeal functions. It consists of interwoven muscle fibres that undergo a wide range of muscular contractions and relaxations whose exact timings and levels of activation are still unknown (Steele and Lieshout 2009).

Computer-aided modelling and simulation of the oropharyngeal structures are beneficial for 3D visualisation, and for the understanding of the associated physiology. Generic biomechanical models of the oral, pharyngeal and laryngeal structures are adopted into the ArtiSynth framework (Lloyd et al. 2012). Further expansion of this generic platform, to encompass individualised information, will allow investigation of the subject-specific variability in the morphology and physiology of the region. It will also facilitate future development of the patient-specific platform, in medical settings, which will be aimed at assisting in the diagnosis and treatment planning of speech and swallowing disorders as suggested by Vranckx and Delaere (2012). Existing biomechanical models of the human tongue (Gerard et al. 2006; Buchaillard et al. 2009; Fang et al. 2009; Sonomura et al. 2011) are generic and do not provide any individualised information. They have been hand sculpted, simplified and made symmetrical to meet the requirements of the simulation platforms. Moreover, their generation workflow is highly manual, tedious, non-trivial and, hence, not suitable for creating subject-specific models. Further automation of modelling and simulation procedures is in high demand, and essential for overcoming the associated prohibitive costs (Neal and Kerckhoffs 2010).

While dynamic imaging can provide insight into the individual physiology of the tongue, its detailed biology should be inspected with static MRI. High-resolution MRI volumes require a long acquisition time, which leads to involuntary movement of the tongue and, hence, introduces severe motion artefacts (Plenge et al. 2012). Two-dimensional acquisition can provide a refined depiction of the tongue in the acquired plane, but the through-plane resolution is low, and inadequate for most of the volumetric analyses. Some previous speech studies (Badin et al. 2002; Engwall 2003; Takano and Honda 2007) adjusted the orientation of the acquisition plane to be orthogonal to the axis of the vocal tract, in order to facilitate the tongue modelling (see Figure 1). Then, 2D tongue contours are manually segmented in each slice and registered to the volume.

In a different approach, super-resolution reconstruction techniques have been introduced to generate isotropic MRI volumes from orthogonal slice stacks acquired sequentially (Peled and Yeshurun 2001; Bai et al. 2004). First, the imaging process is formulated as an observation model.

*Corresponding author. Email: negarm@ece.ubc.ca
Next, the intensity value of each pixel in the volume is estimated, using an optimisation method such as maximum a posteriori (Bai et al. 2004) or least square (LS) (Plenge et al. 2012) iteratively. Recently, Woo et al. (2012) applied an edge-preserving data combination technique based on Markov random field, to build super-resolution volumes of the human tongue. Isotropic resolution of 0.94 mm was reported. Figure 2 shows the reconstruction result, in comparison with the in-plane high-resolution data.

The aforementioned volumetric MRI provides an opportunity for generation of 3D subject-specific biomechanical models of the tongue. Finite element (FE) analysis is the most popular numerical method for simulation of soft-tissue deformation based on the physics of continuum mechanics (Nealen et al. 2006). Few meshing algorithms have been proposed in order to directly generate an FE mesh from a data-set of stacked images. The popular voxel-based method (Keyak et al. 1990) identifies the corresponding voxels by thresholding, before transforming them into cubic elements in the final mesh. The method is fully automated, general and robust; however, it lacks the efficient descriptors for identifying the soft tissues in MRI. Other meshing algorithms rely either on organ contours (Teo et al. 2007) or on the surface mesh (Baghdadi et al. 2005; Bucki et al. 2010), extracted beforehand from the image stacks or volume. Therefore, the modelling process consists in two consecutive phases of segmentation, and FE volume mesh generation. The overall clinical usefulness is highly dependant on the accuracy and degree of automation in each phase.

2. Related work

Accurate delineation of the tongue from MRI remains a challenge, due to the lack of definitive boundary features separating many of the adjacent soft tissues. Manual segmentation produces accurate results, but is prohibitively time consuming and tedious. General-purpose interactive tools can ease the task; however, they still require significant amounts of user interaction. Prior knowledge of shape and appearance, if incorporated effectively, can significantly facilitate dealing with soft-tissue inhomogeneities, noise and loss of contrast in medical images (Heimann and Meinzer 2009). In this regard, shape constraints have been embedded into a level-set framework (Leventon et al. 2000; Tsai et al. 2003; Foulonneau et al. 2009), further equipped with trained
distance maps (Bresson et al. 2006). In addition, statistical methods such as the active shape models (Cootes et al. 1995) have been widely explored to encompass intra- and inter-subject morphological discrepancies. Here, the cardinality of the training set is proportional to the degree of natural variability of the organ shape. For example, Heimann et al. (2007) selected 32 data-sets out of 86 to train their 3D reference model of 2562 equally distributed landmarks, used for segmentation of the liver in CT volumes. Unfortunately, such a large data-set is not currently available for tongue MRI volumes.

As an alternative to statistical methods, prior information may also be formulated in form of a single template, registering to the target image. Saddi et al. (2007) used template-matching procedure as the complementary step to the liver segmentation process, in order to compensate for the limitations of the learning set. Somphone et al. (2008) transformed their binary template, subject to conformity constraints between local patches. In a different approach, Gilles and Pai (2008) used explicit shape representation of the template to segment the musculoskeletal structures out of MR images. The mesh deformation was regularised based on an expanded version of a computer animation technique called shape matching (Muller et al. 2005). The method was shown to efficiently approximate large soft-tissue elastic deformations. For a review on deformable medical image registration, we refer to the detailed survey article by Sotiras et al. (2012).

Despite successful use of prior shape, automatic segmentation is still challenging in low-contrast medical images of the soft tissue. Clinical applications demand an expert’s supervision and control over the process and final result of the segmentation. The effective and minimal interactivity schemes will provide higher reliability, while keeping the cost of interaction reasonable. Freedman and Zhang (2005) combined prior shape and interactivity in a graph cut framework in 2D. Recently, Mory et al. (2012) incorporated user input as inside/outside labelled points to improve the robustness and accuracy of a non-rigid implicit template deformation.

Quite recently, Lee et al. (2013) proposed a semi-automatic seeding method based on random walker algorithm (Grady 2006) for 3D segmentation of the tongue in low-resolution dynamic MRI, where tongue has a uniform intensity. Their proposed method is useful for speech analysis, but random walker requires excessive amount of user interaction for accurate segmentation of a single high-resolution static MRI volume. Other reported works on segmentation of the oropharyngeal structures from MRI data focused on 2D slices. Bresch and Narayanan (2009) proposed an unsupervised regional technique, performed in the frequency domain, to capture the shape of the vocal tract. Peng et al. (2010) used a shape-based variational framework for tracking the tongue contour at its surface in midsagittal dynamic MRI. Eryildirim and Berger (2011) managed to include physically corresponding surface curves of the tongue in their previously introduced Principal Component Analysis-guided method. Although these methods provide valuable information for speech studies, they ignore delineation of the tongue at its base and its contact with the epiglottis, hyoid bone and salivary glands. Segmentation algorithms tend to fail in these areas due to the fusion of the tongue into its neighbouring tissues of similar composition. In addition, as mentioned above, 3D reconstruction of the tongue shape from its sparse 2D-segmented contours is not straightforward.

In this article, we propose a real-time, force-based user interaction platform coupled with a mesh-to-image registration technique as proposed by Gilles and Pai (2008). The proposed method expands the application of such methodology from musculoskeletal structures to highly deformable soft tissues, such as the tongue. The proposed method extends the state-of-the-art by considering delineation of the tongue from the epiglottis, hyoid bone and salivary glands in high-resolution static MRI. Furthermore, we facilitate the transition to a volumetric model by tailoring our segmentation method to use reconstructed MRI volumes over 2D images.

3. Methods

We propose an interactive mesh-to-image registration framework to tackle 3D tongue segmentation from MRI volumes. Both prior shape and intensity are incorporated in the form of a source image volume and its corresponding surface mesh, which is delineated by a dental expert. The choice of the source data-set is arbitrary. We use a discrete surface mesh representation to deal with the regularity and shape constraints. The overall pipeline of the proposed method is shown in Figure 3. The current position of the surface nodes is stored in the Mechanical State module. Loop 1 includes the modules that handle the mesh-to-volume registration. The mesh is deformed according to local intensity similarity between the source and the target volumes. The deformation is regularised using an extended version of shape matching (Gilles and Pai 2008). In loop 2, we deploy an effective minimal user interaction mechanism to help attain higher clinical acceptance. Both loops shown in Figure 3 have access to and are able to update the mechanical state of the mesh simultaneously. This provides real-time visualisation of the surface evolution.

Our proposed method was fully implemented under the simulation open framework architecture (SOFA @www.sofa-framework.org), an open-source modular framework based on C++. This allows for the registration algorithm to be interpreted as a real-time simulation process during which the source model iteratively deforms to match the target configuration starting from its initial position.
3.1 Mesh-to-volume registration

Mesh-to-volume registration is handled through applying the external and internal forces to the vertices of the mesh. The external force, \( F^e(t) \), steers the mesh towards the target boundaries. The internal force, \( F^i(t) \), keeps the mesh regular and close to the prior shape. Instead of summing up the internal and external forces, we use a pair-and-smooth approach, originally proposed by Cachier and Ayache (2001), where the external and internal forces are applied sequentially. This will minimise propagation of noise from image features to the final result. Let \( x(t) \) denote the vector of positions for all the vertices on the surface, at time \( t \). We also define \( x' = x(0) \) as the reference vertex positions. The mesh, as shown in loop 1 of Figure 3, is deformed in two steps:

1. **Intensity profile registration**: The image-based external force, \( F^e(t) \), is calculated for each node as described later in this section; then, the position of each vertex is augmented with its corresponding external force. To augment the positions, the simple relaxation scheme of gradient descent is used with unitary time step: \( x(t) + F^e(t) \).

2. **Shape-matching regularisation**: A smoothing internal force is applied on the augmented position of the vertices, which results into the vector of the regularised goal positions defined as \( \hat{x} \). The details of the smoothing process follow later in this section.

### Intensity profile registration

For each node at position \( x \) on the surface, the external force at time \( t \) is calculated by

\[
F^e(t) = \alpha_e (x' - x(t))
\]

where \( \alpha_e \) is the stiffness and \( x' \) denotes the new location of the node. The search for \( x' \) is performed within a pre-defined range of inward and outward steps at the direction normal to the surface. At each iteration, \( x' \) is selected to be the point which maximises a local similarity measure between the source and target image volumes. Our algorithm matches the 1D gradient intensity profiles of pre-defined length, \( L \), in the direction normal to the surface. Let \( G_{tar}(x) \) be the gradient profile of the target image centred at point \( x \), and \( G_{src} \) denote the corresponding gradient profile in the source image. The optimum value of \( x \) in each time step, denoted by \( x' \), is calculated using the normalised cross-correlation as the similarity metric:

\[
x' = \arg\max_x \left\{ \frac{G_{tar}(x) - G_{tar}^G}{\|G_{tar} - G_{tar}^G\|} \cdot \frac{G_{src} - G_{src}^G}{\|G_{src} - G_{src}^G\|} \right\}
\]

where \( G \) is the average value of \( G \) and \( \langle \cdot \rangle \) and \( \| \cdot \| \) denote the inner product and \( L^2 \) norm, respectively.

### Shape-matching regularisation

To regularise the mesh deformation, we apply the extended version of the shape-matching algorithm previously introduced by Gilles and Pai (2008) in the context of musculoskeletal structures. The underlying mesh is subdivided into overlapping clusters of nodes, defined around each vertex, \( i \), on the surface. The cluster for vertex, \( i \), is defined as

\[
\zeta_i = \{ j : d(x_i, x_j) < s \}
\]

where \( d \) is the Euclidean distance and \( s \) is the predefined cluster size (or radius). Then, for each cluster \( \zeta_i \), the algorithm approximates the local deformation of the nodes with a rigid transform \( T_i \), applied on the central reference position. The LS estimation of \( T_i \) is obtained by

\[
T_i = \arg\min_{T} \sum_{j \in \zeta_i} m_j \left\| T x'_j - x_j - F^i \right\|
\]

where \( m_j \) represents the mass weight of each particle in the cluster. This in turn will update the goal position of each node in the cluster to be \( \hat{x} = Tx' \).

Due to the overlapping nature of the clusters, each vertex may obtain different goal positions from the
different clusters it may belong to, which is subsequently combined into an average position. The final goal position is used to calculate the corresponding internal forces which are then averaged and applied to all the vertices of each cluster. Here, shape matching acts as an elastic force that is proportional to the strain, whereas updating the reference positions at each time step will simulate plastic deformations.

For the initialisation mode, we model the underlying mesh with one cluster. Hence, the bodily movement of the mesh would be purely rigid, containing three translational and three rotational degrees of freedom (DOFs). If desired, we enable the user to guide the initialisation towards what he/she may deem as a better position by simple mouse-click and drag. This will insert a spring force from the mesh towards the cursor. This initialisation scheme compensates for large displacements between the initial and final tongue positions (see Figure 4). At any time, the user can make the transition to the deformation mode by increasing the number of clusters, through entering the desired number in a dialogue box and pressing a button.

3.2 User interaction

We incorporate an effective minimal user interaction mechanism to guarantee a satisfactory result to the end user. The procedure is shown in loop 2 in Figure 3. At any time during the registration process, the user is free to inspect the orthogonal cut-planes of the deforming mesh, overlaid on the corresponding 2D sections of the target image. The user may provide additional boundary labels by simple clicking in any area where automatic segmentation is deemed inadequate.

Since boundary constraints are handled through forces in SOFA, our interaction is also force based. As soon as a new boundary voxel is clicked, the algorithm searches for the closest surface node on the mesh and inserts a spring force between these two points. The closest points on the surface will get updated in each iteration of the mesh deformation. We empirically used a predefined stiffness of about $10^4$ in all implementations. Stiffness values of higher order of magnitude may cause instability and hence are avoided.

4. Materials

We used MRI scans of 18 normal subjects. For each data-set, three sagittal, coronal and axial stacks of MRI slices were acquired with the tongue at the rest position using a T2-weighted Turbo Spin Echo pulse sequence. The MRI scanner was a Siemens 3.0 T Tim Treo system with an eight-channel head and neck coil. The size of each data-set was $256 \times 256 \times z$ ($z$ ranges from 10 to 24) with 0.94 mm $\times$ 0.94 mm in-plane resolution and 3 mm slice thickness. An Markov Random Field-based edge-preserving data combination technique was applied to build super-resolution volumes of the tongue with isotropic resolution of 0.94 mm (see Figure 2). Details of data acquisition and reconstruction technique can be found in Woo et al. (2012).

5. Results and discussion

The proposed method was evaluated on 18 normal subjects (8 females, 10 males). All 18 data-sets were manually segmented under the supervision of our dental expert collaborator using the interactive tool TurtleSeg (Top et al. 2011). The results were used both as the ground truth and as the source model in inter-subject registration, as described later in this section. The segmented surface for each volume included all the internal muscles of tongue as well as the digastric, geniohyoid and hyoglossus muscles (see Figure 5). It excluded the hyoid and mandible bone, epiglottis and salivary glands. We cut the mylohyoid, palatoglossus and styloglossus muscles following the contour of the tongue. In addition, the tongue tissue above the line between the epiglottis and hyoid bone was included. The process took about 5–7 h for each data-set.

We categorised the subjects into two groups of female and male anatomy. We noticed that anatomy of one male subject matched better with the female group; therefore, it was excluded from the male group and was added to the female group (F9 in Figures 6 and 7). For each data-set in each group, the segmentation was repeated by iterating the source on other members of the corresponding group, resulting in $(9 \times 8) \times 2$ experiments in total. In each case, the dental expert was asked to interact with the
segmentation for a minimum of 1 min and maximum of
3 min. The distance and the volume overlap between
the result ($A$) and the ground truth ($B$) were calculated before
and after the interaction. We used the dice coefficient as a
measure of the volume overlap, reported as a percentage:

$$\text{Dice}(A,B) = 2 \frac{|A \cap B|}{|A| + |B|} \times 100.$$  \hspace{1cm} (5)

For calculating the distance between the two surfaces,
we used the modified Hausdorff (MH) distance as the
measure of object matching (Dubuisson and Jain 1994):

$$\text{MH}(A,B) = \max\{d(A,B), d(B,A)\},$$

$$d(A,B) = \frac{1}{|A|} \sum_{a \in A} d(a,B), \quad d(a,B) = \min_{b \in B} d(a,b),$$  \hspace{1cm} (6)

where $N_a$ is the set of all the nodes, $a$, on the surface $A$;
$d(a,b) = |a - b|$ is the simple Euclidean distance between
vertices $a$ and $b$. Figure 6 shows the dice measure calculated
before and after expert interaction, for both groups of the
male and the female subjects. The average dice, measured on
all the data-sets in the male group, improved from
87.15 ± 1.65 to 90.37 ± 0.42 after expert interaction. The
mean of the inter-subject standard deviation (STD) also
dropped from 1.02 ± 0.28 to 0.29 ± 0.07. The average
overlap in the female group is 87.23 ± 1.58, before, and
90.44 ± 0.42, after interaction. The mean of the measured
STD also changes from 0.80 ± 0.16 to 0.29 ± 0.10. Figure 7
shows the average MH distance. The mean of the measure in
the male group changes from 2.06 ± 0.16 to
1.62 ± 0.10 mm after interaction. The mean of the measured
STD also changes from 0.15 ± 0.04 to 0.070.02 mm. For the
female group, the mean of MH distance is 2.06 ± 0.21 mm,

![Figure 5. Ground truth segmented by dental expert in TurtleSeg (Top et al. 2011). (a) Mid-axial, (b) mid-sagittal and (c) mid-coronal
slices shown with (d) 3D volume of the tongue. Salivary glands (labels 1–3), hyoid bone (label 4) and epiglottis (label 5) are excluded.](image)

![Figure 6. Average dice coefficient for subjects in the male (M) and female (F) groups, before (light grey) and after (dark grey) expert
interaction time of 2 ± 1 min.](image)
before, and $1.59 \pm 0.12$, after interaction. The mean of the STD also decreases from $0.13 \pm 0.04$ to $0.06 \pm 0.01$ mm. Figure 8 shows 3D representation of the result for subject number 5 as the source and subject number 2 as the target.

The manual segmentation used as the ground truth is likely to be fuzzy and uncertain in problematic regions, e.g. at the boundary of the hyoid bone and the salivary glands (see Figure 5). However, modelling requisites prohibit the segmentation to include bones and non-muscle soft tissues. To assess such uncertainty, we designed an experiment in which the same dental expert was asked to repeat the same manual segmentation after 10 days, while avoiding referring to the first set. The result showed a volume overlap of dice = 91% and a surface-to-surface distance of $MH = 1.52$ mm between the two manually segmented volumes. We argue that this uncertainty imposes an upper and lower limit of about the same values, 91% and 1.5 mm on the achievable volume overlap and surface-to-surface distance by the automated segmentation. In fact, while...
expert interaction resulted in about 7% improvement for dice of as low as 83% (0.6 mm decrease in distance value), the improvement was less than 1% for the dice values as high as 90% (0.3 mm decrease in distance value). Same ambiguities cause the user interaction not to be efficient after a certain time, justifying our choice of restricted interaction time for reporting the results.

### 5.1 Implementation details

All the parameters were fixed in all the experiments. We unified the number of surface nodes to 2502 in order to capture the sub-millimetre details of the tongue’s shape. For the intensity profile registration module, the length of the profiles was set to 50 pixels, centred on the investigated voxel. The search range was five voxels, inward and outward, in the direction normal to the mesh surface. The stiffness coefficient, $\alpha_e$, was set to 1. For shape-matching regularisation, we set the number of clusters to 300. To attain high flexibility, the radii of all clusters were set to 1.

*Sensitivity analysis*

In order to justify our choice of parameters, we performed a sensitivity analysis on four main parameters of the algorithm: number and radius of clusters as well as search range and stiffness coefficient. For each experiment, we changed the value of one of the parameters, while the others were fixed to their default values. To minimise the control variables, we excluded the initialisation and user interaction stages, running the analysis only for the mesh-to-volume registration. In each group (e.g. male/female), we chose the target image with the worst segmentation accuracy measured in the previous experiments (e.g. M7 and F6). Other members of the group were iteratively used as the source of registration. Hence, we conducted $(1 \times 8) \times 2$ experiments for each value of each parameter. The average and variance of the dice and MH measures, for the male and female subjects, are shown in Figures 9 and 10, respectively.

- **Number of clusters**: Proposed default value: 300. Small number of clusters result in a more rigid model and decrease the segmentation accuracy. More clusters will add more DOFs to the deformation but values greater than 600 cause failure in convergence.
- **Radius of clusters**: Proposed default value: 1. Radii of less than 10 produce similar results. The deformable model will become more rigid as the radius (e.g. overlap) of the clusters increases; hence, the segmentation accuracy decreases.
- **Search range**: Proposed default value: 5. Greater search range will enable the model to deform more in each time step. However, for values more than 10 pixels inward/outward, the model will have difficulties in convergence and will fluctuate back and forth close to the optimum solution.

![Sensitivity analysis graphs](image-url)  
*Sensitivity analysis on four main parameters of the proposed algorithm: dice and MH measures for subject number 7 in male group.*
- **Stiffness coefficient**: Proposed default value: 1. Low values of stiffness decrease the effect of image-based forces and hence are not adequate for matching the model to the target image. Values higher than 1.5 will introduce instabilities to the model.

6. Conclusions

In this article, we tackled the challenging problem of semi-automated 3D segmentation of the tongue from isotropic MRI volumes. Previous works included delineation of the tongue contours at its surface in 2D MRI slices. We adapted an inter-subject registration framework using a shape-matching-based regularisation technique. The method was combined with an instant force-based user interaction mechanism which attracts the model towards user-provided boundary labels. We were able to achieve segmentation accuracy with an average dice coefficient of 90.4 ± 0.4% and an average distance of 2 ± 0.2 mm, within an expert interaction time of 2 ± 1 min. Thus, we conclude that our human-in-the-loop approach using a variation of shape-matching technique proposed by Gilles and Pai (2008) provides an effective method to segment complicated soft tissue areas such as the tongue. Our future work will focus on integrating the proposed segmentation scheme within a more comprehensive biomechanical model suitable for modelling of speech and swallowing.

**Acknowledgements**

The authors would like to thank the collaborators from Dental School at University of Maryland – Professor Maureen Stone and Dr Jonghye Woo – for providing the MRI dataset. Also, our dental expert Dr Ho Beom Kwon is greatly acknowledged for supervising the validation process.

**Funding**

This work was supported in part by NSERC, CHRP and NCE GRAND.

**References**


