Tongue Mesh Extraction from 3D MRI Data of the Human Vocal Tract

Alexander Hewer, Stefanie Wuhrer, Ingmar Steiner, and Korin Richmond

Abstract In speech science, analyzing the shape of the tongue during human speech production is of great importance. In this field, magnetic resonance imaging (MRI) is currently regarded as the preferred modality for acquiring dense 3D information about the human vocal tract. However, the desired shape information is not directly available from the acquired MRI data. In this chapter, we present a minimally supervised framework for extracting the tongue shape from a 3D MRI scan. It combines an image segmentation approach with a template fitting technique and produces a polygon mesh representation of the identified tongue shape. In our evaluation, we focus on two aspects: First, we investigate whether the approach can be regarded as independent of changes in tongue shape caused by different speakers and phones. Moreover, we check whether an average user who is not necessarily an anatomical

Alexander Hewer
Saarbrücken Graduate School of Computer Science, Germany
DFKI Language Technology Lab, Saarbrücken, Germany
Cluster of Excellence Multimodal Computing and Interaction, Saarland University, Germany
e-mail: hewer@coli.uni-saarland.de

Stefanie Wuhrer
INRIA Grenoble Rhône-Alpes, France
e-mail: stefanie.wuhrer@inria.fr

Ingmar Steiner
DFKI Language Technology Lab, Saarbrücken, Germany
Cluster of Excellence Multimodal Computing and Interaction, Saarland University, Germany
e-mail: steiner@coli.uni-saarland.de

Korin Richmond
Centre for Speech Technology Research, University of Edinburgh, UK
e-mail: korin@cstr.ed.ac.uk

M. Breuß et al. (eds.), *Perspectives in Shape Analysis*, Mathematics
and Visualization, DOI 10.1007/978-3-319-24726-7_16

expert may obtain acceptable results. In both cases, our framework shows promising results.

1 Introduction

Shape analysis is of great importance in speech science. In this research area, analyzing and understanding the shape and the motions of the human tongue during the production of speech is of great interest. For example, a tongue model may be integrated into virtual avatars for multimodal spoken interaction or computer-aided pronunciation training. In the latter case, the user can be shown how to move the tongue to produce a specific sound [9]. Furthermore, such a tongue model could be used in articulatory speech synthesis to approximate the vocal tract area function.

Observing and imaging the tongue during speech is a challenging task, since it is inside the mouth and therefore almost completely hidden from view. Thus, traditional imaging modalities based on light, such as photography, are of limited use for acquiring information about the tongue. Currently, magnetic resonance imaging (MRI) can be regarded as the state-of-the-art technique for imaging the human vocal tract. This method is capable of providing 3D information about the inside of the mouth of a speaker without being hazardous or invasive.

The data acquired by MRI has to be further processed to extract the desired shape information, and manually extracting shape information from MRI scans can be a tedious and time-consuming task. This motivates an extended version of our framework [13] that combines image segmentation and template fitting to extract the tongue surface from a 3D MRI scan in a minimally supervised fashion. The only user input required by our method is a sparse set of annotated landmarks. Optionally, the user may additionally crop the MRI scan to the region containing the tongue for improved performance. We demonstrate experimentally that our method is stable with respect to inaccurate landmarks, which implies that a user who is not necessarily an anatomical expert is able to get acceptable results with only minimal input.

It is desirable to represent the extracted tongue surface using a high level representation. In this work, we choose as representation a polygon mesh. This representation has the advantage that it can be directly used in various fields of application, as meshes can be used to produce piecewise linear approximations of scenes of arbitrarily complex geometry and topology. The meshes can be textured and subsequently rendered in real-time to produce photo-realistic images. This even holds for large models, as polygon meshes can be easily represented in a hierarchy of resolutions using subdivision [5, Chapter 1]. Furthermore, polygon meshes are often employed in computer graphics to generate animations of complex objects [5, Chapter 9], and in computer vision to conduct a statistical analysis of a class of shapes, as for example faces [7]. By using polygon models, such deformations and statistical summaries can easily be computed for the extracted tongue surfaces. In speech processing, polygon models of tongues have been used to generate acoustical simulations [4], and using polygon models for our meshes allows us to use the extracted surfaces in existing simulation tools.

Our method uses a single generic template represented by a polygon mesh that was constructed based on an MRI scan by a non-expert. Experiments indicate that our approach has a success rate of 75 percent for the dataset of Adam Baker [3] and the Ultrax project [1]. Furthermore, we show that our method is independent of shape changes caused by different speakers and phones.

This chapter is organized as follows. Section 2 gives an overview of related work and Section 3 describes our framework and elaborates on the motivation behind the design. Section 4 provides background information on the datasets used as the source of the 3D MRI scans in our experiments. It is worth noting that compared to our previous work, we had data from more speakers available. In Section 5, we focus on investigating whether our approach is speaker- and phone-independent, and whether a non-expert user can achieve acceptable results. Finally, Section 6 gives conclusions and discusses open problems.

2 Related work

As it is tedious to manually extract information from MRI scans, a number of methods have been proposed to facilitate this process. Here, we provide a brief overview of recent methods.

The method of Peng et al. [22] aims at identifying the tongue's contour in a 2D mid-sagittal scan. It is based on an active contours approach [17] where a previously trained shape model is used to control the evolution of the contour. This technique was later extended by Eryildirim et al. [10] to align the contour's end points to the corresponding extremities of the tongue. More recently, Raessy et al. [23] showed that it is possible to train oriented active shape models [20] in such a way that they can be used to reliably identify the boundary of the tongue in 2D scans. These methods depend on manually preparing a training set and are restricted to the 2D case.

Lee et al. [16] proposed a framework for extracting the tongue from 3D dynamic MRI in a minimally supervised fashion. The random walker technique [11], which requires a user to manually place some seeds, was used as the base segmentation method. This framework produces a low-level volume segmentation.

Harandi et al. [12] used a template-matching technique to extract a mesh representation of the tongue from 3D MRI scans. A template is extracted from a source scan by an anatomical expert. This template is then fitted to a target scan using color information. Specifically, the mesh points are moved in such a way that the color at the original point in the source scan is similar to the deformed point in the target scan. This approach is limited by requiring an expert to provide the templates.

3 Framework

Our framework consists of three main steps. First, we apply an image segmentation technique to the MRI data to identify the spatial support of the tongue and related tissue.

Second, we extract the surface points of the tissue, thereby reducing the data to a purely geometric representation. This is motivated by the fact that it is relatively easy to combine geometric information from different sources. For example, the surface point cloud obtained from one scan might be incomplete. In this case, the information obtained from a second scan of the same speaker could be used to reconstruct certain missing data by simply adding the corresponding points to the point cloud of the first scan.

Third, we apply a template fitting technique to obtain a polygon mesh representation of the tongue surface from the point cloud. Using such a method has the advantage that we can exploit prior knowledge about the shape of the tongue in the form of a provided template. This is especially useful in situations where the point cloud is noisy, incomplete, or contains additional information other than the tongue.

3.1 Interpretation of a scan as a 3D image

Before discussing our proposed method, we describe how an MRI scan can be turned into a 3D image.

Formally, a scan is given by $g : S \to \mathbb{R}$ where $S \subset \mathbb{R}^3$ is a discrete domain in the form of a rectangular box. The scan domain S contains the positions $\mathbf{x}$ at which the scanner took the measurements. Thus, $g(\mathbf{x})$ represents the density of hydrogen molecules measured by the scanner at coordinate $\mathbf{x}$. Each sample position represents a point on a regular grid with grid spacings h_x, h_y, and h_z.

A 3D image, on the other hand, is given by $f : \Omega \to [0,255]$ where $\Omega \subset \mathbb{R}^3$ is again a discrete domain in the form of a rectangular box. Here, $f(\mathbf{y})$ is the gray-value at voxel coordinate $\mathbf{y}$. In contrast to the sample positions, however, these voxel coordinates are arranged on a Cartesian grid with $h_x = h_y = h_z = 1$.

This means we first have to find a mapping $s : \Omega \to S$ from the voxel coordinates in our image representation to the sample positions of the scan. Here, we can use $\mathbf{y} = (x,y,z)^\top \in \Omega$ as an index to access the vertices of the regular grid in S, as

$$s(\mathbf{y}) := \Big(xh_x,\, yh_y,\, zh_z\Big)^\top . \tag{1}$$

To visualize the measured hydrogen density, we define a quantization operator $q : \mathbb{R} \to [0,255]$ that maps the observed densities to 256 values. This allows us to interpret the scan as 3D image $f : \Omega \to [0,255]$ where

$$f(\mathbf{y}) = q\big(g(s(\mathbf{y}))\big) \tag{2}$$

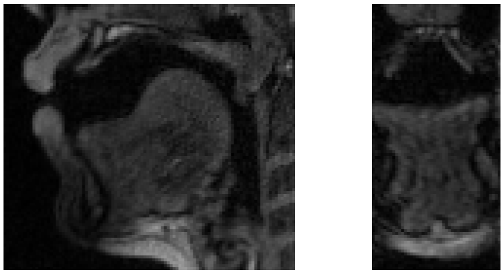


Fig. 1 Two different slice types of a 3D image showing the human vocal tract. **Left:** Sagittal slice. **Right:** Coronal slice.

can be seen as the quantized gray-value representation of the hydrogen density at sample position $s(\mathbf{y})$.

In the following, we assume that the data was recorded in a standard sagittal manner, and refer to an (x,y)-plane of an MRI scan as a sagittal slice and to a (y,z)-plane of an MRI scan as a coronal slice. Both types of slices are shown in Figure 1.

3.2 Image segmentation

The first step of our method aims to identify the spatial support of the tongue. That is, we wish to divide Ω into an object region Ω_O and a background region Ω_B. The object region Ω_O should contain points that are related to the tongue. However, it is also allowed to contain regions that belong to other organic tissue. This relaxation is necessary as in some images no boundary may be detectable between the tongue and other tissues with which it is in contact, such as the palate. The background region Ω_B consists of parts of the scan we have no interest in. These are, for example, bones, air, or other tissue not related to the tongue.

Figure 1 demonstrates that an object can be distinguished from the background by using color information. This motivates the use of image segmentation techniques that make use of color information to extract Ω_O.

As we aim to apply our method to large datasets, the segmentation method must satisfy two requirements. First, the required manual input from the user should be minimal. Second, the segmentation method should be robust. To satisfy both requirements, we compute segmentations using the method by Chan and Vese [8]. This method is robust and generates smooth boundaries between Ω_O and Ω_B, which can later be used to derive clean surface normals.

The method by Chan and Vese requires as initialization a closed contour C that separates Ω into Ω_O and Ω_B. In our approach, this initial contour can be computed automatically: Given a sparse set of manually annotated landmarks L as described in Section 5.1, a sphere can be placed at the centroid of these landmarks in Ω. Alternatively, it can be positioned at the center of Ω if the image mainly shows the tongue, as for example in Figure 2.

The approach evolves the initial contour C such that the gray-value variance inside the regions is minimized, i.e.

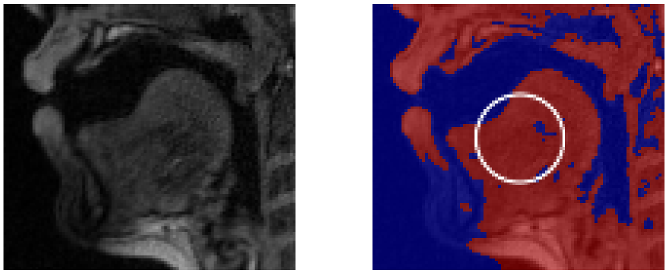


Fig. 2 Example result of Chan-Vese in 2D. **Left:** Sagittal slice. **Right:** Resulting segmentation. Ω_O is colored in red, Ω_B in blue. The initial contour used is shown as a white circle.

$$E_{\mathrm{CV}}(C) = \sum_{\mathbf{x}\in\Omega_O} \left(f(\mathbf{x}) - \mu_{\Omega_O}\right)^2 + \sum_{\mathbf{x}\in\Omega_B} \left(f(\mathbf{x}) - \mu_{\Omega_B}\right)^2 + \lambda\,\mathrm{length}(C), \tag{3}$$

where Ω_O and Ω_B are the regions induced by C and μ_X represents the average gray-value in region X. The method has a regularizer weighted by $\lambda > 0$ that tries to minimize the length of the contour. To minimize the energy, we apply the standard scheme of Chan and Vese. That is, we start with a continuous version of the energy that uses a level set representation [21] of the contour, and subsequently derive the Euler-Lagrange equation of this energy to set up a gradient descent approach that is discretized using a finite differences implicit scheme. Figure 2 shows an example result in 2D that used a circle as the initial contour.

Note that the remainder of our method is independent of the selected segmentation method, and any segmentation method can be freely selected if this is advantageous for a specific dataset. In our preliminary experiments [13], we also explored a graph cut method [6] for segmentation. However, we did not explore this option further as approaches of the graph cut family require a significant amount of manual input, rendering them impractical when processing large datasets.

3.3 Surface point extraction

Given a partition $\Omega = \Omega_O \cup \Omega_B$, we compute the surface information by extracting surface points $P^* := \{\mathbf{p}_i\}$ of Ω_O and normals $N := \{\mathbf{n}_i\}$ for P^*, such that $\mathbf{n}_i$ is the normal at $\mathbf{p}_i$. Surface points $\mathbf{p}_i$ are points of Ω_O that have at least one neighboring point $\mathbf{q}$ in Ω_B. Surface normals are chosen to point towards the outside of Ω_O. Note that due to the relaxation we formulated earlier for Ω_O, P^* may contain surface points belonging to other articulators than the tongue. Furthermore, P^* is a subset of Ω, i.e., the surface information was computed in the image domain. The template fitting, however, should operate on the domain of the observed vocal tract to be anatomically correct. Thus, we apply the mapping from Equation (1) to obtain the correct surface information P as

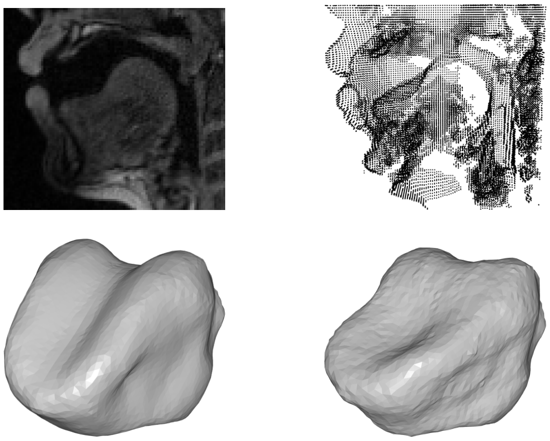


Fig. 3 Example result of the template fitting method. **Top row:** Sagittal slice of the used MRI scan (**left**) and obtained point cloud (**right**). **Bottom row:** Template used in our approach (**left**) and result of the template fitting (**right**).

$$P := \{s(\mathbf{p}) \mid \mathbf{p} \in P^*\}. \tag{4}$$

The surface P consists of a loose collection of points, as shown in Figure 3. Furthermore, the point cloud may be missing information and may contain data other than the tongue. Therefore, surface reconstruction approaches like the Poisson reconstruction [14] may produce undesirable results. To avoid this problem, in the following, we utilize the information that a subset of P forms part of the surface of a tongue.

3.4 Template fitting

We use a template fitting technique [25] to jointly find the subset of P representing the tongue and a polygon mesh representation of the tongue surface. That is, we deform a template mesh $M := (V, F)$ to match the point cloud data P. We use a vertex-face representation of meshes, i.e., $V := \{\mathbf{v}_i\}$ denotes the vertex set of the mesh with $\mathbf{v}_i \in \mathbb{R}^3$ and F its face set. To obtain a deformation, the approach computes a set $A := \{A_i\}$ where $A_i : \mathbb{R}^3 \to \mathbb{R}^3$ is a rigid body motion for the vertex $\mathbf{v}_i$ by minimizing the energy

$$\begin{aligned} E_{\text{Def}}(A) = {} & \alpha \frac{1}{|V^*|} \sum_{v_i \in V^*} \left(\operatorname{dist}_D \left(A_i(\mathbf{v}_i), \arg\min_{\mathbf{p}_j \in P} \| A_i(\mathbf{v}_i) - \mathbf{p}_j \| \right) \right) \\ & + \beta \frac{1}{|V|} \sum_{v_i \in V} \left(\sum_{v_j \in \mathcal{N}(v_i)} \operatorname{dist}_S \left(A_i, A_j \right) \right) \\ & + \gamma \frac{1}{|L|} \sum_{(\mathbf{v}_i, \mathbf{q}_i) \in L} \left(\operatorname{dist}_L \left(A_i(\mathbf{v}_i), \mathbf{q}_i \right) \right). \end{aligned} \tag{5}$$

This energy consists of three terms. Each term is weighted by a non-negative value, α, β, or γ, that is normalized according to the number of participating vertices in the respective term. This normalization makes it easier to compare the influences of the different terms.

The data term $\operatorname{dist}_D(\cdot)$ measures the distance between the transformed vertex $A_i(\mathbf{v}_i)$ and the normal plane at its nearest neighbor. This term is minimized when the template is close to the point cloud P. In our implementation, this term is only evaluated at $V^* \subseteq V$ to increase robustness to noise. In particular, a vertex $\mathbf{v}_i$ is ignored if the Euclidean distance between $\mathbf{v}_i$ and its nearest neighbor is too large or if the angle between the outer normals of $\mathbf{v}_i$ and its nearest neighbor is too large. This commonly used heuristic [2, 18] is meant to distinguish valid data observations from invalid ones. Additionally, we do not consider vertices that are part of the landmark set L to avoid distorting any manually provided correspondences.

The deformation smoothness term $\operatorname{dist}_S(\cdot)$ measures the difference in rigid body motion A_i between $\mathbf{v}_i$ and the vertices of the neighborhood $\mathcal{N}(\mathbf{v}_i)$ that consists of the one-ring neighbors of $\mathbf{v}_i$ and vertices of the mesh within distance of $2 \cdot \operatorname{res}(M)$ from $\mathbf{v}_i$ where $\operatorname{res}(M)$ is the average edge length of the template mesh M. The minimization of $\operatorname{dist}_S(\cdot)$ encourages the template to preserve its overall shape during deformation, which helps to keep the mesh away from data points that do not belong to the surface of the tongue and allows missing parts to be filled in smoothly. This term is active at all vertices.

Finally, the landmark term $\operatorname{dist}_L(\cdot)$ is optional. This term computes the squared Euclidean distance between pairs of manually annotated vertices $\mathbf{v}_i \in V$ and corresponding coordinates $\mathbf{q}_i \in \mathbb{R}^3$ that are contained in a set of landmarks $L := \{(\mathbf{v}_i, \mathbf{q}_i)\}$. Note that the coordinates $\mathbf{q}_i$ do not have to be contained in P. By minimizing this term, the approach will move the selected vertices to the user-provided coordinates.

We discover that minimizing both the data and the smoothness terms will move the mesh to a subset of P that resembles a tongue-like surface.

We follow a similar strategy as Wuhrer et al. [25] to obtain a minimizer A of the energy. Before performing the optimization, we perform a rigid alignment of the template. This step uses the user-provided landmarks and the point cloud to find a good scale and position for the template.

The energy given in Equation (5) is not differentiable with respect to A, which prevents us from minimizing it directly. Therefore, we perform the optimization by minimizing a series of differentiable energies $E^t_{\text{Def}}(A^t)$ where $t \in [1, t_{\max}]$. The energy E^t_{Def} differs from the original energy E_{Def} in the following way: In E^t_{Def}, we use the minimizer of the previous energy in the series to transform the vertex in $\mathrm{dist}_D(\cdot)$: $A_i^{t-1}(\mathbf{v}_i)$. This means that arg $\min_{\mathbf{p}_j \in P}(\cdot)$ no longer depends on A^t. Thus, the energy becomes differentiable and we can use a quasi-Newton technique [19] to compute the minimizer. Moreover, for $t_{\max} > 1$, the weight β of the smoothness term changes in each iteration. Given a base value β, the weights β^t used in iteration t are computed as

$$\beta^t = 2\beta - (t-1)\frac{\beta}{t_{\max}-1}. \qquad (6)$$

This means that we start the optimization by promoting smooth transformations. The weight is then gradually reduced until we arrive at the base weight β in the last iteration.

After the minimization of the last energy, we obtain the sought transformations A as $A^{t_{\max}}$. Note that we use the identity $A_i^0(\mathbf{v}_i) = \mathbf{v}_i$ as A^0 that is needed in the first energy E_{Def}^1. Furthermore, we apply a coarse-to-fine strategy to cope with large deformations.

Figure 3 illustrates an example of the template fitting.

4 Datasets

This study is evaluated on a large dataset of 12 speakers, and extends our previous work [13], which only considered data from a single speaker. We use two MRI datasets to validate our method, that of Adam Baker [3], and the full dataset from the Ultrax project [1].

The Baker dataset contains static 3D MRI scans of a male speaker. 25 of these scans are speech related and show vocal tract configurations for different phones. This data was acquired as part of the Ultrax project, but released separately.

The remainder of the Ultrax dataset consists of static 3D MRI scans of 11 adult speakers. Seven of these speakers are female and four are male. While scanning, the subjects, who were all phonetically trained, were asked to sustain the articulatory configurations for a given phone for around 20 seconds. Prompts were displayed to the subject using a laptop connected to video-goggles. Each subject recorded scans for the following phone set [i, e, ɛ, a, ɑ, ʌ, ɔ, o, u, ʉ, ə, s, ʃ], with an additional scan for the pose at rest. Simultaneous audio recordings were made using a FOMRI-III fiber optic microphone. This microphone is specially designed for use in MRI scanners, using both a pair of microphones and adaptive noise cancellation algorithms to reduce the level of MRI scanner noise. Though it is not possible to remove the scanner noise entirely, the use of this microphone does make it possible to monitor and verify the subject's phone production acoustically. The Ultrax dataset also contains other types of MRI scans for all subjects, but those were not used in this work.

The scans were acquired using a Siemens Verio 3T scanner at the Clinical Research Imaging Centre in Edinburgh. Each scan comprises 44 sagittal slices with a thickness of 1.2 mm and an image size (whole head) of 320 × 320 pixels in the image domain. In the scan domain, we have distances of $h_x = h_y = 1.1875\,\mathrm{mm}$

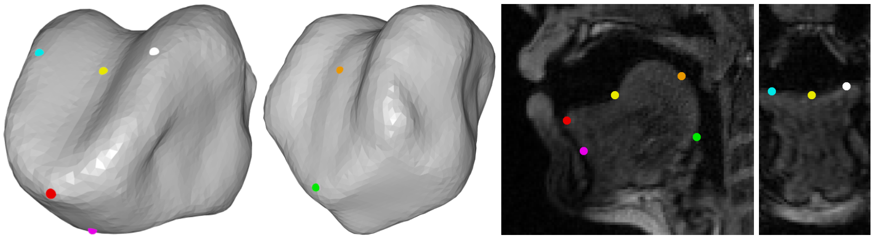


Fig. 4 Placement of the landmarks. **Left:** Selected vertices for the landmarks on the template. The left image shows a front view of the template, the right one a view from the back. **Right:** Sagittal and coronal slice showing an example of the corresponding user-provided landmarks on an MRI scan.

and $h_z = 1.2\,\text{mm}$, corresponding to a voxel size of $1.1875 \times 1.1875 \times 1.2\,\text{mm}^3$. The scans were acquired with an echo time of $0.93\,\text{ms}$ and a repetition time of $2.36\,\text{ms}$.

5 Evaluation

The focus of this section is on investigating whether our approach can be regarded as independent of shape changes caused by different speakers and phones. To show this independence, we demonstrate that is possible to obtain satisfying results across different speakers and phones by always applying the same procedure. To this end, all parameters except for the landmarks are fixed for all scans.

In the following, we first outline how the template is created and how the scans are prepared. We then describe experiments to evaluate the stability of the weights in the template fitting, investigate whether our approach is applicable to different speakers and phones, and analyze the robustness of our approach to erroneously placed landmarks.

5.1 Template creation

The template is manually extracted from a scan of the Baker dataset. After the extraction, we adjust the mesh to be symmetric to remove this particular bias towards the original speaker. Note that the template only models the upper part of the tongue surface and does not include its sublingual part. The template consists of 5864 vertices and 11724 faces, and is shown in Figure 4.

We select seven vertices as landmarks. These vertices and an example of the corresponding user-provided coordinates on an MRI scan are shown in Figure 4. Five landmarks are distributed on a sagittal slice that is located roughly at the center of the tongue. Three of these landmarks are located at feature points that are relatively easy to locate for an average user, namely the tongue root near the epiglottis and the pharynx (green landmark), the tongue tip (red landmark), and the position where the tongue surface connects to its sublingual part (pink landmark). The remaining two landmarks in the mid-sagittal slice are placed at approximately $\frac{1}{3}$ and $\frac{2}{3}$ of the distance from the tongue tip to the root, corresponding to the tongue blade (yellow landmark) and back (orange landmark), respectively. We believe that using this feature-free approach to select the tongue blade and back facilitates the landmark placement. The tongue blade landmark serves as anchor for two additional lateral landmarks that may be positioned using a coronal slice. These are located near the left (blue landmark) and right (white landmark) boundaries of the tongue's upper surface and serve to add lateral information to the landmark set.

Note that not all landmarks are required for our approach. If the user does not provide coordinates for a subset of the landmarks, these landmarks will simply be ignored in the optimization process.

5.2 Scan selection and preparation

We consider the data of all available speakers to ensure high variance with respect to speaker-specific anatomy. To obtain a high variance of intra-speaker tongue shape, scans corresponding to the three corner vowels [ɑ, i, u] are considered for each speaker. These vowels show the tongue in different extreme positions, e.g. as far back and low in the mouth as possible for [ɑ] [15]. We discovered that one speaker showed a high activity of the soft palate leading to contacts with the tongue. Therefore, we removed scans of this specific speaker from further processing. Furthermore, we removed one scan from another speaker because a part of the tongue was not visible.

After this selection process, the data is pre-processed using three steps. First, each scan is cropped to a region of interest containing the vocal tract.

Second, each scan is segmented automatically using the Chan-Vese method. Here, we use $\lambda = 140$ and initialize C to a sphere of radius 15 located at the center of the image representation of the cropped scan. We found that this approach failed to properly segment the scans of one speaker, and all scans of this speaker were removed from further processing. After these steps, 29 point clouds derived from the scans were available for further experiments.

Third, we manually select the landmark coordinates in each scan. To facilitate this task, we developed a graphical user interface that allows landmarks to be placed on the image representation of the scan. Subsequently, the landmark positions are mapped to the scan domain. In our experiments, we encountered scans where the placement of the two lateral landmarks posed a problem. Due to contact with other tissue, the left and right boundaries of the tongue's upper surface were difficult to identify. We found that these landmarks are not always needed to obtain acceptable results. For our experiments, we use the 2 lateral landmarks for only 13 of the 29 scans.

Note that this workflow may be modified. In particular, it is possible to omit the cropping step, thereby reducing the amount of manual pre-processing required of the user. Working with the full scans produces the same results as working with the cropped scans if the Chan-Vese method is initialized after pre-aligning the scans based on the provided landmarks. However, working with the cropped scans decreases the processing time of the segmentation method and the memory requirements for computing the point cloud.

5.3 Experiments

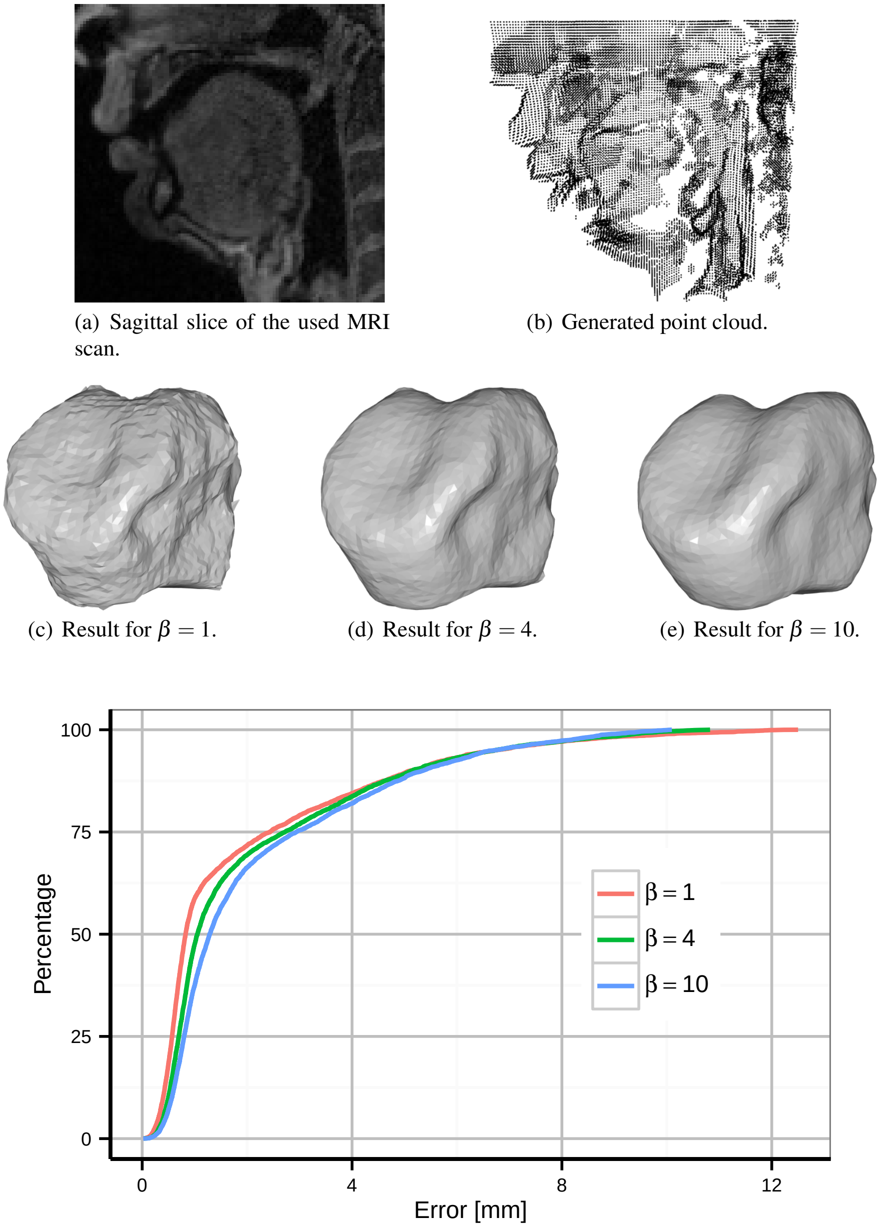
As no ground truth is available, we evaluate the results by computing the Euclidean distances between vertices on the deformed template and their nearest neighbors in the point cloud. Since our template is incomplete, we ignore vertices at the bottom of the mesh, as they are not part of the tongue's boundary. To quantitatively summarize the results, we compute cumulative error functions. For a qualitative evaluation, we show the visual quality of some of the results.

In all following experiments, the parameters $\alpha = 1$ and $t_{\text{max}} = 20$ are fixed. In the data term, we use the same heuristic as [25] to identify valid data observations: We consider only vertices of the template mesh M whose nearest neighbor in the point cloud is at distance at most $5 \cdot \text{res}(M)$ and whose normal deviates at most 60 degrees.

Influence of parameters

We first evaluate the stability of the weights β and γ used in the optimization. This evaluation consists of two parts. First, we check if there is a weight β that produces acceptable results for all scans by setting $\gamma = 0$ and testing the ten weights $\beta = 1, 2, \dots, 10$. In this experiment, the landmarks are used only for rigid alignment.

The parameter value $\beta_{\text{optimal}} = 4$ represents a good compromise between closeness to the data and smoothness of the resulting mesh. This can be seen in Figure 5, which shows the results for an example scan from the Baker dataset. On the one hand, low weights for β lead to overfitting, which produces a very noisy mesh. On the other hand, high weights for β reduce the amount of alignment because the smoothness term has too much influence. Note that the very large distances visible in the cumulative error function are due to holes in the corresponding point cloud and can therefore be disregarded. We encountered 13 scans where this choice for β produced suboptimal results. The poor performance in 4 of those scans was related to palate contacts of the tongue or segmentation issues. In the remaining 9 scans, the poor performance stems from template fitting related problems. For example, the tongue tip of the template was aligned to the front palate region in some results.



(a) Sagittal slice of the used MRI scan.

(b) Generated point cloud.

(c) Result for $\beta = 1$.

(d) Result for $\beta = 4$.

(e) Result for $\beta = 10$.

(f) Cumulative error functions for the different results.

Fig. 5 Example showing how the weight β of the smoothness term affects the result.

Additionally, in some scans, parts of the template were not aligned to the data, as shown in Figure 6(c).

Second, we analyze whether activating the landmark energy in equation (5) can improve the results for fixed β_{optimal}. Specifically, we consider weights $\gamma = 0.1, 0.2, \ldots, 1$. Hence, for this experiment, the landmarks were used in the template fitting.

Figure 6 shows that even using small values of γ can improve the results significantly. The figure shows a particular scan from the Ultrax dataset where activating the landmark energy drastically improves the mesh alignment. On our dataset, the value $\gamma_{\text{optimal}} = 0.1$ led to the best results. For this parameter setting, 6 of the 9 scans that had template fitting problems for $\gamma = 0$ are aligned correctly.

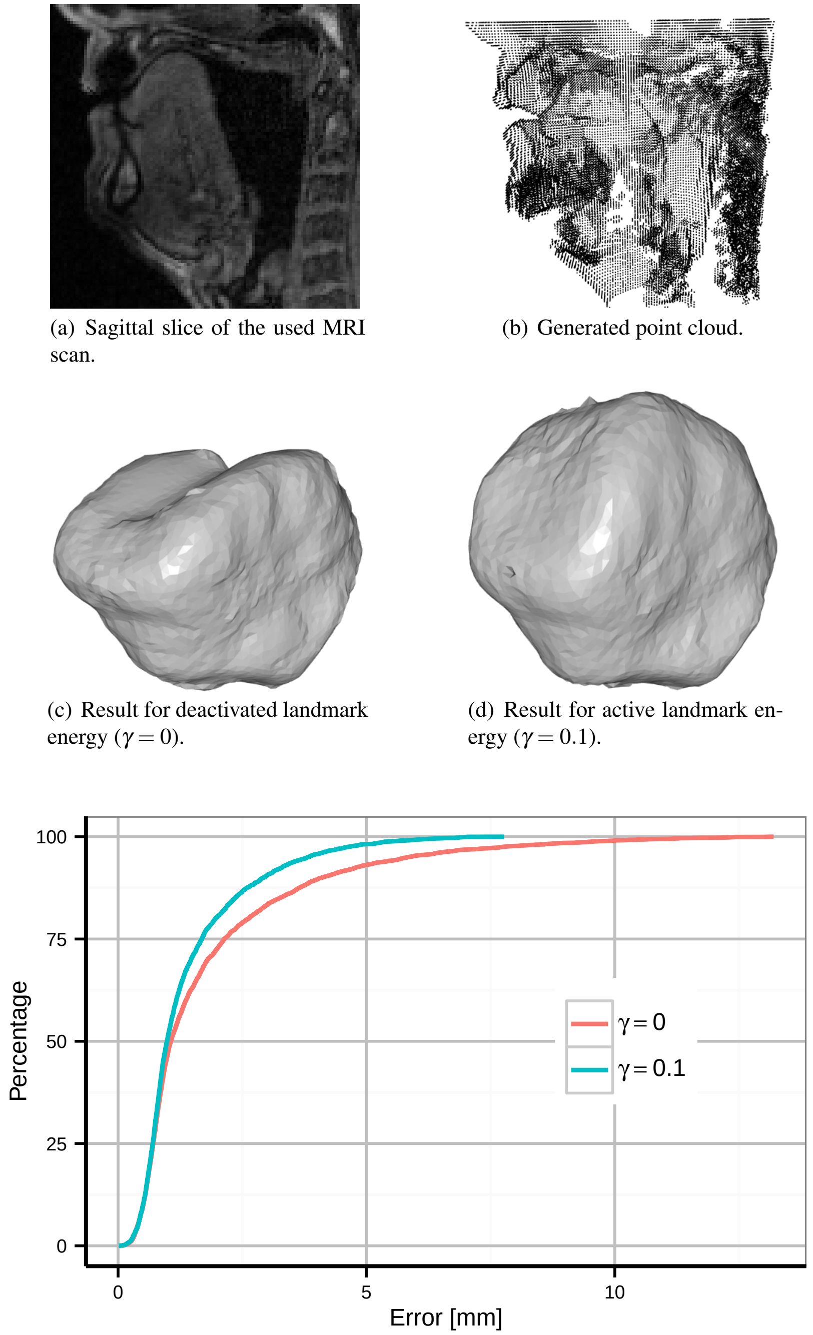
Evaluation of independence of speakers and phones

We now evaluate the template fitting results obtained for parameters $\beta_{\text{optimal}} = 4$ and $\gamma_{\text{optimal}} = 0.1$ across different speakers and phones. For these parameter settings, our approach was successful for 22 of the 29 considered scans. These 22 scans include scans from all 10 speakers for which scan preparation was successful and scans from all three considered phones. To evaluate whether the method is biased towards specific speakers or phones, we consider the set of cumulative error plots across different phones and speakers. To avoid large distances originating from potential holes in the point cloud, we only consider distances below 5 mm in the error computation. Figure 7(a) shows the distribution of cumulative error plots for different phones, and Figure 7(b) shows the distribution of cumulative error plots for different speakers. Note that all cumulative error plots are similar, and hence the variance between the plots is low. This shows that for our dataset, there is no significant bias towards any specific speaker or phone, and leads us to conclude that our approach is speaker- and phone-independent.

Evaluation of noisy landmark placement

In the final experiment, we analyze the robustness of our approach against errors in the coordinates of the landmarks provided by the user. To this end, we add Gaussian noise with mean 0 mm and standard deviation 5 mm to each component of the original coordinates to simulate the input of an inexperienced user. We only consider the scans where our framework succeeded and used the optimal weights $\beta = \beta_{\text{optimal}}$ and $\gamma = \gamma_{\text{optimal}}$.

Errors in the landmarks do not have a significant effect on the results. In all but one of the tested scans, our approach obtains acceptable results even when noisy landmarks are used. Figure 8 shows a representative example of a deformed template computed using noisy landmarks. Note that the shape of the deformed templates obtained with clean and noisy landmarks is globally quite similar and only leads to



(a) Sagittal slice of the used MRI scan.

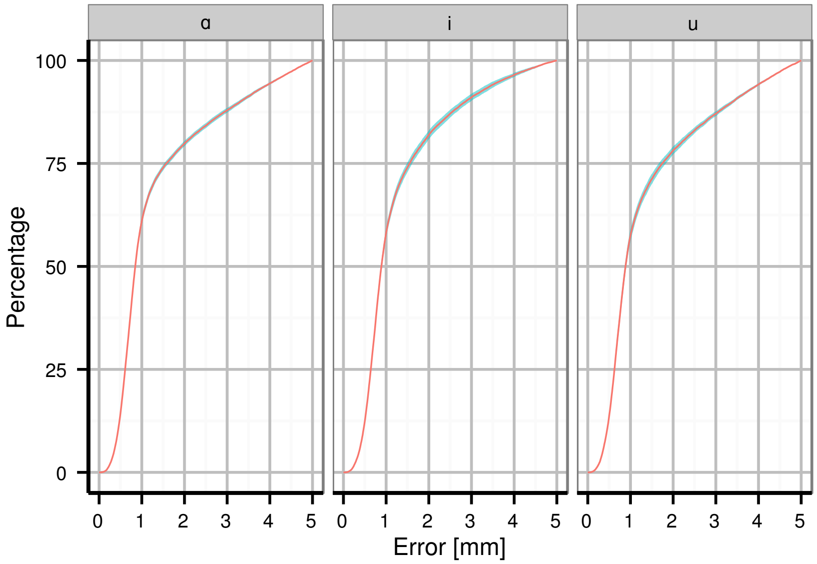
(b) Generated point cloud.

(c) Result for deactivated landmark energy ($\gamma = 0$).

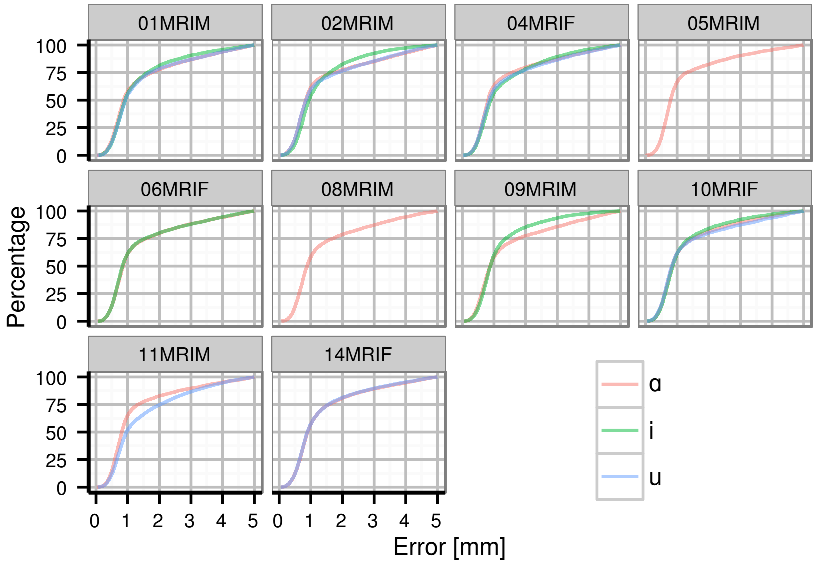
(d) Result for active landmark energy ($\gamma = 0.1$).

(e) Cumulative error functions for the two results.

Fig. 6 Example showing how the landmark energy can help to improve the result.

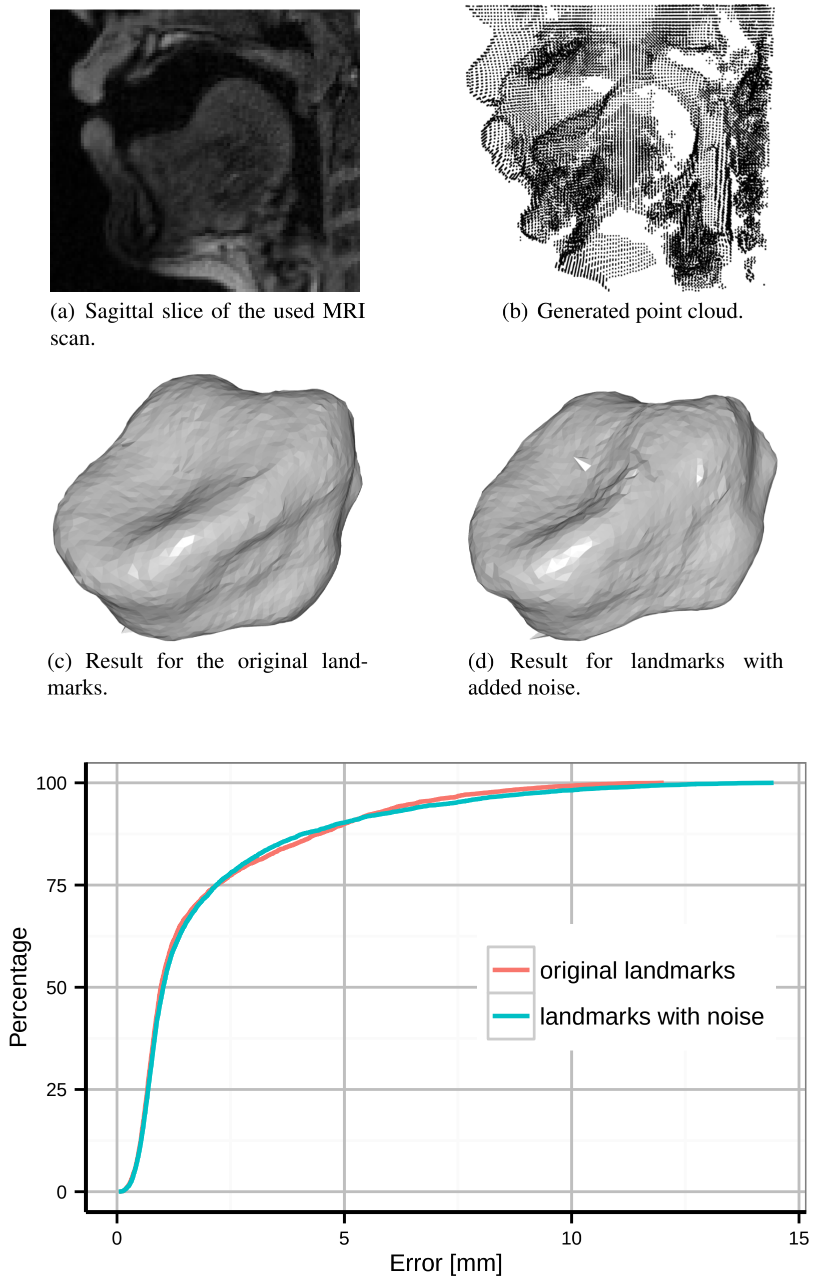


(a) Cumulative error of the results grouped by phone. The plot shows the mean error (line) and the standard error (ribbon) of all results belonging to the corresponding phone.



(b) Cumulative error grouped by speaker. Missing lines indicate that no result was obtained for the specific phone.

Fig. 7 Visualizations of the cumulative error for the 22 scans where our approach succeeded.



(a) Sagittal slice of the used MRI scan.

(b) Generated point cloud.

(c) Result for the original landmarks.

(d) Result for landmarks with added noise.

(e) Cumulative error functions for the two results.

Fig. 8 Example showing the effect of noise in the landmarks.

localized differences. However, we encountered one scan where the noisy landmarks lead to a suboptimal result.

5.4 Discussion

Our approach succeeded in 75 percent of the selected scans for a fixed template and fixed parameter settings. Furthermore, the proposed framework did not show any significant bias towards a specific phone or speaker, which indicates that it is phone- and speaker-independent. Here, we want to note that in the study of Harandi et al. [12] only the speaker-independence of their approach was analyzed. In particular, they only considered the tongue in the resting position and evaluated their method across 18 speakers. Moreover, our approach is robust against errors in the landmarks provided by the user. Thus, even an inexperienced user may obtain acceptable results using our method.

The observed failure cases stem from three main causes. Issues with the segmentation approach forced us to discard data from certain speakers completely, or prevented our framework from producing acceptable results. Using more than one segmentation technique may help to overcome these problems. Multiple segmentation results could be generated, and the user could then select the best one to use in the subsequent processing steps.

Furthermore, for scans where a contact between tongue and palate occurred, finding surface information of the tongue in the contact area is difficult because it may not be visible, which leads to a hole in the point cloud. Note that if we reconstruct the hard palate surface in this region, it may be used to represent the portion of the tongue surface in contact with the palate. For a point cloud P where such a hole is present due to a contact in the region of the hard palate, we explored the following approach to reconstruct the palate. First, a scan is selected where the hard palate is clearly visible, and the subset of points H representing the palate surface is extracted. Second, the hard palate is manually aligned to match the vocal tract configuration in P, which results in the set of transformed points H^*. Note that this alignment is easy to perform manually because the hard palate can only undergo rigid body motions. Third, P and H^* are merged into a single point cloud, which is used in the template fitting. This palate reconstruction can improve the results in cases where palate contact results in incomplete point clouds.

Finally, for the scans where the template fitting failed, we suspect that using more landmarks could help to align the template correctly to the point cloud.

6 Conclusion

In this chapter, we presented a minimally supervised approach to extract mesh representations of the human tongue from MRI data of the vocal tract. The experiments performed revealed promising results, as the presented approach leads to results of high quality in 75 percent of our tests. An important feature of the approach is its independence with respect to changes in tongue shape due to different speakers and phones. Furthermore, the approach is robust to noise in the manually placed landmarks.

We leave the following open problems for future work. A palate reconstruction could help to significantly increase the number of scans that can be processed successfully by our approach. Hence, it is important to facilitate the process of palate reconstruction. We plan to replace the process of manually aligning the palate surface to the MRI data with a rigid alignment approach based on landmarks that are not necessarily located on the tongue.

Our template fitting could be improved by including more information, such as the sublingual part of the tongue, more annotated landmarks, or typical MR-values at the vertices. Such modifications may improve the performance of the template fitting.

Moreover, the evaluation of our approach could be made more thorough by using more datasets and comparing the results to other methods in literature. However, datasets in literature are in general not easy to access due to privacy concerns for the recorded subjects.

For the future, we also think that it would be worthwhile to explore the performance of robust unsupervised methods, like for example [24], in the segmentation part of the framework. Detecting the position of the landmarks automatically would be another interesting modification. Both improvements could make the framework more accurate and further reduce the input required by the user or even make it fully automatic.

Acknowledgements This study uses data from work supported by EPSRC Healthcare Partnerships Grant number EP/I027696/1 (“Ultrax”).

References

[1] Ultrax: Real-time tongue tracking for speech therapy using ultrasound (2014). URL http://www.ultrax-speech.org/. Accessed 5 May 2015

[2] Allen, B., Curless, B., Popović, Z.: The space of human body shapes: reconstruction and parameterization from range scans. ACM Transactions on Graphics **22**(3), 587–594 (2003). DOI 10.1145/1201775.882311

[3] Baker, A.: A biomechanical tongue model for speech production based on MRI live speaker data (2011). URL http://www.adambaker.org/qmu.php. Accessed 5 May 2015

[4] Blandin, R., Arnela, M., Laboissière, R., Pelorson, X., Guasch, O., Hirtum, A.V., Laval, X.: Effects of higher order propagation modes in vocal tract like geometries. Journal of the Acoustical Society of America **137**(2), 832–843 (2015). DOI 10.1121/1.4906166

[5] Botsch, M., Kobbelt, L., Pauly, M., Alliez, P., Levy, B.: Polygon Mesh Processing. A K Peters/CRC Press (2010)

[6] Boykov, Y., Funka-Lea, G.: Graph cuts and efficient ND image segmentation. International Journal of Computer Vision **70**(2), 109–131 (2006). DOI 10.1007/s11263-006-7934-5

[7] Brunton, A., Salazar, A., Bolkart, T., Wuhrer, S.: Review of statistical shape spaces for 3D data with comparative analysis for human faces. Computer Vision and Image Understanding **128**, 1 – 17 (2014). DOI 10.1016/j.cviu.2014.05.005

[8] Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Transactions on Image Processing **10**(2), 266–277 (2001). DOI 10.1109/83.902291

[9] Engwall, O.: Can audio-visual instructions help learners improve their articulation? - an ultrasound study of short term changes. In: 9th Annual Conference of the International Speech Communication Association (Interspeech), pp. 2631–2634 (2008)

[10] Eryildirim, A., Berger, M.O.: A guided approach for automatic segmentation and modeling of the vocal tract in MRI images. In: European Signal Processing Conference (EUSIPCO), pp. 61–65 (2011)

[11] Grady, L.: Random walks for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(11), 1768–1783 (2006). DOI 10.1109/TPAMI.2006.233

[12] Harandi, N.M., Abugharbieh, R., Fels, S.: 3D segmentation of the tongue in MRI: a minimally interactive model-based approach. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **3**(4), 178–188 (2015). DOI 10.1080/21681163.2013.864958

[13] Hewer, A., Steiner, I., Wuhrer, S.: A hybrid approach to 3D tongue modeling from vocal tract MRI using unsupervised image segmentation and mesh deformation. In: 15th Annual Conference of the International Speech Communication Association (Interspeech), pp. 418–421 (2014)

[14] Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Eurographics Symposium on Geometry Processing (SGP), pp. 61–70 (2006). DOI 10.2312/SGP/SGP06/061-070

[15] Ladefoged, P.: A Course in Phonetics, 2nd edn. Harcourt Brace Jovanovich (1982)

[16] Lee, J., Woo, J., Xing, F., Murano, E.Z., Stone, M., Prince, J.L.: Semi-automatic segmentation of the tongue for 3D motion analysis with dynamic MRI. In: IEEE 10th International Symposium on Biomedical Imaging (ISBI), pp. 1465–1468 (2013). DOI 10.1109/ISBI.2013.6556811

[17] Li, C., Kao, C.Y., Gore, J.C., Ding, Z.: Implicit active contours driven by local binary fitting energy. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–7 (2007). DOI 10.1109/CVPR.2007.383014

[18] Li, H., Adams, B., Guibas, L.J., Pauly, M.: Robust single-view geometry and motion reconstruction. ACM Transactions on Graphics **28**(5), 175:1–175:10 (2009). DOI 10.1145/1618452.1618521

[19] Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Mathematical Programming **45**(1-3), 503–528 (1989). DOI 10.1007/BF01589116

[20] Liu, J., Udupa, J.K.: Oriented active shape models. IEEE Transactions on Medical Imaging **28**(4), 571–584 (2009). DOI 10.1109/TMI.2008.2007820

[21] Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. Journal of Computational Physics **79**(1), 12–49 (1988). DOI 10.1016/0021-9991(88)90002-2

[22] Peng, T., Kerrien, E., Berger, M.O.: A shape-based framework to segmentation of tongue contours from MRI data. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 662–665 (2010). DOI 10.1109/ICASSP.2010.5495123

[23] Raeesy, Z., Rueda, S., Udupa, J.K., Coleman, J.: Automatic segmentation of vocal tract MR images. In: IEEE 10th International Symposium on Biomedical Imaging (ISBI), pp. 1328–1331 (2013). DOI 10.1109/ISBI.2013.6556777

[24] Witten, D.M.: Penalized unsupervised learning with outliers. Statistics and its Interface **6**(2), 211–221 (2013). DOI 10.4310/SII.2013.v6.n2.a5

[25] Wuhrer, S., Lang, J., Tekieh, M., Shu, C.: Finite element based tracking of deforming surfaces. Graphical Models **77**, 1–17 (2015). DOI 10.1016/j.gmod.2014.10.002

Index