A hybrid approach to 3D tongue modeling from vocal tract MRI using unsupervised image segmentation and mesh deformation

Alexander Hewer[1–4], *Ingmar Steiner*[1–3], *Stefanie Wuhrer*[1]

[1]Cluster of Excellence Multimodal Computing and Interaction, Saarland University, Germany
[2]DFKI Language Technology Lab, Saarbrücken, Germany
[3]Computational Linguistics & Phonetics, Saarland University, Germany
[4]Saarbrücken Graduate School of Computer Science, Germany

{hewer|steiner}@coli.uni-saarland.de, swuhrer@mmci.uni-saarland.de

Abstract

Vocal tract magnetic resonance imaging (MRI) has become one of the preferred imaging modalities for the analysis of human speech production. However, the raw image data must be segmented before further analysis can take place. This paper describes a hybrid approach to extract a 3D tongue model from 3D or 2D MRI scans of the vocal tract during speech, which combines unsupervised image segmentation with a mesh deformation technique. An efficient, minimally supervised segmentation algorithm can also be used as an alternative to provide a robust fallback in certain isolated cases. Both image segmentation algorithms produce a point cloud, which is completed and registered by deforming a template mesh to the data. Since the mesh deformation can be applied even with a sparse point cloud, it is possible to extract realistic 3D tongue shapes even from the 2D video frames of real-time MRI. Our approach is applied to several sets of available MRI data and yields promising results.

Index Terms: vocal tract MRI, image segmentation, 3D tongue model

1. Introduction

Analyzing the vocal tract, particularly articulators such as the tongue, during speech is of great interest in the areas of speech science and speech processing. Following technological advances in recent years, magnetic resonance imaging (MRI) can now be regarded as the state-of-the-art modality for imaging the vocal tract, due to its non-invasive and non-hazardous nature. However, since the resulting data contains the entire field of view, image segmentation is required before the shape and movements of the articulators can be processed for analysis. In particular, the extraction of high-level representations of the vocal tract surface is desirable. An example of such a representation is a polygon mesh, which can be used in various fields of application, such as computer graphics (talking heads or augmented-reality computer-assisted pronunciation training) or even articulatory speech synthesis, where it may be used to approximate the vocal tract area function.

1.1. Related Work

Extracting information about the shape of the tongue from MRI scans is an active field of research.

Peng et al. [1] employed an approach based on active contours [2] to find the contour of the tongue in a 2D mid-sagittal scan, using a previously trained shape model to control the evolution of the contour. Eryildirim and Berger [3] extended this approach to align the contour's end points to the corresponding extremities of the tongue. More recently, Raeesy et al. [4] demonstrated that oriented active shape models [5] can be trained to reliably identify the boundary of the tongue in 2D MRI scans. These methods rely on manually preparing a training set and are limited to the 2D case.

Lee et al. [6] presented a framework for minimally supervised tongue segmentation from 3D dynamic MRI. They used the random walker approach [7] as the base segmentation technique, which requires seeds manually provided by the user. Moreover, this approach only provides access to a low-level volume segmentation which has to be further processed.

Harandi et al. [8] employed a template-matching technique to generate a mesh representation of the tongue from 3D MRI scans. They used a mesh created by an expert from a source scan as their template, which is then deformed using color information to match a target scan. Specifically, they moved the mesh points in such a way that the color at the undeformed point in the source scan is similar to the deformed point in the target scan. Again, the approach is limited by requiring an expert to provide the templates.

1.2. Our contribution

In this work, we present a two-step approach that can be used in a minimally supervised way to obtain a mesh representation of the tongue from an MRI scan. First, we apply a segmentation method to extract the surface points of the tongue, without relying on a specific approach. Next, we use a template matching technique to align a mesh to this generated point cloud that operates exclusively on geometric information. The template is a generic mesh

automatically extracted from one 3D MRI scan. Thus, we do not rely on training a shape model. We also do not require an expert for creating the templates. Furthermore, by only using geometric information in the template matching step, we are able to use the same template mesh for scans of different persons and from different scanners. As the segmentation approach can be freely selected, we can choose for each scan the most suitable method.

2. Methods

In this section, we briefly outline the procedure used to obtain a mesh representation of the tongue contained in a selected MRI scan of the vocal tract.

We may interpret a scan as an image $f : \Omega \to \mathbb{R}^+$. Here, $\Omega \subset \mathbb{R}^n$ is a discrete set of the points where the measurements were taken by the MRI scanner with $n \in \{2,3\}$, and $f(\vec{x})$ represents the gray-value at point $\vec{x} \in \Omega$.

To extract the surface points of the tongue, we first find a partition of $\Omega = \Omega_O \cup \Omega_B$. The object region Ω_O is required to contain all points belonging to the tongue. However, it is allowed to consist also of regions related to other organic tissue. This relaxation is needed because in some scans no boundary may be detected between tongue and adjacent tissue. The background region Ω_B then represents the locations we have no interest in, e.g., air, or other tissue not belonging to the tongue. We observe that the color information in a scan can be exploited to distinguish soft tissue from material like air and bones, which motivates the idea to use a segmentation technique based on color information. In the following sections, we briefly present two different approaches used for this task.

2.1. Chan-Vese algorithm

The main approach used for image segmentation is the method of Chan and Vese [9]. Although it can be seen as a minimally supervised approach because it requires an initialization, we later on use it in an unsupervised fashion by always initializing it in the same way for all experiments. The needed initialization is a rough contour C that separates Ω into two regions: the region Ω_O enclosed by C and the region Ω_B that is always located outside of C. Essentially, the approach then evolves this provided contour such that the gray-value variance inside each region is minimized:

$$E_{\mathrm{CV}}(C) = \sum_{X \in \{\Omega_O, \Omega_B\}} \left(\sum_{\vec{x} \in X} \left(f(\vec{x}) - \mu_X \right)^2 \right) \qquad (1)$$

where Ω_O and Ω_B are the regions induced by C and μ_X represents the average gray-value in region X. As the average gray-value is global information, this approach can be considered a global method. In order to minimize the energy, we apply a standard scheme [9].

2.2. Graph cut algorithm

The graph cut technique [10] is a minimally supervised segmentation method that we use as a robust fallback alternative to our main method. It requires the user to provide two annotation sets $O \subset \Omega$ and $B \subset \Omega$ with $O \cap B = \emptyset$. O contains points the user wants to be part of Ω_O and B the ones that should be contained in Ω_B. In essence, the approach then finds Ω_O with $O \in \Omega_B$ and Ω_B with $B \in \Omega_B$ such that the gray-value similarity between neighboring points belonging to the same set is maximized:

$$E_{\mathrm{GC}}(\Omega_O, \Omega_B) = \sum_{X \in \{\Omega_O, \Omega_B\}} \left(\sum_{\vec{x} \in X} \sum_{\vec{y} \in \mathcal{N}(\vec{x}, X)} \psi\big(f(\vec{x}), f(\vec{y})\big) \right) \qquad (2)$$

where $\mathcal{N}(\vec{x}, X)$ are the neighbors of $\vec{x}$ contained in X and $\psi(a,b) := \exp(-|a-b|)$ in our case. Thus, we see that the graph cut technique is a local method because it finds the partition by using neighborhood information. We use the algorithm in [11] to obtain Ω_O and Ω_B as the maximizer of (2). However, in situations where the maximizer is not unique, the algorithm may output a partition $\Omega = \Omega_O \cup \Omega_B \cup \Omega_U$ with $\Omega_U \neq \emptyset$. In order to get the desired partition $\Omega = \Omega_O \cup \Omega_B$, we post-process the unassigned points in Ω_U as follows: Add $\vec{x} \in \Omega_U$ to Ω_O if its nearest neighbor $\min_{\vec{y} \in \Omega \setminus \Omega_U} d(\vec{x}, \vec{y})$ is located inside Ω_O with $d(\vec{x}, \vec{y})$ denoting a distance measure between $\vec{x}$ and $\vec{y}$. Otherwise, add $\vec{x}$ to Ω_B.

2.3. Mesh deformation

After obtaining a partition $\Omega = \Omega_O \cup \Omega_B$, we can compute the surface information as follows: First, we extract the surface points $P := \{\vec{p}_i\}$ of Ω_O, i.e., points $\vec{p}_i \in \Omega_O$ that are adjacent to at least one point $\vec{q} \in \Omega_B$. Additionally, we compute surface normals $N := \{\vec{n}_i\}$ for P such that $\vec{n}_i \in N$ is the surface normal at $\vec{p}_i \in P$. In order to eliminate the ambiguity of surface normals, we choose them in such a way that they are pointing towards the inside of Ω_O. We remark that P may also contain surface points belonging to other articulators than the tongue, due to the relaxation we formulated earlier for Ω_O.

Finally, we use the method of Wuhrer et al. [12] to deform a template mesh $M := (V, F)$ to match the point cloud data P. Here, $V := \{\vec{v}_i\}$ denotes the vertex set of the mesh with $\vec{v}_i \in \mathbb{R}^3$ 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 $\vec{v}_i$ by minimizing the following energy:

$$E_{\mathrm{Def}}(A) = \sum_{v_i \in V} \Bigg(\alpha \operatorname{dist}_D\left(A_i(\vec{v}_i), \operatorname*{arg\,min}_{\vec{p}_j \in P} \|A_i(\vec{v}_i) - \vec{p}_j\|\right) + \beta \sum_{v_j \in \mathcal{N}(v_i)} \operatorname{dist}_S\left(A_i, A_j\right) \Bigg) \qquad (3)$$

The term $\operatorname{dist}_D(\cdot)$ weighted by $\alpha > 0$ measures the distance between the transformed vertex $A_i(\vec{v}_i)$ and the normal plane at its nearest neighbor. Minimizing $\operatorname{dist}_D(\cdot)$ will move the mesh towards the point cloud P. The second term, $\operatorname{dist}_S(\cdot)$ weighted by $\beta > 0$, will generate energy if the rigid body motion A_i differs from the ones in

the neighborhood $\mathcal{N}(\vec{v}_i)$ around $\vec{v}_i$. Its minimization will prevent an alignment to data points that would distort the shape of the template too much, which helps to keep the mesh away from points not belonging to the surface of the tongue. Finally, we apply the found rigid body motions in A to the corresponding vertices to obtain the deformed mesh.

In order to find a minimizer A, we use a similar strategy like [12]: As a preprocessing step, we perform an automatic rigid alignment to find a good position for the template that is near the point cloud. Afterwards, we minimize a series of energies $E^t_{\text{Def}}(A^t)$ where $t \in [1, t_{\max}]$. The energy E^t_{Def} differs from E_{Def} in the following way: In E^t_{Def}, the transformed vertex in $\text{dist}_D(\cdot)$ is computed by using the minimizer of the previous energy: $A^{t-1}_i(\vec{v}_i)$. As $\arg\min_{\vec{p}_j \in P}(\cdot)$ then does not depend on A^t, the energy becomes differentiable and we can use a quasi-Newton technique [13] to compute the minimizer. In contrast to [12], we fix the weight β for all E^t_{Def} because the weights used in the experiments are not high enough to apply their relaxation technique. Eventually, we obtain A as $A^{t_{\max}}$. For A^0 that is needed in E^1_{Def}, we are using the identity: $A^0_i(\vec{v}_i) = \vec{v}_i$. Furthermore, the optimization process is embedded in a coarse-to-fine strategy to account for large deformations [12].

3. Experiments

In order to assess the performance of our proposed framework, we performed experiments on three different datasets where we always applied the following procedure: First, we manually place the template at a position centered in the tongue region of the currently selected dataset. Next, we apply the Chan-Vese method to each scan of the dataset, always with the same initial contour, an n-dimensional axis-aligned ellipsoid ($n \in \{2,3\}$) located at the center of the scan. The length of its semi-axis $\vec{a}_i$ is given by $\|\vec{a}_i\| = 15h_i$ with $i \in [1,n]$ where h_i is the distance between two points in the current dataset along that axis. Finally, we align our generic mesh (consisting of 17 091 vertices) to each point cloud by applying the template matching method. Here, we briefly present the datasets, the experiments, and the parameters used for the template matching.

A demonstration video is provided as supplementary material in the proceedings.

3.1. Static 3D MRI

The dataset of Baker [14] contains static 3D MRI scans of a male speaker producing 25 different phonemes, as well as several non-speech vocal tract configurations. Each scan consists of 44 sagittal slices with a spatial resolution of 320×320 pixels (corresponding to a pixel size of $1.1875 \times 1.1875\,\text{mm}^2$) and a slice thickness of 1.2 mm. Furthermore, for each scan the identified mid-sagittal slice is provided. As the scans show the entire head of the speaker, we cropped each slice to a 100×90 pixel region of interest containing only the vocal tract. Furthermore, we only considered the speech related scans.

First, we processed the whole 3D scans where we used $\alpha = 1$, $\beta = 3$ and $t_{\max} = 10$ in the template matching approach.

Afterwards, we used only the identified mid-sagittal slices in our framework to obtain the meshes. In this experiment, we used the same parameters as in the first one.

3.2. Movie of Real-time 2D MRI

We used supplementary material from Niebergall et al. [15] as our second dataset.[1] It consists of a video of real-time 2D MRI showing the mid-sagittal slice of a male speaker uttering a short text, with a temporal resolution of 33 ms and a duration of 21 s. Each video frame represents an upsampled version of a 128×128 pixels scan with a pixel size of $1.5 \times 1.5\,\text{mm}^2$.

Here, we processed frames 76 to 434 where we set $\alpha = 1$, $\beta = 4$ and $t_{\max} = 20$ in the template matching.

3.3. Corpus of real-time 2D MRI

The third data set is the updated dynamic 2D MRI recording for subject M1 in the USC-TIMIT database [16], which has a temporal resolution of 43.148 ms. Each frame shows the mid-sagittal slice of a single speaker and consists of 68×68 pixels with a pixel size of $2.9 \times 2.9\,\text{mm}^2$. Since the scans suffer from vignetting artifacts (i.e., the corners of the field of view are significantly darker than the center region), we cropped each frame to a selected region of interest to remove a large part of this effect. Additionally, we applied a flat field correction by manually creating a correction image and compositing it with the cropped frames.

This time, we processed all frames and used $\alpha = 1$, $\beta = 4$ and $t_{\max} = 5$.

4. Results and Discussion

4.1. Experiment on 3D Data

In the case of static 3D MRI, we performed a qualitative evaluation of the 17 results where our framework succeeded as follows: For each mesh, we computed at each vertex the error by measuring the distance to its nearest neighbor in the corresponding point cloud data. Finally, we computed the cumulative error for all meshes. We see in Figure 2 that our method produces satisfying results: Approximately 70 percent of the errors are below 2 mm.

However, phonemes like /k/ or /q/ lead to problems in our framework: First of all, their associated tongue shapes differ very much from our generic mesh, which may keep the template matching from aligning it correctly. Moreover, the Chan-Vese method fails to detect a part of the tongue-palate boundary where they are in contact, because it is not clearly visible in the data in these cases. However, it is able to identify the boundary between palate and nasal cavity, which is then included in

[1] http://upload.wikimedia.org/wikipedia/commons/4/4a/Real-time_mri_speaking_30fps.ogv

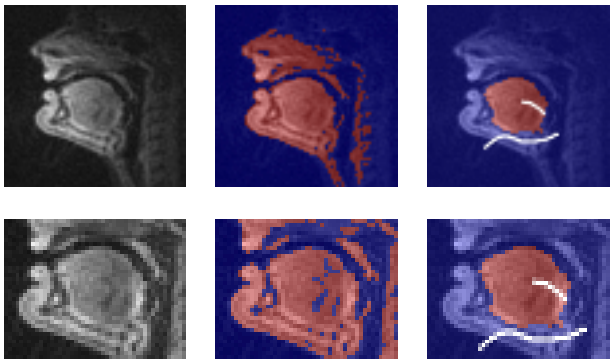

Figure 1: Segmentation results for frame 28 of the USC-TIMIT dataset with (**top row**) and without (**bottom row**) vignetting artifacts. Ω_O is shown in red and Ω_B in blue. **Left column**: Input images. **Center column**: Chan-Vese results with standard initialization. **Right column**: Graph cut results with annotations for respective regions colored in white.

the point cloud. This causes the template matching to move the mesh to the boundary between nasal cavity and palate.

4.2. Experiments on 2D Data

In the 2D case, we first evaluated the quality of the meshes obtained from the mid-sagittal slice where we proceeded like in the 3D case: We selected the results for the same 17 scans and computed the cumulative error by comparing the meshes to the corresponding full point cloud data of the 3D scan. We see in Figure 2 that the results are slightly worse than in the first experiment but still acceptable.

Afterwards, we processed the real-time 2D MRI datasets and investigated the temporal evolution of the derived meshes. Here, we discovered that the mesh sequence was suffering from severe temporal noise, which can be explained by the fact that we did not exploit any temporal coherence between the frames. We also saw in this evolution that mostly the mid-sagittal region was moving. Moreover, we also encountered intermittent frames where the mesh alignment was incorrect. Both problems can be seen as the consequence of using very sparse data in the template matching.

In the case of the USC-TIMIT dataset, we faced a segmentation related issue: As the Chan-Vese method makes use of global information to find a segmentation, it has trouble to cope with the globally changing illumination in each frame that is caused by vignetting artifacts. However, the minimally-supervised graph cut approach that uses local information is able to produce a suitable segmentation with only two broad annotations. We observe in Figure 1 that removing the vignetting artifacts improves the result of Chan-Vese with respect to the extractable surface information of the tongue, which allows us to apply our standard unsupervised procedure. As expected, the flat field correction has almost no effect on the

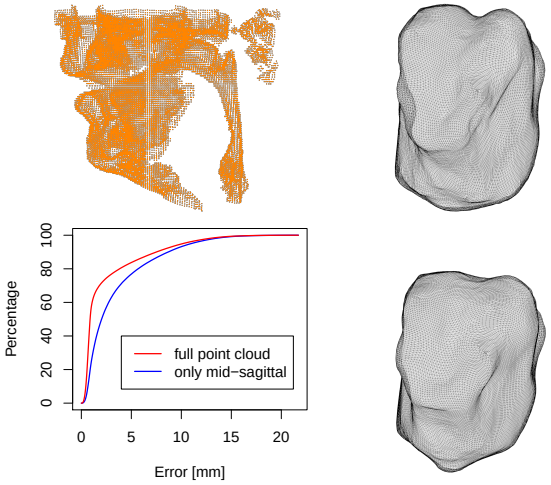

Figure 2: (**Top**) Point cloud extracted from [α] scan in [14] and template mesh deformed to fit. (**Bottom**) Cumulative errors for the two experiments on the Baker dataset and template mesh deformed using *only mid-sagittal* slice.

performance of the graph cut method.

5. Conclusion and Future Work

In this work, we presented a minimally supervised approach to extract polygon tongue meshes from vocal tract MRI scans. We saw that the proposed method was able to produce promising results. However, we also identified some issues which will be addressed in the future.

For example, we are currently investigating how to handle problematic phonemes like /k/. Here, we plan to reconstruct the missing palate information by using a second point cloud from a different vocal tract configuration or a full palate trace obtained by electromagnetic articulography (EMA) [17].

Moreover, we are planning to enhance the framework to train a statistical model with 3D data, which can then be used in the template matching to improve the results with sparse point clouds (e.g., 2D scans). Stone et al. [18] successfully used a similar approach to predict coronal tongue profiles from mid-sagittal tongue contours. Additionally, we also want to explore a more data-driven approach for improving the results in 2D: In particular, we want to combine lateral EMA information with a point cloud originating from a mid-sagittal scan, which may provide the template matching with sufficient information to align both the mid-sagittal and the lateral region of the mesh.

In order to cope with the temporal noise we encountered for 2D real-time MRI, we are thinking about applying a temporal smoothing. We also want to investigate if we can improve the results by sharing information between consecutive frames: Here, we plan to use the result of the current frame as the template for the next frame, since it resembles the shape of the tongue in the next frame more closely than the generic template.

References

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

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

[3] A. Eryildirim and M.-O. Berger, "A guided approach for automatic segmentation and modeling of the vocal tract in MRI images," in *European Signal Processing Conference (EUSIPCO)*, 2011. [Online] http://hal.inria.fr/inria-00630642/

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

[5] J. Liu and J. K. Udupa, "Oriented active shape models," *IEEE Transactions on Medical Imaging*, vol. 28, no. 4, pp. 571–584, 2009. doi:10.1109/TMI.2008.2007820

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

[7] L. Grady, "Random walks for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1768–1783, 2006. doi:10.1109/TPAMI.2006.233

[8] N. M. Harandi, R. Abugharbieh, and S. Fels, "3D segmentation of the tongue in MRI: a minimally interactive model-based approach," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 2014. doi:10.1080/21681163.2013.864958

[9] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001. doi:10.1109/83.902291

[10] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient ND image segmentation," *International Journal of Computer Vision*, vol. 70, no. 2, pp. 109–131, 2006. doi:10.1007/s11263-006-7934-5

[11] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004. doi:10.1109/TPAMI.2004.60

[12] S. Wuhrer, J. Lang, M. Tekieh, and C. Shu, "Finite element based tracking of deforming surfaces," 2013, arXiv:1306.4478.

[13] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical programming*, vol. 45, no. 1-3, pp. 503–528, 1989. doi:10.1007/BF01589116

[14] A. Baker. (2011) A biomechanical tongue model for speech production based on MRI live speaker data. [Online] http://www.adambaker.org/qmu.php

[15] A. Niebergall, S. Zhang, E. Kunay *et al.*, "Real-time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction," *Magnetic Resonance in Medicine*, vol. 69, no. 2, pp. 477–485, 2013. doi:10.1002/mrm.24276

[16] S. Narayanan, A. Toutios, V. Ramanarayanan *et al.*, "USC-TIMIT: A database of multimodal speech production data," USC, Tech. Rep., 2013. [Online] http://sail.usc.edu/span/usc-timit/usctimit_report.pdf

[17] P. Hoole and A. Zierdt, "Five-dimensional articulography," in *Speech motor control: New developments in basic and applied research*, B. Maassen and P. van Lieshout, Eds. Oxford University Press, 2010, ch. 20, pp. 331–349. doi:10.1093/acprof:oso/9780199235797.003.0020

[18] M. Stone, M. Epstein, M. Li, and C. Kambhamettu, "Predicting 3D tongue shapes from midsagittal contours," in *Speech production: Models, phonetic processes, and techniques*, J. Harrington and M. Tabain, Eds. New York, NY: Psychology Press, 2006, ch. 18, pp. 315–330.