A multilinear tongue model derived from speech related MRI data of the human vocal tract

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 51 (2018) 68–92

www.elsevier.com/locate/csl

A multilinear tongue model derived from speech related MRI data of the human vocal tract☆

Alexander Hewer[a,b,c,*], Stefanie Wuhrer[d], Ingmar Steiner[a,b], Korin Richmond[e]

[a] *Cluster of Excellence "Multimodal Computing and Interaction", Saarland University, Saarbrücken, Germany*
[b] *German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany*
[c] *Saarbrücken Graduate School of Computer Science, Saarbrücken, Germany*
[d] *INRIA Grenoble Rhône-Alpes, France*
[e] *Centre for Speech Technology Research, University of Edinburgh, UK*

Received 14 July 2017; received in revised form 10 December 2017; accepted 15 February 2018
Available online 21 February 2018

Abstract

We present a multilinear statistical model of the human tongue that captures anatomical and tongue pose related shape variations separately. The model is derived from 3D magnetic resonance imaging data of 11 speakers sustaining speech related vocal tract configurations. To extract model parameters, we use a minimally supervised method based on an image segmentation approach and a template fitting technique. Furthermore, we use image denoising to deal with possibly corrupt data, palate surface information reconstruction to handle palatal tongue contacts, and a bootstrap strategy to refine the obtained shapes. Our evaluation shows that, by limiting the degrees of freedom for the anatomical and speech related variations, to 5 and 4, respectively, we obtain a model that can reliably register unknown data while avoiding overfitting effects. Furthermore, we show that it can be used to generate plausible tongue animation by tracking sparse motion capture data.

Keywords: Tongue; Vocal tract; MRI; Statistical model; Shape analysis

1. Introduction

1.1. Motivation

As one of the main articulators, the tongue plays a major role in human speech production. In speech science, it is therefore of great interest to understand its shape and motion during speech articulation. To this end, we want to derive a tongue model that offers the following features: it is a three-dimensional articulatory model that only uses a few degrees of freedom (DoF) that influence the shape of the tongue. Furthermore, the DoF of the model are split

☆ This paper has been recommended for acceptance by Shrikanth Narayanan.

* Corresponding author.
E-mail address: hewer@coli.uni-saarland.de (A. Hewer), stefanie.wuhrer@inria.fr (S. Wuhrer), steiner@coli.uni-saarland.de (I. Steiner), korin@cstr.ed.ac.uk (K. Richmond).

http://dx.doi.org/10.1016/j.csl.2018.02.001

into two sets: One set controls the anatomical shape of the tongue and the other, the speech related tongue pose. This is important because the articulation strategy may depend on the anatomy of the speaker (Johnson et al., 1993; Ladefoged and Broadbent, 1957; Honda et al., 1996; Brunner et al., 2009; Fuchs et al., 2008; Rudy and Yunusova, 2013; Weirich and Fuchs, 2013; Weirich et al., 2013; Yunusova et al., 2012). As shape representation of the model, we select a polygon mesh that can easily be used in various fields of applications: in computer graphics, such meshes are used to generate animations of complex objects (Botsch et al., 2010) or to model objects of highly complex geometry and topology. Additionally, polygon models have been used in speech processing to generate acoustical simulations (Blandin et al., 2015).

The model itself can be used to equip virtual avatars for multimodal spoken interaction with a more natural animation of the tongue. In this regard, we note that it is of vital importance to synthesize the correct motion for the speech audio: McGurk and MacDonald (1976) found that inconsistencies between visible mouth motions and audible speech may cause the speech to be perceived incorrectly. Moreover, information about tongue motion can be applied in computer-aided pronunciation training (CAPT) to provide the user with visual information on how to move the tongue to produce a specific sound (Engwall, 2008). It can also be employed in an articulatory speech synthesis framework to help approximate the vocal tract area function or it can be used to estimate the full tongue shape from sparse data, such as electromagnetic articulography (EMA) measurements. Finally, the model can also help to perform speaker normalization, that is, investigate only shape variations of an articulation that is independent of the speaker anatomy. In particular, such a model allows speaker adaptation, which is useful in the aforementioned areas of applications. For example, in audiovisual speech synthesis, CAPT, and articulatory speech synthesis, it is vital to replicate the speaker's specific tongue shape to match the remaining anatomy; the tongue should not leave the mouth or penetrate the palate during articulation. Additionally, for CAPT, the speaker's tongue anatomy influences the articulatory strategy of the speaker. Providing the incorrect strategy could confuse the subject, especially if real-time feedback was provided from EMA data. In the case of estimating the tongue shape from EMA, using the wrong anatomy of the tongue may keep the model from registering the EMA data correctly.

For completeness, we want to mention another class of tongue models, so-called biomechanical models. Such models aim at simulating the entire tongue body, including the internal muscle activities. This property is useful for, e.g., simulating laryngoscopy (Rodrigues et al., 2001), investigating the consequences of surgery (Buchaillard et al., 2007), or for studying muscle activation during speech (Buchaillard et al., 2009; Wu et al., 2014). However, for our target application areas, such a model may be regarded as too complex.

1.2. Methodology

We want to derive DoF that are speech related. Thus, we need to analyze meshes extracted from actual speech production data. However, most of the articulators are contained inside the human mouth and therefore partially or completely hidden from view. This means that traditional imaging modalities based on light, e.g., photography, are of limited use for acquiring the desired shape information for analysis. Currently, magnetic resonance imaging (MRI) can be regarded as the state-of-the-art technique for investigating the interior of the human vocal tract during speech. It is non-invasive and non-hazardous to the subject, and in contrast to ultrasound (US) or EMA, it is able to provide dense volumetric measurements. Moreover, a lot of previous work has focused on adapting the MRI method to the needs of speech research. The main issue in early studies (Baer et al., 1991) was the long acquisition time, which forced subjects to maintain the vocal tract configuration for a long time with brief interruptions: one scan took around 3 minutes.

Subsequent advances in MRI scanners made it possible to acquire 3D time-evolving models of the vocal tract (Foldvik, 1993; Shadle et al., 1999) and 2D MRI movies with up to 5 frames per second (fps) (Demolin et al., 2000). An approach to real-time magnetic resonance imaging (rtMRI) recording of the vocal tract with synchronized audio was presented by Narayanan et al. (2004), which offered a frame rate of up to 24 fps and thus enabled the examination of the dynamics of fluent speech using MRI. This method also applied noise cancellation to deal with the scanner noise. More recent methods (Kim et al., 2009; Scott et al., 2012; Niebergall et al., 2013; Burdumy et al., 2015; Fu et al., 2015; Elie et al., 2016; Lingala et al., 2016; 2017) further reduced the acquisition time and improved the quality of obtained scans. For example, Lingala et al. (2017) reported rtMRI scanning at 83 fps for a single slice, or 27 fps for three slices.

As a lot of speech production data can nowadays be obtained by means of MRI techniques in a short amount of time, the mesh extraction process from the individual scans should be as minimally supervised as possible, as doing it manually takes a lot of time and typically requires anatomical expertise. For the analysis of extracted meshes, statistical methods like principal component analysis (PCA) can be used. Such a method provides access to a statistical model that uses a low-dimensional subspace to represent the shape of the corresponding object, and is also able to estimate the plausibility of such a shape. This approach requires a labeled database that shows the object under consideration in many different shapes that are related to the motion to be observed. Polygon meshes already have been successfully used for statistical shape analysis of, e.g., human bodies (Allen et al., 2003), faces (Blanz and Vetter, 1999), or tongues (Badin and Serrurier, 2006). As we want to separate the anatomy related variations from the ones related to the speech related tongue pose, we use a method that leads to a so-called multilinear model. This statistical model type is able capture those different types of variation separately.

1.3. Related work

A sizable body of research has focused on analyzing the vocal tract shape during speech production using different modalities, including X-ray, electropalatography (EPG), cineradiography (CR), ultrasound, cone beam computed tomography (CBCT), real-time magnetic resonance imaging, or computed tomography (CT). In Table 1, we provide an overview of previous studies.

Even some of the earliest studies aimed at analyzing the anatomical and speech related shape variations by using multiple subjects; Harshman et al. (1977) investigated these variations in 2D X-ray data. Nowadays, this imaging modality is no longer used for this purpose, due to the dangers of the ionizing radiation involved. Narayanan et al. (1995, 1997) analyzed shape variabilities using 3D MRI data. Analysis on 2D MRI was conducted by Hoole et al. (2000), Ananthakrishnan et al. (2010), and Valdés Vargas et al. (2012b,a). Zheng et al. (2003) performed this analysis on sparse sets of 65 points that were manually extracted from 3D MRI scans. Kaburagi (2015) used PCA to analyze the vocal tract area functions of ten speakers obtained from MRI. The work by Woo et al. (2015b) used

Table 1
Overview of several studies that have investigated shape variabilities of the vocal tract. We list the modality (or modalities) used, the analyzed data representation, and the number of subjects taking part in the corresponding study.

Work	Modality	Analyzed data	Subjects
Mermelstein (1973)	X-ray	2D contours	1
Harshman et al. (1977)	X-ray	2D contours	5
Baer et al. (1991)	MRI	Vocal tract area functions and shapes	4
Stone and Lele (1992)	US	Fitted polynomial functions	1
Narayanan et al. (1995)	MRI	Shapes	4
Stone and Lundberg (1996)	3D US + EPG	Interpolated meshes	1
Tiede et al. (1996)	MRI	Cross-section shapes	1
Narayanan et al. (1997)	MRI + EPG	Shapes	4
Badin et al. (1998)	MRI + CR	Meshes	1
Engwall and Badin (1999)	MRI	Meshes + 2D contours	1
Engwall (2000)	MRI	Meshes	1
Hoole et al. (2000)	MRI	2D contours	9
Kröger et al. (2000)	MRI	Vocal tract area functions	1
Beautemps et al. (2001)	CR + labio-film	2D contours	1
Badin et al. (2002)	MRI + video	Meshes	1
Zheng et al. (2003)	MRI	Sparse 3D point clouds	5
Badin and Serrurier (2006)	MRI + CT	Meshes	1
Geng and Mooshammer (2009)	EMA	Flesh points	7
Ananthakrishnan et al. (2010)	MRI	2D contours	3
Valdés Vargas et al. (2012b)	MRI	2D contours	7
Toutios and Narayanan (2015)	rtMRI	2D contours	1
Kaburagi (2015)	MRI	Vocal tract area functions	10
Woo et al. (2015b)	Dynamic MRI	Images	18
Woo et al. (2015a)	MRI	Deformation fields	20
Stone et al. (2016)	MRI	Muscle architectures	14
Fang et al. (2016)	MRI + CBCT	Meshes	1
Serrurier et al. (2017)	MRI	2D contours	11

dynamic MRI to build a spatio-temporal atlas of the vocal tract. Woo et al. (2015a) analyzed a high resolution atlas of the vocal tract using PCA. In the study by Stone et al. (2016), the muscle architectures of different subjects were investigated. Speaker normalization was performed by Geng and Mooshammer (2009) and Serrurier et al. (2017).

Shape variations related to the anatomy of the subject are also of interest in the field of biomechanical models: Bijar et al. (2016) presented an atlas-based automatic approach to generate subject-specific finite element tongue meshes. Harandi et al. (2017) used cine MRI to derive speaker-specific biomechanical models.

For our purposes, we need to analyze the anatomical and speech related variations in 3D meshes. Initial work investigating these variations obtained from MRI data of 9 speakers was presented by Hoole et al. (2003), but neither evaluated nor published (Hoole, personal communication). Moreover, work that focused on the speech related shape variations of a more dense 3D representation of the tongue required manual annotation of the MRI data, which makes it less feasible for large collections of data. Work exists that aims at facilitating tongue shape extraction from MRI data. However, such approaches are often limited because they are restricted to 2D (Peng et al., 2010; Eryildirim and Berger, 2011; Raeesy et al., 2013), produce only a low-level volume segmentation (Lee et al., 2013), or require an anatomical expert to provide the tongue templates (Harandi et al., 2014).

1.4. Our contribution

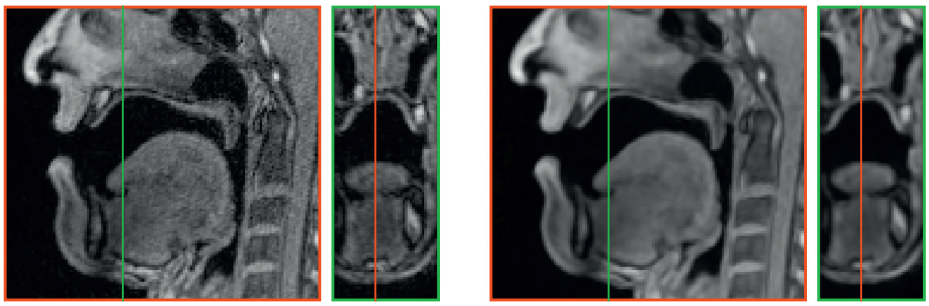
In this paper, we present a significant extension of our previous work (Hewer et al., 2014). Originally, we combined an image segmentation method and a template matching approach to extract tongue meshes from MRI data in a minimally supervised way. The new features of our framework can be summarized as follows: we use an image denoising method to deal with possibly corrupt data. Moreover, we modify the template matching approach to better handle volumetric point clouds. Additionally, the user can provide landmarks to guide the template matching process. Furthermore, we integrate a strategy for restoring any tongue surface information that might be missing due to contact between the hard palate and tongue. This improvement increases the number of tongue shape configurations we can register. Additionally, the framework is augmented by use of a bootstrapping strategy, which refines the quality of the obtained shape meshes. Finally, it can now be used to derive a multilinear statistical model that captures almost the entire complex 3D surface geometry of the tongue and allows the anatomy and pose related variations to be modified separately.

We use our new framework to register speech related tongue shapes of 11 speakers and examine the obtained model with respect to its compactness, generalization, and specificity properties. In the case of the specificity analysis, we investigate those parts of the tongue surface mesh that play an important role during human articulation. The results of our experiments motivate us to choose a model with 5 DoF for the anatomy and 4 for the speech related tongue pose. Moreover, we successfully use the obtained model for tracking motion capture data of the tongue.

The remainder of the paper is organized as follows: in the next section, we start by describing how surface information of the vocal tract can be extracted from a given 3D MRI scan by denoising it and applying an image segmentation approach. We proceed by discussing the modified template matching approach in Section 3 and also present the used templates of our approach. Section 4 is dedicated to describing how we estimate a tongue mesh from the surface information by using the template fitting. In this part, we present the bootstrapping strategy used, and our approach to restore missing tongue surface information that is caused by contact between tongue and hard palate. Next, we turn to the multilinear tongue model in Section 5. In this section, we outline how the acquired mesh collection can be aligned to only contain speech and anatomy related tongue shape variations, and how the model is derived. We then turn to the evaluation of our approach in Section 6, where we apply it to MRI scans of two datasets. Afterwards, we investigate the validity of our obtained mesh collection by means of a preference test in Section 7. Furthermore, we conduct experiments to evaluate the compactness, generalization, and specificity properties of the acquired model in Section 8. In Section 9, we use the model for tracking speech related motion capture data of a new speaker. Finally, we conclude in Section 10 and outline possible future work.

2. Extracting surface information from MRI

As a first step, we want to extract a point cloud $Q := \{(\mathbf{q}_i, \mathbf{n}_i)\}$ from an MRI scan that contains the surface points $\mathbf{q}_i$ and the associated normals $\mathbf{n}_i$ of the major articulators and related tissue. We use a purely geometric representation



(a) Raw MRI images (b) Smoothed MRI images

Fig. 1. Raw MRI scan (a) and smoothed version (b). The left image of each pair shows a sagittal slice, the other one a coronal slice.

of this surface information because it is easy to combine two point clouds into a single one. This is helpful in situations where we want to restore missing information in a point cloud Q that is present in another cloud Q^*.

As we are using image processing methods, we briefly describe how we treat a volumetric MRI scan as a 3D image. We may represent an MRI scan as a function

$$s : \Omega \to [s_{\min}, s_{\max}] \tag{1}$$

where $s_{\min}$ and $s_{\max}$ are real values. Here, $\Omega \subset \mathbb{R}^3$ is a discrete rectangular domain consisting of the sample positions where the scanner took the measurements. These coordinates are arranged on a regular grid with grid spacings h_x, h_y, and h_z. We say that $s(\mathbf{q})$ represents the measured nuclear magnetic resonance (NMR)[1] at sample position $\mathbf{q} \in \Omega$. This scan can be interpreted as a gray-scale 3D image

$$f : \Omega \to [0, 255] \tag{2}$$

by applying a quantization operator to the NMR values that scales them to a standard (8bit) gray-scale. We decided to use a standard visualization where bright and dark indicate a high and low NMR, respectively.

Fig. 1 shows two typical visualizations of such a representation: a sagittal slice and a coronal one showing an (x, y)-plane and a (y, z)-plane of the scan image, respectively. As in general the original scan field of view contains much more information than just the vocal tract, we usually crop it to a selected region of interest. This reduces the memory requirements and the processing time of our framework. By inspecting the scan, we observe that the data is degraded due to measurement noise. As a remedy, we apply a 3D variant of edge-enhancing diffusion (Weickert, 1998) to the image. An example result of the approach can be inspected in Fig. 1b. We see that the noise was removed and structural information like edges were preserved and enhanced.

We now want to extract a point cloud Q of the desired surface information from the denoised MRI scan. First, we detect the spatial support of the region whose surface information we want to derive. That is, we want to find a partition

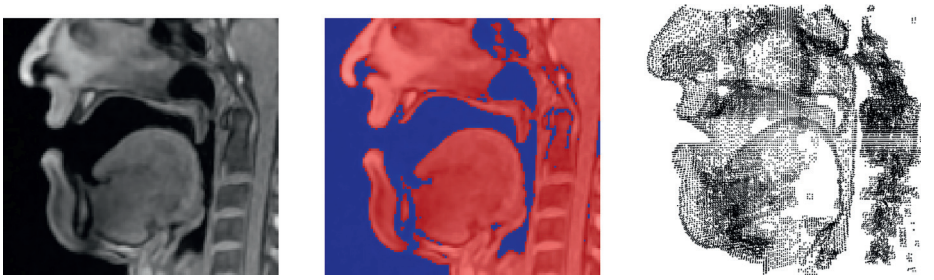
$$\Omega = \Omega_O \cup \Omega_B \tag{3}$$

such that Ω_O contains the region of the major articulators and related tissue and $\Omega_B = \Omega \backslash \Omega_O$ everything else. By inspecting the denoised data, we notice that tissue can be distinguished from non-tissue, such as air and bone, by using color information. This observation motivates the use of image segmentation methods that make use of such a feature. In our case, we decided to use the method introduced by Otsu (1979) to perform this task as it is fully automatic. An example segmentation can be seen in Fig. 2b.

As we are interested in the shape information of the surface, we proceed by extracting the surface points of the tissue from the obtained partition. We call $\mathbf{q}_i \in \Omega_O$ a surface point if at least one of its neighbors is part of Ω_B. Additionally, we use the partition to estimate normal information at the extracted surface points.

The obtained surface points and associated normals are then assembled in a point cloud. An example of such a point cloud can be inspected in Fig. 2c.

[1] Correlated with hydrogen molecule density, i.e., high for soft tissue, low for bone and air.



(a) One sagittal slice (b) Slice segmentation (c) Resulting point cloud

Fig. 2. Extracting surface information from an MRI scan: one sagittal slice (a), its segmentation (b), and a rendering of the resulting point cloud (c). For visibility reasons, the cloud was clipped and decimated.

3. Template matching

Next, we want to estimate the surface of the desired articulator from such a point cloud Q. Here, we use a polygon mesh $M := (V, F)$ as the surface representation. The set $V := \{\mathbf{v}_i\}$ with $\mathbf{v}_i \in \mathbb{R}^3$ is called the vertex set of the mesh. The other set, F, is the face set of our mesh.

We observe that a point cloud Q is a loose collection of points. In general, this collection contains more information than the desired articulator and there might be holes in the point cloud with missing data. However, a subset of Q implicitly represents the surface of the desired articulator.

In order to identify this subset and estimate the articulator shape from it, we can apply a template fitting technique.

Given a template mesh $M = (V, F)$ that resembles the desired articulator and a point cloud Q, it finds a set $A := \{A_i\}$ where $A_i : \mathbb{R}^3 \to \mathbb{R}^3$ is a rigid body motion for the vertex $\mathbf{v}_i \in V$, such that the deformed mesh $M^* = (V^*, F)$ with $V^* := \{A_i(\mathbf{v}_i)\}$ is near the point cloud data Q.

The template matching finds this set A of deformations by minimizing the energy

$$E_{\text{Def}}(A) = \alpha\, E_{\text{data}}(A) + \beta\, E_{\text{smooth}}(A) + \gamma\, E_{\text{landmark}}(A) \tag{4}$$

The data term

$$E_{\text{data}}(A) := \frac{1}{|V^L|} \sum_{\mathbf{v}_i \in V^L} \| A_i(\mathbf{v}_i) - \arg\min_{\mathbf{p}_j \in Q} \| A_i(\mathbf{v}_i) - \mathbf{p}_j \| \|^2 \tag{5}$$

measures the distance between the deformed vertices $A_i(\mathbf{v}_i)$ and their nearest neighbors $\mathbf{p}_j$ in the point cloud Q. Thus, it is minimized if applying A to the mesh moves it towards some points in the point cloud. In this term, V^L refers to the set of vertices that are not landmarks. The smoothness term

$$E_{\text{smooth}}(A) := \frac{1}{|V|} \sum_{v_i \in V} \left(\sum_{v_j \in \mathcal{N}(v_i)} \| A_i - A_j \|^2 \right) \tag{6}$$

evaluates the differences between the rigid body motion A_i at vertex $\mathbf{v}_i$ and the motions A_j in its neighborhood $\mathcal{N}(\mathbf{v}_i)$. This means that it penalizes deformations that alter the original shape of the template. Finally, the landmark term

$$E_{\text{landmark}}(A) := \frac{1}{|L|} \sum_{(\mathbf{v}_i, \mathbf{p}_i) \in L} \| A_i(\mathbf{v}_i) - \mathbf{p}_i \|^2 \tag{7}$$

produces energy in proportion of how many correspondences between deformed landmark vertices $A_i(\mathbf{v}_i)$ and user-provided target points $\mathbf{p}_i$ of the landmark set $L := \{(\mathbf{v}_i, \mathbf{p}_i)\}$ are violated by the deformation. We remark that the used target points are not required to be part of the generated point cloud. As a convention, we always set the weight α to

1 in order to interpret the other parts of the energy in terms of the data nearness assumption: for example, using a value of $\beta = 10$ means that the smoothness term is ten times more important than the data term.

As the energy in (4) is not differentiable due to the data term, it is usually optimized by minimizing a series of energies $E^t_{\text{Def}}(A^t)$ where $t \in [1, t_{\max}]$. Each energy uses adapted weights β^t and γ^t:

$$\beta^t = \beta - (t-1)\frac{\beta - \beta_{\min}}{t_{\max} - 1} \tag{8}$$

$$\gamma^t = \gamma - (t-1)\frac{\gamma - \gamma_{\min}}{t_{\max} - 1} \tag{9}$$

where $\beta_{\min}$ and $\gamma_{\min}$ are set by the user.

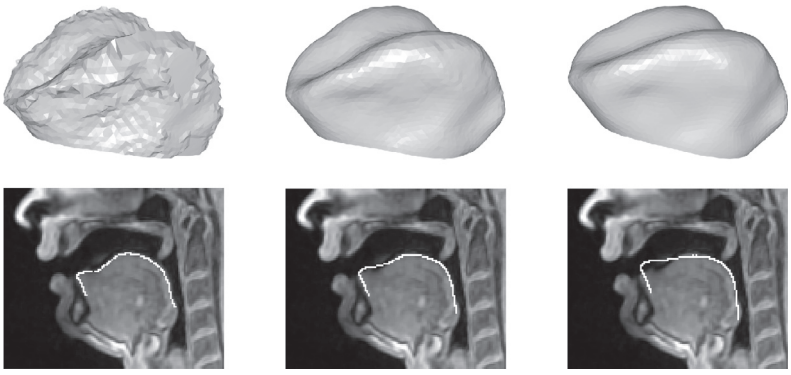
These parameters have to be carefully selected as they influence the last energy that is optimized. A value of $\beta_{\min}$ that is too high forces the approach to preserve the shape of the original template, which leads to an underfitting. Setting the value too small, on the other hand, causes an overfitting that produces many local shape artifacts on the resulting mesh. Fig. 3 shows results for different values of $\beta_{\min}$.

A similar statement holds true for $\gamma_{\min}$: using too small a value could move the template away from the desired landmark locations during the optimization. A value that is too high might overfit the landmark positions, which could cause problems if landmarks are wrongly placed, and lead to spike-like artifacts. In our approach, we mitigate such effects of wrongly placed landmarks by applying a smoothing operation after the template matching: we replace the measured rigid body motion A_i at the corresponding vertex with the average of the rigid body motions of vertices that are connected via an edge to this vertex. In Fig. 4, the effect of $\gamma_{\min}$ on the mesh can be inspected. Furthermore, we show how the smoothing operation improves the result. It is important to avoid noise or artifacts on the mesh because we want to prevent our tongue model from modeling this noise.

Originally, we used a standard heuristic (Allen et al., 2003; Li et al., 2009) to distinguish valid data observations from invalid ones in the optimization of E_{data}. In particular, we say that $\mathbf{q}$ is a valid data point candidate for a deformed vertex $A_i(\mathbf{v}_i)$ if the Euclidean distance between both is not too large and if their normals do not differ too much from each other.

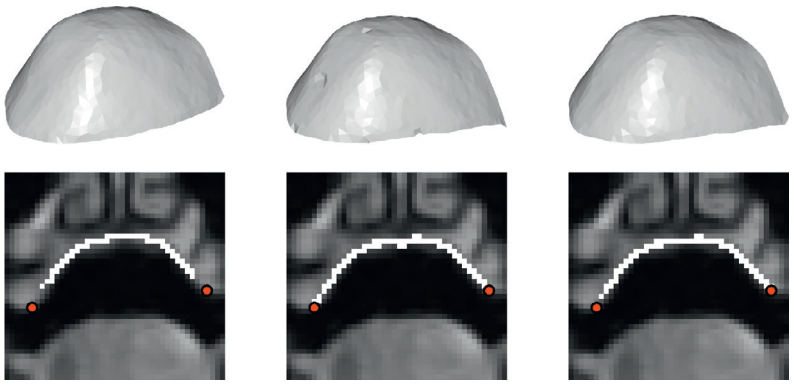
In this paper, we have modified this nearest neighbor heuristic somewhat: we now collect all valid data point candidates within a fixed radius and then select the best candidate that lies below the current mesh surface. If no such candidate exists below the surface, we will select the best one above it. This modification is intended to prevent the template mesh from getting stuck at unrelated points in the volumetric cloud during the optimization.

In our framework, we use two templates: one for the tongue and one for the hard palate. Both templates were extracted from MRI data by means of medical imaging software (Rosset et al., 2004). Afterwards, we made the templates symmetric to remove this particular bias towards the original speaker by mirroring the respective mesh at a selected center plane.



(a) $\beta_{\min} = 0.1$ (b) $\beta_{\min} = 5$ (c) $\beta_{\min} = 100$

Fig. 3. Effect of $\beta_{\min}$ on the resulting mesh. A low value (a) leads to an overfitting of the data and a very noisy mesh, whereas a high value causes underfitting and produces a very smooth result (c). Choosing an appropriate value provides a good compromise between data nearness and mesh quality (b).



(a) $\gamma_{\min} = 0$ (b) $\gamma_{\min} = 10$ (c) $\gamma_{\min} = 10$ and post smoothing

Fig. 4. Effect of $\gamma_{\min}$ on the resulting mesh. Setting it to 0 prevents the template matching from reaching the provided landmarks shown as red dots (a). Using the value 10 aligns the template to the wanted positions, but leads to spike-like artifacts (b). Applying a smoothing afterwards removes these spikes while keeping the template close to the landmarks (c).

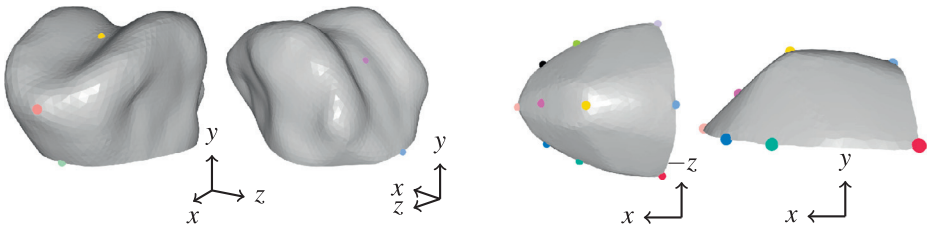


Fig. 5. Used templates with landmarks of the tongue (left) and hard palate (right).

The palate template consists of 994 vertices and 1828 faces with an average edge length of 1.4 mm. The tongue template contains 3100 vertices and 6102 faces with an average edge length of 1.8 mm. In our case, the tongue template does not contain the sublingual part. This means that the part below the line from the jaw to the epiglottis is missing, as well as the part below the tongue tip that is negligible for speech production.

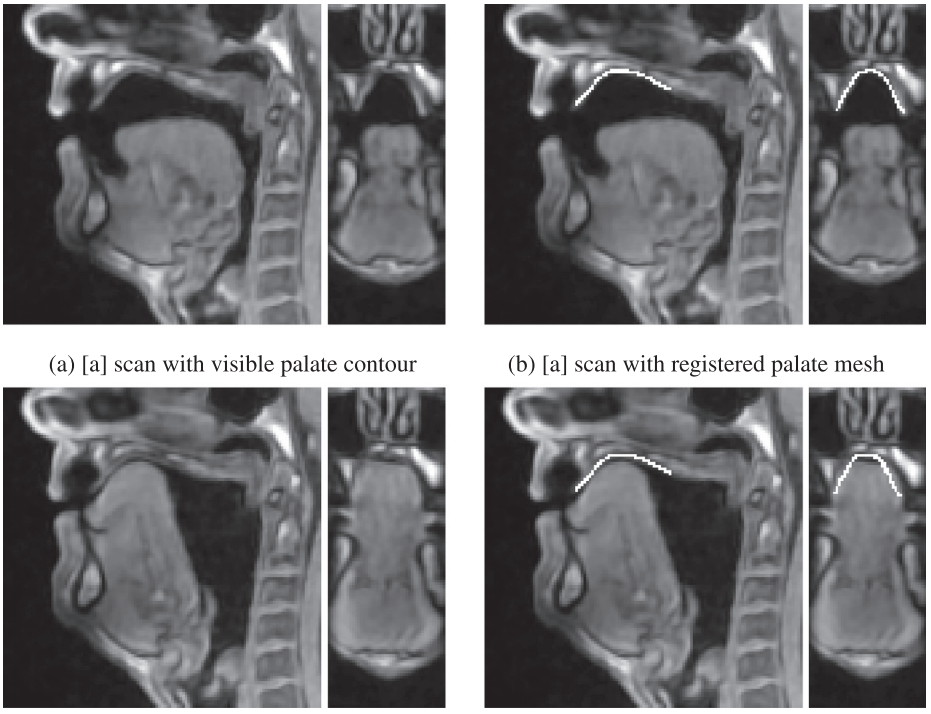
Both templates can be inspected in Fig. 5 together with the landmarks used.

4. Tongue and palate shape estimation

We first estimate the palate shape for each MRI scan. We select a scan for each speaker where the hard palate is clearly visible and perform template matching. In general, using a single template might produce sub-optimal results in some matching cases. In order to improve the results, we set up an iterative bootstrapping approach. In each iteration, we first compute a PCA model of the palate (Hewer et al., 2015) by using the results of the previous iteration. The palate model is derived in a way similar to the multilinear model in Section 5. This model is then fitted to each point cloud and the results are afterwards used as the initialization for the template matching.

After we have acquired the hard palate mesh for each speaker, we align this mesh to each scan of that speaker. This procedure serves the purpose of restoring tongue surface information that is missing due to contacts between tongue and palate as shown in Fig. 6.

In this context, we have to address the issue that the corresponding speaker might have moved between the scans. Fortunately, as the hard palate can only undergo rigid body transformations, we only have to estimate this type of motion. However, as the palate surface information might be partly missing, we fall back to NMR information for this task. To this end, we define the NMR profile set $E(M,f) \subset \mathbb{R}^{\ell}$ of a mesh M in a scan f. A profile $\mathbf{e}^i(M, f) \in E(M,$



(a) [a] scan with visible palate contour

(b) [a] scan with registered palate mesh

(c) [i] scan with palatal contact

(d) [i] scan with hard palate surface restored

Fig. 6. Palate reconstruction: scan of speaker pronouncing [a] (a), and registered hard palate surface (b). Palatal contact during the pronunciation of [i] (c), and result of restoring hard palate surface information (d). It should be observed that in (c), the hard palate appears to merge with the tongue, which presents a potential pitfall for manual annotation. The palate reconstruction, on the other hand, helps to identify the true palate contour.

f) is a vector such that its entries are given by

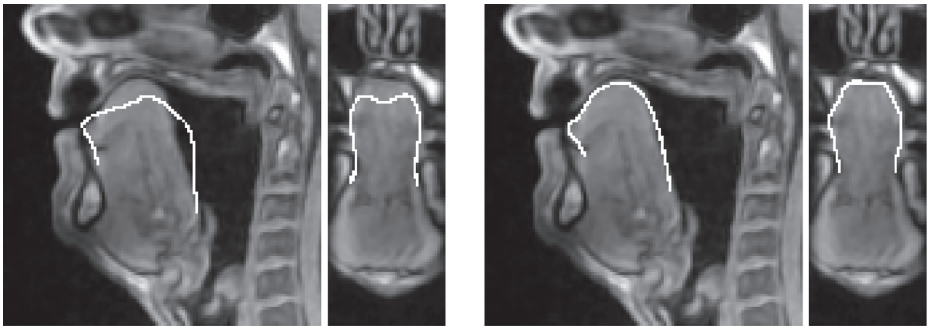
$$\mathbf{e}_j^i(M,f) = f(\mathbf{v}_i + j\ d\ \mathbf{n}_i) \tag{10}$$

where $\mathbf{v}_i$ is a mesh vertex, $\mathbf{n}_i$ its corresponding normal, and d the chosen sampling distance. We start above the palate surface in order to avoid taking samples in the possible contact area between tongue and palate. We can estimate the rigid body motion A for aligning a palate mesh M obtained from a scan f to a scan g by maximizing the energy:

$$E_{\text{palate}}(A) = \sum_{i \in J(V)} \text{NCC}\left(\mathbf{e}^i(M,f), \mathbf{e}^i\left(A(M), g\right)\right) \tag{11}$$

where $J(V)$ is the index set of the vertex set V, NCC the normalized cross-correlation between its operands, and $A(M)$ the transformed mesh. We decided to use the normalized cross-correlation (NCC) as a similarity measure because it is known to be robust against noise and brightness differences. Furthermore, the NCC between NMR profiles was already successfully used in a nearest neighbor heuristic for template matching (Harandi et al., 2014). Results of this alignment approach can be seen in Fig. 6. In this figure, we also present the result for a scan without contact between tongue and palate, to show the actual contour of the hard palate of the speaker, which is not visible in the other scan.

We now inject this aligned palate mesh information into the point cloud of the corresponding scan in order to restore missing tongue surface information by using the palate surface as a replacement. Additionally, we use the aligned mesh as a boundary to remove points in the point cloud above the palate that are unrelated to the tongue. Finally, we use template matching to extract the tongue shape from the corresponding modified point cloud. As in the palate case, we use a bootstrapping strategy to refine the results. This time, we use a multilinear model in each



(a) Initial template (b) Bootstrapping result

Fig. 7. Effect of the bootstrapping strategy. Initial template matching has trouble correctly registering the tongue shape (a). Bootstrapping improves the result (b).

iteration as a statistical prior that is described in the next section. The benefits of this bootstrapping operation can be seen in Fig. 7.

5. Multilinear tongue model

Having obtained a collection of tongue meshes, we now want to derive a function

$$M : S \times P \rightarrow \mathbb{M} \tag{12}$$

where $\mathbb{M}$ is a set of meshes. The set $S \subseteq \mathbb{R}^{\tilde{m}}$ consists of coordinates $\mathbf{s}$ that describe a speaker's anatomical tongue shape. The set $P \subseteq \mathbb{R}^{\tilde{n}}$ contains coordinates $\mathbf{p}$ that determine the shape for a specific speech related tongue pose. We call S the speaker subspace and P the pose subspace of the model. Meshes $M \in \mathbb{M}$ should have the same face set as our tongue template mesh. Their vertex sets $V(\mathbf{s}, \mathbf{p})$, however, may differ from the original template with respect to their vertex positions.

5.1. Preparing the training mesh collection

Deriving the function in (12) implies we want to analyze only the anatomical and speech related variations in our mesh collection, which means we have to remove all other variations present. The Procrustes alignment technique (Dryden and Mardia, 1998) is a method suitable for this task as it may be used to remove any translational and rotational differences among the meshes in the collection. However, applying this technique directly to the acquired tongue meshes might eliminate critical information related to, e.g., the speech related tongue pose. This is, for example, due to the fact that the tongue also undergoes translational and rotational motions because it is connected to the lower jaw.

As a remedy, we apply the Procrustes alignment to the hard palate meshes we obtained earlier to remove translational and rotational differences between the speakers that are unrelated to the tongue motion. The results are afterwards used as a reference to align the tongue meshes. To this end, we use a speaker's palate mesh that was earlier aligned to the corresponding scan. We then estimate the rigid transformation that maps this aligned palate mesh to its Procrustes variant and apply the same motion to the corresponding tongue mesh. By doing so, we remove any translational and rotational differences related to head motion or position differences without eliminating any speech or anatomy-specific information.

Finally, we have to ensure that for each speaker the meshes for all selected poses are available. This might be needed because some MRI scans may be corrupt or even missing in the used dataset. In these cases, we reconstruct a missing pose shape of a speaker by averaging available data: first, we compute the average shape of all meshes that are present for the speaker. Afterwards, we compute the mean shape of all meshes that are available for this specific pose from the other speakers. Finally, both meshes are averaged again. In the literature, there are more sophisticated

methods to restore missing information, such as HALRTC (Liu et al., 2013). In our case, however, this averaging approach was sufficient.

5.2. Model construction

In order to derive our desired function in (12), we need to analyze the anatomical and speech related variations separately. In previous work (Harshman et al., 1977; Hoole et al., 2000; 2003; Ananthakrishnan et al., 2010; Valdés Vargas et al., 2012b; 2012a; Zheng et al., 2003), the PARAFAC method (Harshman, 1970), also known as CANDECOMP, has often been used to perform this analysis. This method decomposes a tensor into a sum of r rank-1 tensors where r is provided by the user. Therefore, this technique can be regarded as an extension of the singular value decomposition to tensors. However, there are reports in the literature of certain issues with this method: Hoole et al. (2000) found that it might be difficult to find reliable solutions; Valdés Vargas et al. (2012a) pointed out that the PARAFAC decomposition requires numerous components to describe the observed data in a satisfactory way, which limits its usefulness as a dimensionality reduction method; moreover, De Silva and Lim (2008) discovered that the associated standard approximation problem is mathematically ill-posed, which can lead to the problem of diverging components in a numerical setting.

Another suitable method is the Tucker decomposition (Tucker, 1966), which is sometimes also called higher-order singular value decomposition (HOSVD). This method computes the orthonormal spaces of a tensor associated with its modes. It may be regarded as a more flexible variant of the PARAFAC method (Kiers and Krijnen, 1991) and has previously been used to analyze 2D tongue shape data (Valdés Vargas et al., 2012b). To avoid the issues with PARAFAC, we decided to use the Tucker decomposition to analyze our data. We follow the approach of Bolkart and Wuhrer (2015) who used it to analyze the variations of human faces in different expressions. To this end, we first turn our tongue meshes into feature vectors by serializing the vertex sets V into vectors $\mathbf{f}_i$. Then, we compute the mean μ, and center the vectors. Afterwards, we organize those centered vectors in a tensor $A \in \mathbb{R}^{m\times n\times k}$. Here, we refer to the first mode of the tensor as the *speaker* mode where m represents the number of speakers, to the second mode as *pose* mode with n being the number of different tongue poses, and to the third mode as the *vertex* mode with k representing the dimension of the vectors $\mathbf{f}_i$.

The HOSVD makes use of the fact that A can be decomposed as follows:

$$A = C \times_1 U_1 \times_2 U_2 \tag{13}$$

In our case, the row vectors of $U_1 \in \mathbb{R}^{m\times m}$ are coordinates in our speaker space S that determine the anatomical shape for each of the original speakers. A similar observation applies to $U_2 \in \mathbb{R}^{n\times n}$ where the row vectors are coordinates in the pose space P. The tensor $C \in \mathbb{R}^{m\times n\times k}$ is the core tensor of the decomposition that acts as a link between S and P. The operation $C \times_n U$ is called the nth mode multiplication of the tensor C with the matrix U.

The core tensor is the multilinear model we can use to create our function in (12): essentially, given $\mathbf{s} \in S$ and $\mathbf{p} \in P$, we can use C to generate serialized vertex sets that represent the generated shape as follows:

$$v(\mathbf{s}, \mathbf{p}) = \mu + C \times_1 \mathbf{s} \times_2 \mathbf{p} \tag{14}$$

By letting $V(\mathbf{s}, \mathbf{p})$ be the vertex set reconstructed from $v(\mathbf{s}, \mathbf{p})$, we can finally define our function as:

$$M(\mathbf{s}, \mathbf{p}) \coloneqq (V(\mathbf{s}, \mathbf{p}), F) \tag{15}$$

where F is the face set of our original template. We remark that the dimensionality of the speaker and pose subspaces can be truncated to remove shape variations that may be considered negligible or related to noise. This means that our subspaces have dimensionalities $\widetilde{m} \leq m$ and $\widetilde{n} \leq n$.

5.3. Model fitting

We can use this derived model to register data, for example a point cloud Q. This time, we want to optimize for the parameters $\mathbf{s} \in S$ and $\mathbf{p} \in P$ that best describe the speaker anatomy and tongue pose that is represented in the data. To this end, we minimize the following energy:

$$E_{\text{Fit}}(\mathbf{s}, \mathbf{p}) = \alpha\, E_{\text{data}}(\mathbf{s}, \mathbf{p}) + \gamma\, E_{\text{landmark}}(\mathbf{s}, \mathbf{p}) \tag{16}$$

where the data and landmark terms are equivalent in their modeling idea to their counterparts in the template matching case. Furthermore, we use the same nearest neighbor heuristic and optimization approach as in the template matching. This time, the weights for both terms remain constant during the optimization of the energy series. However, if the corresponding neighbor for each vertex is known, they can be set directly and only one energy has to be minimized in that case. This is the case, for example, if our target is a tongue mesh of the same size, such that a one-to-one correspondence between model mesh and target mesh exists.

It is common to limit the admissible values for **s** and **p** to avoid highly unlikely shapes. In particular, we limit each entry of **s** and **p** individually to an interval

$$[m_i - h\ \sigma_i, m_i + h\ \sigma_i] \tag{17}$$

where σ_i is the standard deviation of the corresponding variation in the used mesh collection, and m_i is the corresponding entry of the mean coordinate in the respective subspace. Finally, $h \in \mathbb{R}^+$ is a scale factor.

The above energy can also be used to fit a PCA model: in this case, the energy depends only on one parameter.

6. Deriving a tongue model

Our next goal is to apply the described framework to MRI data and obtain a tongue model. We do not undertake a quantitative evaluation of the tongue meshes extracted from the MRI scans, a decision that is necessitated by the fact that we are working with real-world data, and thus have no ground truth reference available. There exists the possibility of manually annotating the MRI scans to create a reference solution, but this procedure is very time consuming and expensive if the number of scans in the dataset is very large. Moreover, such hand-labeling is error prone and the anatomical expert(s) involved may introduce a subjective bias. Instead, we chose to perform a qualitative analysis: we inspected the results manually as a post-hoc analysis in order to decide whether they are acceptable. In particular, we projected the registered tongue meshes onto to the corresponding scans and verified that the mesh surface was close to the true tongue contour. Videos showing such projections slice by slice can be found in the files S1–S7 of the supplementary material.

6.1. Background information about data used

We use two datasets to derive our model: the dataset of Baker (2011) and the full dataset of the Ultrax project (Ultrax: Real-time tongue tracking for speech therapy using ultrasound, 2014), which provides us with data of 12 speakers in total.

The Ultrax dataset consists of static MRI scans of 11 adult speakers of British English where 7 are female and 4 are male. All speakers are phonetically trained and were recorded while sustaining the vocal tract configuration for different phones for around 20 s. For each speaker, 13 speech related scans are available that correspond to the phone set [i, e, ɛ, a, ɑ, ʌ, ɔ, o, u, ʉ, ə, s, ʃ].

The Baker dataset was recorded as part of the Ultrax project, but released separately. It contains 25 scans of one male speaker that are speech related and represent different articulatory configurations.

The data was recorded at the Clinical Research Imaging Centre in Edinburgh using a Siemens Verio 3T scanner; the scans were acquired with an echo time of 0.93 ms and a repetition time of 2.36 ms. The individual scans consist of 44 sagittal slices with a thickness of 1.2 mm and a slice size of 320 × 320 pixels. The grid spacings are $h_x = h_y = 1.1875$ mm and $h_z = 1.2$ mm.

For our analysis, we decided to exclude one speaker of the Ultrax dataset that showed a high activity of the soft palate, which caused problems in our framework. Furthermore, we use the whole phone set that was recorded for the Ultrax data. However, the Baker dataset is lacking scans for the phones [a, ɔ, ʉ, ə, s, ʃ] where the shape information has to be reconstructed.

In total, we are using the shape information of 11 speakers with 13 different tongue shape configurations.

In Fig. 8, some of these shapes can be seen. This means that we arrive at a tensor $A \in \mathbb{R}^{11\times13\times9300}$ where the dimension of the vertex mode is determined by the vertex count of the tongue template we are using.

It is important to state that a tongue model derived from this data might lack some typical tongue pose related variations because the phonetic coverage of the underlying dataset can be considered incomplete. Due to these data related constraints, we believe that this model's applicability to some of the application areas we described earlier,

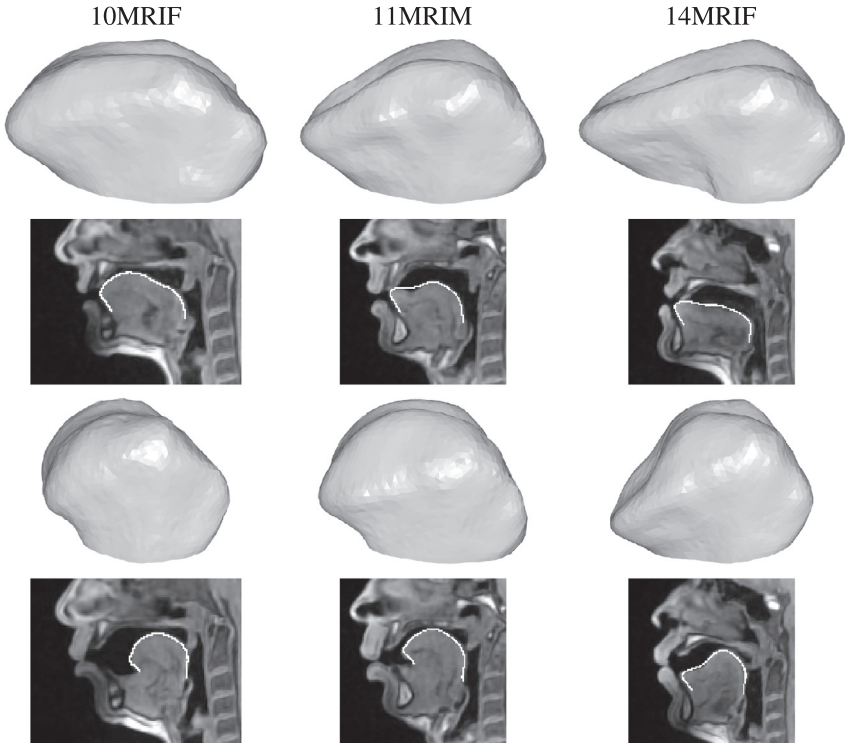


Fig. 8. Shape variabilities among the speakers 10MRIF, 11MRIM, and 14MRIF in the dataset used. The images show the extracted mesh above a sagittal slice of the corresponding scan for the phones [s] (top) and [u] (bottom).

may be limited. However, assembling and using MRI recordings is a demanding task: First, it requires access to an MRI scanner, along with specialized equipment and technical staff experienced in performing such recordings. Second, appropriate speakers have to be found who are phonetically trained and whose articulation is not impacted by the MRI scanner. Moreover, the use and distribution of acquired medical imaging data is governed by strict and extensive data privacy protection: For example, the raw data cannot normally be published, and using it for research purposes requires the explicit consent of the corresponding speaker. Under these considerations, we argue that even a limited dataset is a valuable resource and can lead to useful results.

6.2. Applied settings

In the following, we describe the settings we applied in our framework to extract the mesh collection from the given data. In general, the settings were chosen in such a way that the results were visually satisfactory. This means that we wanted the meshes to be close to the tongue contour that is visible in the MRI scan. Moreover, we tried to reduce overfitting artifacts, such as spikes or noise, on the resulting shapes.

In the case of template matching, we used $\alpha = 1$, $\beta = 10$, $\beta_{\text{min}} = 6$, and $\gamma = 10$. Thus, we start with a high weight for the smoothness and landmark terms to drive the template to the correct neighborhood at the beginning of the optimization. The template matching for the tongue used $\gamma_{\text{min}} = 0$ to dampen the effects of falsely placed landmarks. We used $\gamma_{\text{min}} = 10$ for the palate matching to ensure that its extremities were correctly aligned. For the model fitting that is applied during the bootstrapping, we used $\alpha = \gamma = 1$. In the nearest neighbor heuristic, we set the search radius to 4 mm and limited the maximally allowed angle difference between the normals to 60°. The optimization for the template matching used a series of 40 energies, the one for the model fitting applied a series of 10 energies to find the minimizer. For the palate alignment, we decided to use sufficiently long profiles with a length of $\ell = 15$ and a sampling distance of $d = 1$ mm.

In the bootstrapping strategy, we applied the following iteration amounts: we used one iteration for the hard palate and 5 iterations for the tongue. For the scale factor h in the model fitting, we used 0.5 for the tongue and 1 for the palate in order to prevent overfitting.

The landmarks needed for the hard palate and the tongue were placed on the MRI scans by a user who is not an anatomical expert.

7. Evaluation of the mesh extraction process

We mentioned earlier that no ground truth is available for our dataset. Before analyzing and using the acquired tongue model, however, we wanted to evaluate the validity of the meshes we extracted from the MRI data. To this end, we designed a web-based preference test and elicited the opinion of speech experts. The goal of this experiment was to investigate whether the experts agreed with our own informal assessment of the acquired meshes.

7.1. Experiment setup

We prepared the following data for the experiment: for each of the 137 scans in our dataset, we created three versions of the same sagittal slice. One version showed the unannotated slice. The second version showed the slice with the tongue mesh contour after the initial template matching. The last version visualized the tongue mesh contour after the final bootstrapping. Afterwards, we randomly partitioned our scan set into 4 subsets of roughly equal size. These partitions were then randomly assigned to the participants such that overall, each scan was seen by 3–4 participants.

15 speech experts took part in the experiment. On average, they had 11 years of research experience with speech production data. Each participant was asked to view all scans of the assigned partition and to select the preferred annotated version of the shown sagittal slices. During the experiment, the individual methods that produced the results were hidden from the participants. Moreover, in order to prevent the participants from detecting any pattern in the presentation, the two annotated versions were always displayed in random order.

7.2. Results

The evaluation revealed that in 83.85% of the cases, the bootstrap result was preferred by the participants. We proceeded by investigating how these preferences were distributed among the different scans that were shown.

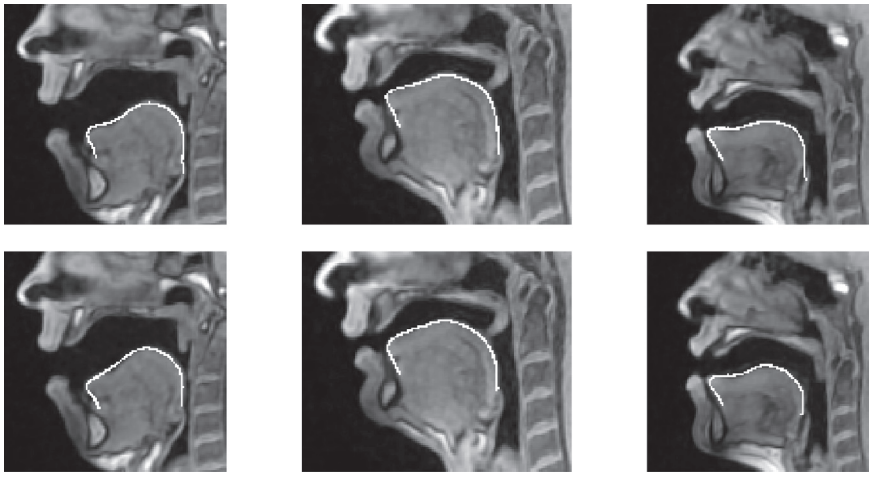
A plot summarizing our findings can be seen in Fig. 10. We see that for 19 scans, in 50% or more cases, the initial template matching was preferred over the bootstrapping result. For these individual scans, we inspected the displayed slices and discovered that the initial and bootstrapping versions were very similar. Moreover, the bootstrapping results seemed to slightly underestimate the tongue shape in the MRI scan in these cases, which might have caused the participants to choose the initial result. Examples of such cases are shown in Fig. 9.

7.3. Discussion

Overall, we draw two conclusions from the obtained results of the experiment. On the one hand, the relatively high acceptance rate of the obtained bootstrap results among the consulted experts justifies our decision to derive a tongue model from the corresponding mesh collection. On the other hand, however, we see also that there is still some room for improvement of our approach, which is conditioned on speaker-specific anatomy.

8. Model analysis

It is common to evaluate such statistical models by analyzing their compactness, generalization, and specificity (Styner et al., 2003) in order to find the optimal subspace dimensionality.



(a) [ɑ] scan of 11MRIM (b) [ʌ] scan of 01MRIM (c) [ə] scan of 14MRIF

Fig. 9. Examples for scans where the participants preferred the initial template matching result (top row) over the bootstrapping one (bottom row).

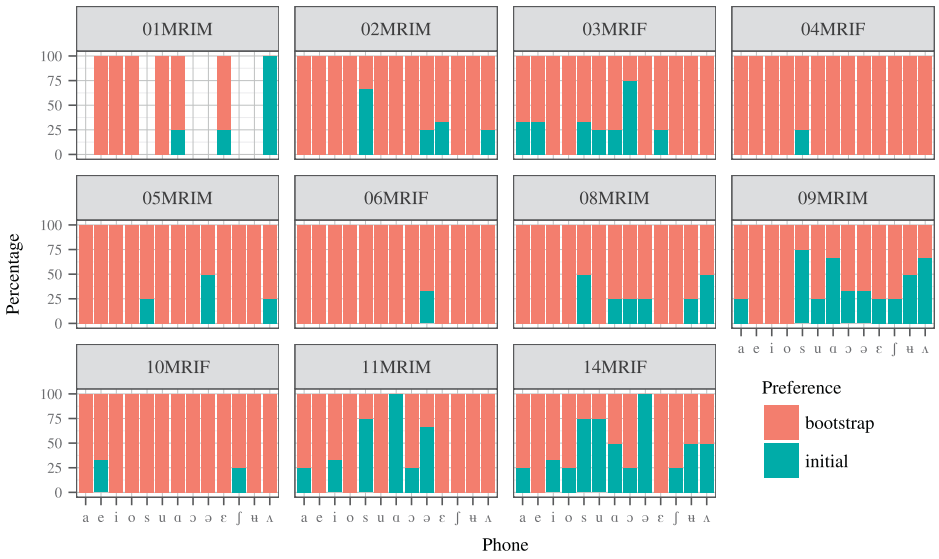


Fig. 10. Results of the preference test for each considered scan. Note that not all phonemes are available in the data for speaker 01MRIM. We grouped the scans by speaker to improve the visualization.

8.1. Compactness

Compactness investigates how much the individual components of **s** and **p** contribute to the description of the used training data. In Fig. 11, we see that using $\widetilde{m}=5$ is sufficient to represent 91% of data variability. Approximately the same holds for $\widetilde{n}=4$.

8.2. Generalization

Generalization measures how well the model can register data that was not part of the training. To evaluate the speaker generalization, we designed the following experiment: for each speaker, we derived a tongue model from the meshes of all other speakers and registered the tongue meshes of the excluded speaker. Afterwards, we measured

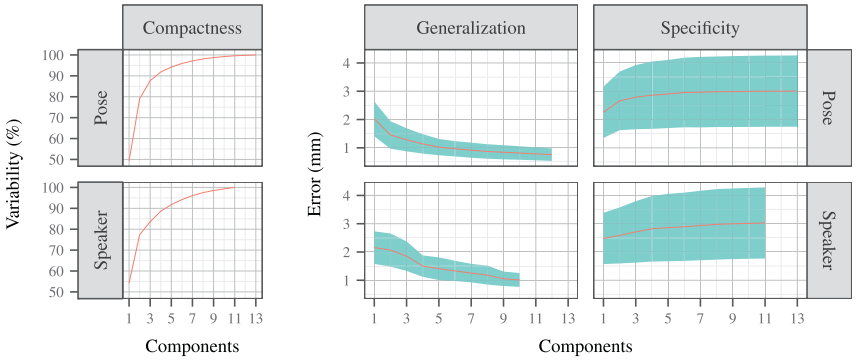


Fig. 11. Compactness (left), generalization (center), and specificity (right) of the model for the pose (top) and speaker subspace (bottom). For the generalization and specificity, we visualize the mean (red line) and the standard deviation (blue ribbon) of the experiments.

the average Euclidean distance between the registered meshes and the original ones. Additionally, we analyzed the fitting results for different values of $\widetilde{m}$. The dimensionality of the pose subspace was fixed to $\widetilde{n}=4$ during these experiments to prevent overfitting caused by this subspace.

In the analysis of the pose generalization, we used a similar approach: for each phone, we derived a tongue model from the meshes of all other phones and registered the tongue meshes of the excluded phone. In this case, the dimensionality of the speaker subspace was fixed to $\widetilde{m}=5$. The results of these experiments are depicted in Fig. 11. During this evaluation, we used the scale factor $h=2$ in the model fitting optimization. We observe that increasing the subspace dimensionality leads to better fitting results. The results of the generalization experiments show that only a few components of **p** and **s** are needed to reliably register unseen data, which implies that the model can adapt to new tongue anatomies or poses. In particular, for **p**, 3 components are enough to reach an average error that is slightly above the average measurement resolution of 1.2 mm of the MRI scan data. For **s**, 7 components are needed to reach this level of precision. Furthermore, we observe that a high number of components leads to errors below the average measurement resolution of the scan data, which can be considered as overfitting. We observe that the pose subspace has better generalization abilities than the speaker subspace. We suspect this might be related to redundancies in our training data: for example, the phone pairs [ʌ, ʉ], [e, i], and [e, ɛ] are similar to each other with respect to shape (Ladefoged, 1982). This means that excluding one still provides the model with enough information to capture the related variation.

8.3. Specificity

Specificity tries to assess how much randomly generated tongue shapes of the model differ from valid tongue shapes. This is essentially a measure for determining how specific the model is to the tongue. In particular, we wanted to investigate how large these differences were for the regions of the tongue mesh that are speech related. Fig. 12 shows an overview of those regions. To this end, we designed a few experiments where samples from the two subspaces were drawn randomly in order to generate a random tongue shape. In these experiments, we used the tongue meshes of all speakers as the collection of valid shapes. The first experiment investigated the specificity of the speaker subspace. The pose subspace is again fixed to $\widetilde{n}=4$ and the speaker subspace size was varied. For each value of $\widetilde{m}$, we generated random tongue shapes and evaluated the average Euclidean distance between the created mesh and the closest one in the mesh collection. In this comparison and distance evaluation, a region consisting of all speech related parts was considered. The same experiment was conducted for analyzing the specificity of the pose subspace where the speaker subspace size was set to $\widetilde{m}=5$. The results of both experiments can be inspected in Fig. 11. In these experiments, we see that increasing the subspace dimensionality leads to higher average Euclidean distances, which means that the model is becoming less specific. This could also be seen as an indicator that the higher dimensions are modeling the noise in the training meshes.

Finally, we wanted to find out how much the tongue shapes belonging to specific phones differ from the corresponding ones generated by the model. We performed for each phone the following experiment: we froze the coordinates in the pose subspace to the ones belonging to the given phone. Moreover, we only allowed the generated

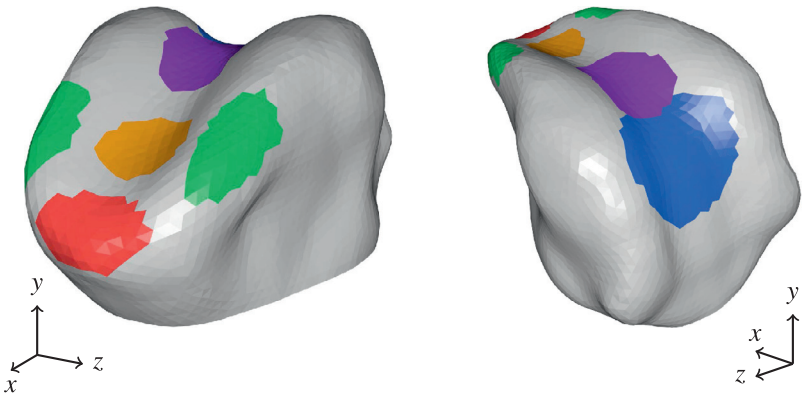


Fig. 12. Speech related regions of the tongue surface: Tongue tip (red), tongue blade (brown), tongue back (violet), tongue dorsum (blue), and the lateral regions (green).

meshes to be compared to meshes belonging to that phone. Then, for each dimensionality of the speaker subspace, we generated samples and computed the average Euclidean distance to the closest mesh. This time, in the distance evaluation and comparison, we used parts of the tongue that are considered critical for this specific phone (Jackson and Singampalli, 2009). For the vowels [i, e, ɛ, a, ɑ, ʌ, ɔ, o, u, ʉ, ə], we selected a region consisting of the tongue blade, tongue back, and tongue dorsum. The area for the sibilants [s, ʃ] contains the tongue tip and the tongue blade. The results of these experiments are shown in Fig. 13. In all specificity experiments, we generated 1000000 samples. In these experiments, we notice that the phone [ʉ] shows a significantly bad result in the fixed phone specificity evaluation, which might be related to its unusual role in the phonology of British English. We suspect that some speakers might have pronounced it inconsistently and applied different strategies, which led to a high variation in the data, which is then integrated into the model.

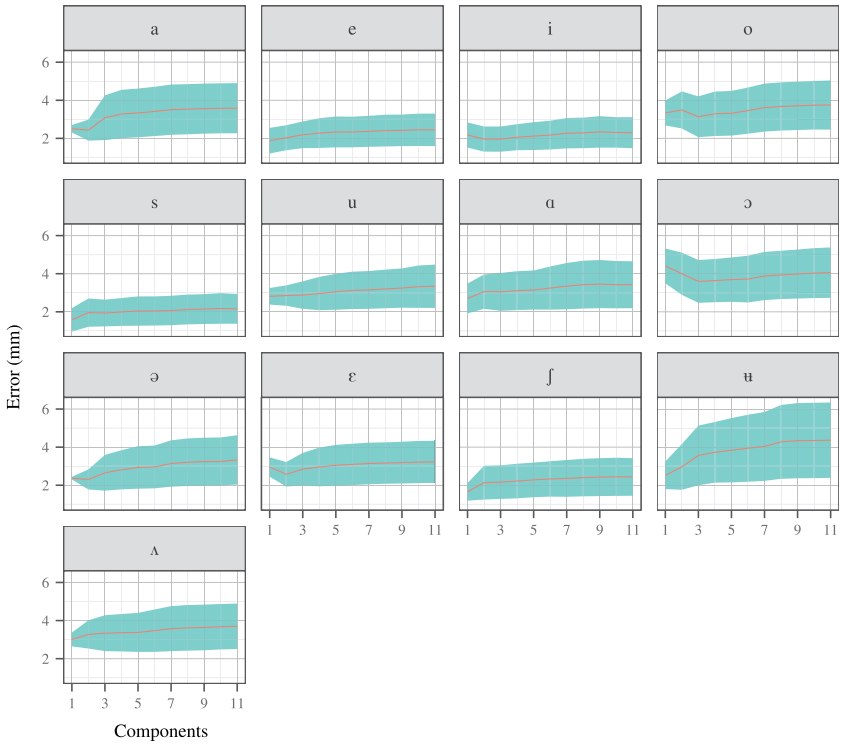


Fig. 13. Specificity results for the individual fixed phone experiments. Plots show mean (red line) and standard deviation (blue ribbon).

Overall, we decided that setting $\widetilde{m}=5$ and $\widetilde{n}=4$ provides a good compromise between specificity, generalization, and compactness. We note that this choice also limits the effects of overfitting.

9. Using the model for tracking motion capture data

After having derived our final model with $\widetilde{m}=5$ and $\widetilde{n}=4$, we investigated if it could be used to reliably track the tongue motion capture data of an unknown speaker and to generate a plausible animation from it. To this end, we decided to use EMA data from a previous study (Steiner et al., 2014). This study focuses on the German language and consists of speakers that are not part of the Ultrax data. EMA uses alternating electromagnetic fields to track the position of coils that are attached to specific points of interest, e.g., on the tongue surface; this modality can provide data with a high temporal resolution, but only gives access to a sparse set of points.

9.1. Data selection and setup for experiments

We selected the following data of the female subject *VP05* in the dataset: one recording that contains repeated consonant-vowel combinations of the consonants [f, s, ʃ, ç, x, ʁ, m, n, ŋ l] and the vowel [u]. Furthermore, we used an EMA recording of the German translation of the "Northwind and the Sun" passage, a standard specimen in phonetic research (International Phonetic Association, 1999). To combine this EMA data with our model, we had to register dynamic speech data of an unknown speaker containing new phonemes, which is a significant challenge. However, we want to point out that the phonetic inventories of the English and German languages are very similar. Thus, we think that the pose subspace of our tongue model should be compatible with motion capture data originating from producing German speech.

The raw EMA data was prepared as follows: first, we smoothed the data to dampen any high-frequency measurement noise. Afterwards, we removed rigid motion originating from head movements by using the three reference coils of the EMA data that were attached to suitable positions on the head. Finally, we used the palate shape and the bite plane of the subject to rigidly align the data to our tongue model in a semi-supervised way.

For our experiments, we chose the 3 coils that were placed at the tongue tip, the tongue blade, and the tongue dorsum, which lay roughly in the mid-sagittal plane of the tongue. We normalized this data, specifically, we projected the positional data into the mid-sagittal plane to guarantee this mid-sagittal property.

For our tracking experiments, we first had to find for each EMA coil a corresponding vertex on our tongue mesh. We used the following semi-supervised approach to determine these correspondences from one frame of the used EMA data: first, we sampled a random tongue shape from our model and initially determined for each coil the nearest neighbor on the mid-sagittal area of the tongue mesh. Then, we iteratively refined these correspondences by fitting the model and updating the nearest neighbors. We repeated the above two steps multiple times and kept the correspondences that achieved the smallest average distance between coil positions and their corresponding vertices. Afterwards, we visually compared the proposed correspondences with a photographic reference of the subject's tongue with the attached coils and reran the above approach until the correspondences were plausible. During the sampling and the correspondence optimization, we used the scale factor $h=1$ to avoid overfitting.

9.2. Experiments

In our experiments, we used the following energy to fit our model to the current EMA data frame:

$$E_{\text{Track}}(\mathbf{s},\mathbf{p})=\alpha\ E_{\text{data}}(\mathbf{s},\mathbf{p})+\beta\ E_{\text{bias}}(\mathbf{s},\mathbf{p})+\gamma\ E_{\text{coherence}}(\mathbf{s},\mathbf{p}) \tag{18}$$

The data term E_{data} measures the distances between the vertex locations and their corresponding EMA coil positions. The bias term E_{bias} penalizes deviations from the mean weights of the model. We added this term to the energy to provide the approach with information about the average tongue shape in order to cope with the sparsity of the data. Finally, the consistency term $E_{\text{coherence}}$ favors a temporal coherence between consecutive frames.

For all experiments, we used $\alpha=1$ and $\beta=\gamma=5$ to provide a good compromise between these model ideas. Furthermore, we set the scale factor h to 5 to give the approach sufficient freedom during the optimization.

In the first experiment, we optimized for $\mathbf{s}$ and $\mathbf{p}$. However, we know that the anatomy of the speaker should remain constant during the recordings. As it is unknown, we used a common approach to estimate the corresponding

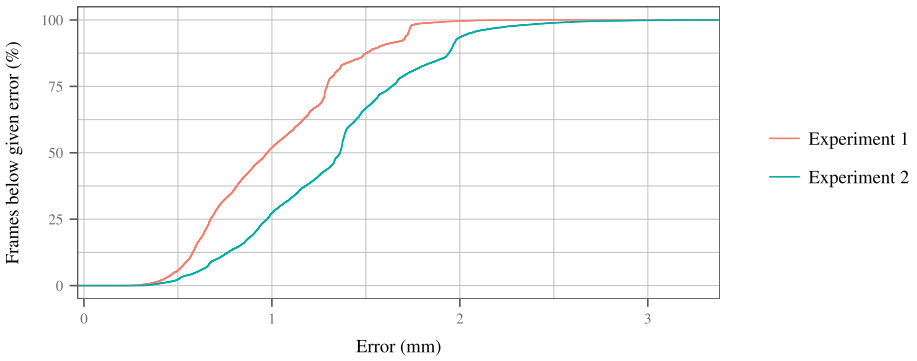


Fig. 14. Cumulative error for the two tracking experiments.

Table 2
Statistics of the weight distribution for the two tracking experiments.

	Experiment 1				Experiment 2			
Weight	Mean	Std. dev.	Min	Max	Mean	Std. dev.	Min	Max
s_1	−0.67	0.36	−1.59	1.35	−0.67	0	−0.67	−0.67
s_2	0.32	0.22	−0.59	0.93	0.32	0	0.32	0.32
s_3	0.26	0.44	−0.84	1.11	0.26	0	0.26	0.26
s_4	0.46	0.60	−1.58	1.49	0.46	0	0.46	0.46
s_5	0.36	0.34	−0.48	2.11	0.36	0	0.36	0.36
p_1	0.35	0.40	−1.41	1.20	0.44	0.37	−0.69	1.11
p_2	−0.82	0.42	−1.92	2.20	−0.68	0.54	−1.91	1.32
p_3	−0.62	0.24	−1.26	0.29	−0.75	0.46	−1.80	0.80
p_4	0.45	0.68	−1.69	1.96	0.70	1.11	−2.81	3.81

weights: we averaged the obtained anatomy weights of the first experiment. For the second experiment, we only optimized for **p** and fixed **s** to the estimated anatomy weights.

For all EMA frames of the results, we computed the distribution of the weights and also the cumulative error. The error in each frame was calculated by measuring the average Euclidean distance between vertex location and corresponding coil position. The error is shown in Fig. 14 and the weight distributions can be inspected in Fig. 15 and Table 2. Additionally, we created for each experiment a video for the "Northwind and the Sun" passage. These videos show the (anonymized) speaker during recording, an animation of the fitted tongue model together with the actual EMA coil positions, and information about the current weights. Both videos can be found in the supplementary material: the video file corresponding to the first experiment is named S9, the second, S8. Note that we show normalized versions of the model weights, i.e., they are shifted and scaled such that x represents the value $m_i + x\ \sigma_i$ where m_i and σ_i have the same roles as in (17).

9.3. Discussion

We observe in the results for the first experiment that we achieve acceptable errors where 62% of the errors are below 1.25 mm. Additionally, we see that all weights are used during the tracking approach, which also means that the anatomy is often also adapted to improve the fitting result. We see that their values approximately stay within the interval $[m_i - 2\ \sigma_i, m_i + 2\ \sigma_i]$. Moreover, we notice that weight 4 of the tongue pose is showing a significantly higher variance than the other pose weights.

Moving to the second experiment, we notice that the errors increased. However, they are still acceptable: 62% of the errors are still below 1.5 mm. This development can be seen as an expected behavior because the approach now only has 4 degrees of freedom instead of 9 to fit the data. Moreover, we see that the variance of the tongue pose weights increased. We might argue here that optimizing all weights as in the first experiment causes the pose weights to be underestimated. Again, the fourth weight shows a higher variance than the others. We suspect that this high

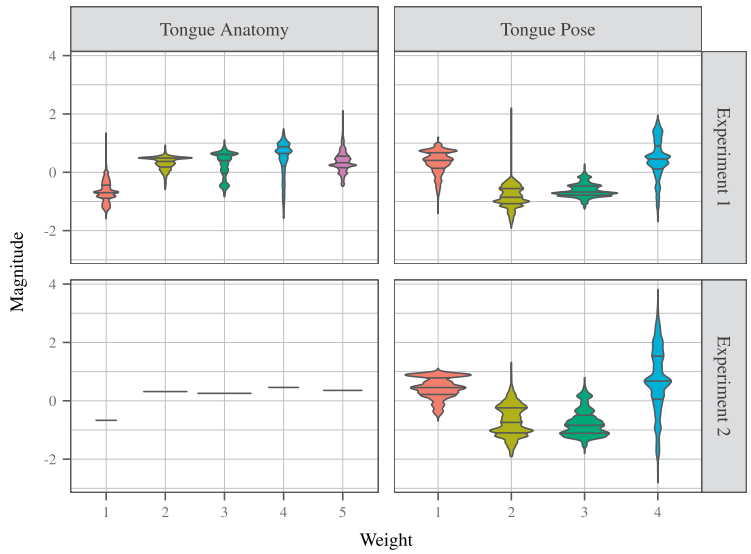


Fig. 15. Weight distribution for the two tracking experiments. The violin plots show the density, with the mean and interquartile range marked by horizontal lines.

variance and high range of achieved values might be caused by the presence of new phonemes or the fact that we are working with EMA data of German speakers.

By inspecting the video material, we notice that in the first experiment, the tongue is sometimes visually changing its anatomical properties: for example, it is shrinking or expanding. However, this is an expected behavior because the anatomy weights were also optimized to improve the result. By fixing the anatomy and optimizing only the tongue pose in the second experiment, we seem to avoid these problems: the anatomy of the tongue seems to stay stable and the transitions between frames also appear more plausible. But we also discover in the video material that our model might be lacking important shape variations: The motion of the tongue tip seems to be very rigid.

Overall, we conclude that the second approach produces more acceptable results than the first despite the slight loss of precision. Moreover, we see that this approach also provides us with information about transition paths in the tongue pose subspace between phonemes. These obtained transition paths could also be used to transfer the tracked motion to another speaker by adjusting the anatomy weights accordingly.

However, we observe that we cannot evaluate from these tracking experiments how close the estimated tongue shape is to the true tongue of the subject, because the true shape is unknown. For such an evaluation, a dedicated study is needed, which assesses the tracking and reconstruction capabilities of the model; this is beyond the scope of the present work. Such an evaluation might make use of an additional modality, such as US or rtMRI, to determine the true tongue contour.

10. Conclusion

In this work, we have presented a multilinear tongue model that was derived from volumetric MRI scans in a minimally supervised way. We also verified the validity of the extracted meshes by conducting a preference test. Afterwards, we observed in the experiments that a model with a low dimensionality can reliably register unknown data with an acceptable precision. Moreover, we explored in two experiments whether the model that was acquired from static data was suitable for tracking sparse motion capture data of the tongue. We found that fixing the anatomical features of the model to a speaker specific shape provided the most acceptable results. However, we also discovered indications that the current multilinear model might be missing some shape variations.

A previous version of our multilinear model was used in a text-to-speech framework that is able to synthesize audio with synchronized tongue motion (Le Maguer et al., 2017). For this purpose, the audio and EMA recordings of the mngu0 articulatory database (Richmond et al., 2011) were used, where we utilized the output tongue pose

weights from a tracking approach similar to the one described in Section 9 as feature vectors for training the synthesis framework.

In the future, we plan to investigate whether more shape variations can be obtained using more data. To this end, we want to use additional datasets in our framework. This implies that we also have to extract the shapes of phones like [g, k] that are characterized by contact with the soft palate. In this regard, we have to address the issue of recovering in the corresponding scans the surface of the soft palate, which can deform in a non-rigid way. Additionally, the datasets we use might differ with respect to the recorded phones, which leads to missing data in our training set. In this case, the simple averaging method for reconstructing missing shapes will no longer be sufficient. We believe that such a model may have access to a sufficient amount of shape variations to be usable in areas of application such as audiovisual speech synthesis, CAPT, or articulatory speech synthesis. Of course, this hypothesis has to be carefully investigated and evaluated in future work.

Moreover, in the current study, we manually selected the parameters for our framework. It may be interesting to investigate in the future if parameters exist that lead to acceptable results in a general setting, which would free the user from the burden of tuning them by hand. Moreover, we could try to improve the approach to extract higher quality meshes.

We are currently preparing a follow-up study to investigate the accuracy of the tongue shape reconstruction from EMA data, by using rtMRI as a reference. In this context, we think it may also be worthwhile to explore whether the derived model can be used to extract realistic 3D tongue motion from 2D rtMRI data recorded in the mid-sagittal plane. In contrast to EMA, this modality provides much richer motion information.

Finally, we could investigate if tongue motion transfer from one speaker to another is possible by adapting the anatomy weights to the new speaker and using the pose weights of the original speaker. It would also be interesting to compare different types of dynamic speech, e.g., whispering, shouting, or expressive speech. Ultimately, this could lead to a more flexible multilinear model that is able to synthesize these different types of speech.

Acknowledgments

This study was funded by the German Research Foundation (grant number EXC 284). It uses data from work supported by EPSRC Healthcare Partnerships (grant number EP/I027696/1).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.csl.2018.02.001

References

Allen, B., Curless, B., Popović, Z., 2003. The space of human body shapes: reconstruction and parameterization from range scans. ACM Trans. Graph. 22 (3), 587–594. doi: 10.1145/1201775.882311.

Ananthakrishnan, G., Badin, P., Vargas, J.A.V., Engwall, O., 2010. Predicting unseen articulations from multi-speaker articulatory models. Interspeech, pp. 1588–1591.

Badin, P., Bailly, G., Raybaudi, M., Segebarth, C., 1998. A three-dimensional linear articulatory model based on MRI data. 3rd ESCA/COCOSDA Workshop on Speech Synthesis (SSW).

Badin, P., Bailly, G., Reveret, L., Baciu, M., Segebarth, C., Savariaux, C., 2002. Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. J. Phonet. 30 (3), 533–553. doi: 10.1006/jpho.2002.0166.

Badin, P., Serrurier, A., 2006. Three-dimensional linear modeling of tongue: articulatory data and models. 7th International Seminar on Speech Production (ISSP), pp. 395–402.

Baer, T., Gore, J.C., Gracco, L., Nye, P.W., 1991. Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. J. Acoust. Soc. Am. 90 (2), 799–828. doi: 10.1121/1.401949.

Baker, A., 2011. A biomechanical tongue model for speech production based on MRI live speaker data. URL http://www.adambaker.org/qmu.php.

Beautemps, D., Badin, P., Bailly, G., 2001. Linear degrees of freedom in speech production: analysis of cineradio- and labio-film data and articulatory-acoustic modeling. J. Acoust. Soc. Am. 109 (5), 2165–2180. doi: 10.1121/1.1361090.

Bijar, A., Rohan, P.-Y., Perrier, P., Payan, Y., 2016. Atlas-based automatic generation of subject-specific finite element tongue meshes. Ann. Biomed. Eng. 44 (1), 16–34. doi: 10.1007/s10439-015-1497-y.

Blandin, R., Arnela, M., Laboissiére, R., Pelorson, X., Guasch, O., Hirtum, A.V., Laval, X., 2015. Effects of higher order propagation modes in vocal tract like geometries. J. Acoust. Soc. Am. 137 (2), 832–843. doi: 10.1121/1.4906166.

Blanz, V., Vetter, T., 1999. A morphable model for the synthesis of 3D faces. 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). ACM Press/Addison-Wesley Publishing Co., pp. 187–194. doi: 10.1145/311535.311556.

Bolkart, T., Wuhrer, S., 2015. 3D faces in motion: fully automatic registration and statistical analysis. Comput. Vision Image Understand. 131, 100–115. doi: 10.1016/j.cviu.2014.06.013.

Botsch, M., Kobbelt, L., Pauly, M., Alliez, P., Levy, B., 2010. Polygon Mesh Processing. A K Peters/CRC Press.

Brunner, J., Fuchs, S., Perrier, P., 2009. On the relationship between palate shape and articulatory behavior. J. Acoust. Soc. Am. 125 (6), 3936–3949. doi: 10.1121/1.3125313.

Buchaillard, S., Brix, M., Perrier, P., Payan, Y., 2007. Simulations of the consequences of tongue surgery on tongue mobility: implications for speech production in post-surgery conditions. Int. J. Med. Rob. Comput. Assist. Surgery 3 (3), 252–261. doi: 10.1002/rcs.142.

Buchaillard, S., Perrier, P., Payan, Y., 2009. A biomechanical model of cardinal vowel production: muscle activations and the impact of gravity on tongue positioning. J. Acoust. Soc. Am. 126 (4), 2033–2051. doi: 10.1121/1.3204306.

Burdumy, M., Traser, L., Richter, B., Echternach, M., Korvink, J.G., Hennig, J., Zaitsev, M., 2015. Acceleration of MRI of the vocal tract provides additional insight into articulator modifications. J. Magn. Reson. Imaging 42 (4), 925–935. doi: 10.1002/jmri.24857.

De Silva, V., Lim, L.-H., 2008. Tensor rank and the ill-posedness of the best low-rank approximation problem. SIAM J. Matrix Anal. Appl. 30 (3), 1084–1127. doi: 10.1137/06066518X.

Demolin, D., Metens, T., Soquet, A., 2000. Real time MRI and articulatory coordinations in vowels. 5th Speech Production Seminar (SSP), pp. 86–93.

Dryden, I.L., Mardia, K.V., 1998. Statistical Shape Analysis. Wiley.

Elie, B., Laprie, Y., Vuissoz, P.-A., Odille, F., 2016. High spatiotemporal cineMRI films using compressed sensing for acquiring articulatory data. 24th European Signal Processing Conference (EUSIPCO), pp. 1353–1357. doi: 10.1109/EUSIPCO.2016.7760469.

Engwall, O., 2000. A 3D tongue model based on MRI data. 6th International Conference on Spoken Language Processing (ICSLP), 3, pp. 901–904.

Engwall, O., 2008. Can audio-visual instructions help learners improve their articulation? - An ultrasound study of short term changes. Interspeech, pp. 2631–2634.

Engwall, O., Badin, P., 1999. Collecting and analysing two- and three-dimensional MRI data for Swedish. KTH Dept. Speech, Music Hearing Q. Prog. Status Rep. 40 (3-4).

Eryildirim, A., Berger, M.-O., 2011. A guided approach for automatic segmentation and modeling of the vocal tract in MRI images. 19th European Signal Processing Conference (EUSIPCO), pp. 61–65.

Fang, Q., Chen, Y., Wang, H., Wei, J., Wang, J., Wu, X., Li, A., 2016. An improved 3D geometric tongue model. Interspeech, pp. 1104–1107. doi: 10.21437/Interspeech.2016-901.

Foldvik, A.K., 1993. A time-evolving three-dimensional vocal tract model by means of magnetic resonance imaging (MRI). 3rd European Conference on Speech Communication and Technology (Eurospeech), pp. 557–559.

Fu, M., Zhao, B., Carignan, C., Shosted, R.K., Perry, J.L., Kuehn, D.P., Liang, Z.-P., Sutton, B.P., 2015. High-resolution dynamic speech imaging with joint low-rank and sparsity constraints. Magn. Reson. Med. 73 (5), 1820–1832. doi: 10.1002/mrm.25302.

Fuchs, S., Winkler, R., Perrier, P., 2008. Do speakers' vocal tract geometries shape their articulatory vowel space? 8th International Seminar on Speech Production (ISSP), pp. 333–336.

Geng, C., Mooshammer, C., 2009. How to stretch and shrink vowel systems: results from a vowel normalization procedure. J. Acoust. Soc. Am. 125 (5), 3278–3288. doi: 10.1121/1.3106130.

Harandi, N.M., Abugharbieh, R., Fels, S., 2014. 3D segmentation of the tongue in MRI: a minimally interactive model-based approach. Comput. Methods Biomech. Biomed. Eng.. doi: 10.1080/21681163.2013.864958.

Harandi, N.M., Woo, J., Stone, M., Abugharbieh, R., Fels, S., 2017. Variability in muscle activation of simple speech motions: A biomechanical modeling approach. J. Acoust. Soc. Am. 141 (4), 2579–2590. doi: 10.1121/1.4978420.

Harshman, R., Ladefoged, P., Goldstein, L., 1977. Factor analysis of tongue shapes. J. Acoust. Soc. Am. 62 (3), 693–707. doi: 10.1121/1.381581.

Harshman, R.A., 1970. Foundations of the PARAFAC Procedure: Models and Conditions for an "Explanatory" Multi-Modal Factor Analysis. 16, . UCLA Working Papers in Phonetics

Hewer, A., Steiner, I., Bolkart, T., Wuhrer, S., Richmond, K., 2015. A statistical shape space model of the palate surface trained on 3D MRI scans of the vocal tract. 18th International Congress of Phonetic Sciences (ICPhS).

Hewer, A., Steiner, I., Wuhrer, S., 2014. A hybrid approach to 3D tongue modeling from vocal tract MRI using unsupervised image segmentation and mesh deformation. Interspeech, pp. 418–421.

Honda, K., Maeda, S., Hashi, M., Dembowski, J.S., Westbury, J.R., 1996. Human palate and related structures: their articulatory consequences. 4th International Conference on Spoken Language Processing (ICSLP). IEEE, pp. 784–787.

Hoole, P., Wismüller, A., Leinsinger, G., Kroos, C., Geumann, A., Inoue, M., 2000. Analysis of tongue configuration in multi-speaker, multi-volume MRI data. 5th Seminar on Speech Production (SSP), pp. 157–160.

Hoole, P., Zierdt, A., Geng, C., 2003. Beyond 2D in articulatory data acquisition and analysis. 15th International Congress of Phonetic Sciences (ICPhS), pp. 265–268.

International Phonetic Association, 1999. Handbook of the International Phonetic Association. Cambridge University Press.

Jackson, P.J.B., Singampalli, V.D., 2009. Statistical identification of articulation constraints in the production of speech. Speech Commun. 51 (8), 695–710. doi: 10.1016/j.specom.2009.03.007.

Johnson, K., Ladefoged, P., Lindau, M., 1993. Individual differences in vowel production. J. Acoust. Soc. Am. 94 (2), 701–714. doi: 10.1121/1.406887.

Kaburagi, T., 2015. Morphological and acoustic analysis of the vocal tract using a multi-speaker volumetric MRI dataset. Interspeech, pp. 379–383.

Kiers, H.A.L., Krijnen, W.P., 1991. An efficient algorithm for PARAFAC of three-way data with large numbers of observation units. Psychometrika 56 (1), 147–152. doi: 10.1007/BF02294592.

Kim, Y.-C., Narayanan, S.S., Nayak, K.S., 2009. Accelerated three-dimensional upper airway MRI using compressed sensing. Magn. Reson. Med. 61 (6), 1434–1440. doi: 10.1002/mrm.21953.

Kröger, B.J., Winkler, R., Mooshammer, C., Pompino-Marschall, B., 2000. Estimation of vocal tract area function from magnetic resonance imaging: preliminary results. 5th Seminar on Speech Production (SSP), pp. 333–336.

Ladefoged, P., 1982. A Course in Phonetics. Second Harcourt Brace Jovanovich.

Ladefoged, P., Broadbent, D.E., 1957. Information conveyed by vowels. J. Acoust. Soc. Am. 29 (1), 98–104. doi: 10.1121/1.1908694.

Le Maguer, S., Steiner, I., Hewer, A., 2017. An HMM/DNN comparison for synchronized text-to-speech and tongue motion synthesis. Interspeech, pp. 239–243. doi: 10.21437/Interspeech.2017-936.

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

Li, H., Adams, B., Guibas, L.J., Pauly, M., 2009. Robust single-view geometry and motion reconstruction. ACM Trans. Graph. 28 (5), 175:1–175:10. doi: 10.1145/1618452.1618521.

Lingala, S.G., Toutios, A., Toger, J., Lim, Y., Zhu, Y., Kim, Y.-C., Vaz, C., Narayanan, S.S., Nayak, K.S., 2016. State-of-the-art MRI protocol for comprehensive assessment of vocal tract structure and function. Interspeech, pp. 475–479. doi: 10.21437/Interspeech.2016-559.

Lingala, S.G., Zhu, Y., Kim, Y.-C., Toutios, A., Narayanan, S., Nayak, K.S., 2017. A fast and flexible MRI system for the study of dynamic vocal tract shaping. Magn. Reson. Med. 77 (1), 112–125. doi: 10.1002/mrm.26090.

Liu, J., Musialski, P., Wonka, P., Ye, J., 2013. Tensor completion for estimating missing values in visual data. IEEE Trans. Pattern Anal. Mach. Intell. 35 (1), 208–220. doi: 10.1109/TPAMI.2012.39.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748. doi: 10.1038/264746a0.

Mermelstein, P., 1973. Articulatory model for the study of speech production. J. Acoust. Soc. Am. 53 (4), 1070–1082. doi: 10.1121/1.1913427.

Narayanan, S.S., Alwan, A.A., Haker, K., 1995. An articulatory study of fricative consonants using magnetic resonance imaging. J. Acoust. Soc. Am. 98 (3), 1325–1347. doi: 10.1121/1.413469.

Narayanan, S.S., Alwan, A.A., Haker, K., 1997. Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. part I. The laterals. J. Acoust. Soc. Am. 101 (2), 1064–1077. doi: 10.1121/1.418030.

Narayanan, S.S., Nayak, K., Lee, S., Sethy, A., Byrd, D., 2004. An approach to real-time magnetic resonance imaging for speech production. J. Acoust. Soc. Am. 115 (4), 1771–1776. doi: 10.1121/1.1652588.

Niebergall, A., Zhang, S., Kunay, E., Keydana, G., Job, M., Uecker, M., Frahm, J., 2013. Real-time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction. Magn. Reson. Med. 69 (2), 477–485. doi: 10.1002/mrm.24276.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst., Man, Cybern. 9 (1), 62–66. doi: 10.1109/TSMC.1979.4310076.

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

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

Richmond, K., Hoole, P., King, S., 2011. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. Interspeech, pp. 1505–1508.

Rodrigues, M.A.F., Gillies, D.F., Charters, P., 2001. A biomechanical model of the upper airways for simulating laryngoscopy. Comput. Methods Biomech. Biomed. Eng. 4 (2), 127–148. doi: 10.1080/10255840008908001.

Rosset, A., Spadola, L., Ratib, O., 2004. OsiriX: an open-source software for navigating in multidimensional DICOM images. J. Digit. Imaging 17 (3), 205–216. doi: 10.1007/s10278-004-1014-6.

Rudy, K., Yunusova, Y., 2013. The effect of anatomic factors on tongue position variability during consonants. J. Speech, Lang., Hearing Res. 56 (1), 137–149. doi: 10.1044/1092-4388(2012/11-0218).

Scott, A., Boubertakh, R., Birch, M., Miquel, M., 2012. Towards clinical assessment of velopharyngeal closure using MRI: evaluation of real-time MRI sequences at 1.5 and 3 T. Brit. J. Radiol. 85 (1019), e1083–e1092. doi: 10.1259/bjr/32938996.

Serrurier, A., Badin, P., Boë, L.-J., Lamalle, L., Neuschaefer-Rube, C., 2017. Inter-speaker variability: speaker normalisation and quantitative estimation of articulatory invariants in speech production for French. Interspeech, pp. 2272–2276. doi: 10.21437/Interspeech.2017-1126.

Shadle, C.H., Mohammad, M., Carter, J.N., Jackson, P.J.B., 1999. Multi-planar dynamic magnetic resonance imaging: new tools for speech research. 14th International Congress of Phonetic Sciences (ICPhS), pp. 623–626.

Steiner, I., Knopp, P., Musche, S., Schmiedel, A., Braun, A., Ouni, S., 2014. Investigating the effects of posture and noise on speech production. 10th International Seminar on Speech Production (ISSP), pp. 417–420.

Stone, M., Lele, S., 1992. Representing the tongue surface with curve fits. 2nd International Conference on Spoken Language Processing (ICSLP), pp. 875–878.

Stone, M., Lundberg, A., 1996. Three-dimensional tongue surface shapes of English consonants and vowels. J. Acoust. Soc. Am. 99 (6), 3728–3737. doi: 10.1121/1.414969.

Stone, M., Woo, J., Lee, J., Poole, T., Seagraves, A., Chung, M., Kim, E., Murano, E.Z., Prince, J.L., Blemker, S.S., 2016. Structure and variability in human tongue muscle anatomy. Comput. Methods Biomech. Biomed. Eng. 1–9. doi: 10.1080/21681163.2016.1162752.

Styner, M.A., Rajamani, K.T., Nolte, L.-P., Zsemlye, G., Székely, G., Taylor, C.J., Davies, R.H., 2003. Evaluation of 3D correspondence methods for model building. 18th International Conference on Information Processing in Medical Imaging (IPMI), pp. 63–75. doi: 10.1007/978-3-540-45087-0_6.

Tiede, M.K., Yehia, H., Vatikiotis-Bateson, E., 1996. A shape-based approach to vocal tract area function estimation. 1st ETRW on Speech Production Modeling, pp. 41–44.

Toutios, A., Narayanan, S.S., 2015. Factor analysis of vocal-tract outlines derived from real-time magnetic resonance imaging data. 18th International Congress of Phonetic Sciences (ICPhS).

Tucker, L.R., 1966. Some mathematical notes on three-mode factor analysis. Psychometrika 31 (3), 279–311. doi: 10.1007/BF02289464.

Ultrax: Real-time tongue tracking for speech therapy using ultrasound 2014.

Valdés Vargas, J.A., Badin, P., Ananthakrishnan, G., Lamalle, L., 2012. Articulatory speaker normalisation based on MRI-data using three-way linear decomposition methods. 29e Journées d'Études sur la Parole (JEP), pp. 529–536.

Valdés Vargas, J.A., Badin, P., Lamalle, L., 2012. Articulatory speaker normalisation based on MRI-data using three-way linear decomposition methods. Interspeech, pp. 2186–2189.

Weickert, J., 1998. Anisotropic Diffusion in Image Processing. Teubner.

Weirich, M., Fuchs, S., 2013. Palatal morphology can influence speaker-specific realizations of phonemic contrasts. J. Speech, Lang., Hearing Res. 56 (6), S1894–S1908. doi: 10.1044/1092-4388(2013/12-0217).

Weirich, M., Lancia, L., Brunner, J., 2013. Inter-speaker articulatory variability during vowel-consonant-vowel sequences in twins and unrelated speakers. J. Acoust. Soc. Am. 134 (5), 3766–3780. doi: 10.1121/1.4822480.

Woo, J., Lee, J., Murano, E.Z., Xing, F., Al-Talib, M., Stone, M., Prince, J.L., 2015. A high-resolution atlas and statistical model of the vocal tract from structural MRI. Comput. Methods Biomech. Biomed. Eng. 3 (1), 47–60. doi: 10.1080/21681163.2014.933679.

Woo, J., Xing, F., Lee, J., Stone, M., Prince, J.L., 2015. Construction of an unbiased spatio-temporal atlas of the tongue during speech. 24th International Conference on Information Processing in Medical Imaging (IPMI), pp. 723–732. doi: 10.1007/978-3-319-19992-4_57.

Wu, X., Dang, J., Stavness, I., 2014. Iterative method to estimate muscle activation with a physiological articulatory model. Acoust. Sci. Technol. 35 (4), 201–212. doi: 10.1250/ast.35.201.

Yunusova, Y., Rosenthal, J.S., Rudy, K., Baljko, M., Daskalogiannakis, J., 2012. Positional targets for lingual consonants defined using electromagnetic articulography. J. Acoust. Soc. Am. 132 (2), 1027–1038. doi: 10.1121/1.4733542.

Zheng, Y., Hasegawa-Johnson, M., Pizza, S., 2003. Analysis of the three-dimensional tongue shape using a three-index factor analysis model. J. Acoust. Soc. Am. 113 (1), 478–486. doi: 10.1121/1.1520538.

Alexander Hewer received his BSc in Computer Science in 2009, and his MSc degrees in Computer Science and Visual Computing in 2012 and 2013, respectively, from Saarland University. Currently, he is a PhD candidate in the *Saarbrücken Graduate School of Computer Science*, Germany.

Stefanie Wuhrer received her Diplom-Mathematik (FH) in 2005 from the University of Applied Sciences Stuttgart, Germany. She received her MSc and PhD in Computer Science in 2006 and 2009, respectively, from Carleton University, Canada. Since 2015 she is a research scientist with Morpheo team at INRIA Grenoble Rhône-Alpes, France.

Ingmar Steiner received his Masters degree in Phonetics from the University of Bonn in 2004, and his PhD in Phonetics from Saarland University in 2010. He currently leads an Independent Research Group in Saarland University's Cluster of Excellence *Multimodal Computing and Interaction*.

Korin Richmond received his MA (Hons) in Linguistics and Russian, his MSc in Cognitive Science and Natural Language Processing, and his PhD on inversion mapping in 1995, 1997, and 2002, respectively, from the University of Edinburgh, UK. He is currently a lecturer at the Centre for Speech Technology Research.