A Multilinear Tongue Model Derived from Speech Related MRI Data of the Human Vocal Tract

Alexander Hewer, Stefanie Wuhrer, Ingmar Steiner, Korin Richmond

A Multilinear Tongue Model Derived from Speech Related MRI Data of the Human Vocal Tract

Alexander Hewer[1,2,3], Stefanie Wuhrer[4], Ingmar Steiner[1,2], and Korin Richmond[5]

[1]*Cluster of Excellence "Multimodal Computing and Interaction", Saarland University, Saarbrücken, Germany*
[2]*German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany*
[3]*Saarbrücken Graduate School in Computer Science, Saarbrücken, Germany*
[4]*INRIA Grenoble Rhône-Alpes, France*
[5]*Centre for Speech Technology Research, University of Edinburgh, UK*

Abstract

We present a multilinear statistical model of the human tongue that captures anatomical and tongue pose related shape variations separately. The model was derived from 3D magnetic resonance imaging data of 11 speakers sustaining speech related vocal tract configurations. The extraction was performed by using a minimally supervised method that uses as basis an image segmentation approach and a template fitting technique. Furthermore, it uses 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 experiments concluded that limiting the degrees of freedom for the anatomical and speech related variations to 5 and 4 respectively produces a model that can reliably register unknown data while avoiding overfitting effects.

Keywords: tongue, vocal tract, MRI, statistical model, shape analysis.

1 Introduction

The tongue as one of the main human articulators plays an important role in speech production. In speech science, it is therefore of great interest to understand its shape and motion during human articulation. The results of such an analysis could be used to derive a tongue model that is able to replicate those shape changes by manipulating a few parameters. Ideally, these parameters should be separated into two sets: One set should control the appearance of the tongue that is related to the anatomy of the speaker. The other set adapts the shape to the sound that should be produced. Such a tongue model would also provide insight into how a change of anatomy affects the articulation for speech production.

Areas of application for a tongue model are, for example, virtual avatars for multimodal spoken interaction or computer-aided pronunciation training. In the latter case, the user can be provided with visual information on how to move the tongue to produce a specific sound (Engwall, 2008). Additionally, such a model might be employed in an articulatory speech synthesis framework to approximate the vocal tract area function. Finally, a tongue model can also simply be used as a prior to register new data that is possibly very sparse.

We notice that we need data about the tongue shape during speech production to derive such a model. 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. 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 observed speaker and in contrast to ultrasound or electromagnetic articulography (EMA), it is able to provide dense volumetric measurements of the vocal tract. Moreover, there is work on adapting the MRI method to the needs of speech research: Here, advances have been made to improve the measurement time (Kim et al., 2009) and quality of the acquired scans (Lingala et al., 2016).

The measured MRI data only contains raw image data and has to be further processed to extract the desired shape information. In our case, a suitable shape representation is given by a polygon mesh. Such a representation offers the advantage that it can be directly used in different fields. For example, in computer graphics such meshes are used to generate animations of complex objects (Botsch et al., 2010, Chapter 9) or to model objects of highly complex geometry and topology. Furthermore, polygon models have been used in speech processing to generate acoustical simulations (Blandin et al., 2015). More importantly, they have been used to perform a statistical analysis of a class of shapes, like for example human bodies (Allen et al., 2003), faces (Blanz and Vetter, 1999), or tongues (Badin and Serrurier, 2006).

Ideally, the extraction process should be at least minimally supervised, as doing it manually takes a lot of time and might require the expertise of an anatomical expert.

Afterwards, the collection of estimated meshes can be analyzed to derive a statistical tongue model. Such a model offers the advantage that for each generated tongue shape its probability can be measured. This is helpful in situations where the tongue model is used as a statistical prior.

In literature, a lot of research has focused on analyzing the tongue shape during speech production. The works by Engwall (2000), Badin and Serrurier (2006), and Fang et al. (2016) used 3D MRI data of a single speaker to analyze the speech related shape variations by using principal component analysis (PCA) or linear component analysis (LCA) in the case of Engwall. They annotated the contour of the tongue manually in the scan data and used a mesh as shape representation.

There also exists research that aims at analyzing the anatomical and speech related shape variations separately: Harshman et al. (1977) investigated these variations in 2D X-Ray data. We note that this image modality is nowadays no longer used for this purpose due to its known negative health effects. Analysis on 2D MRI was conducted by Hoole et al. (2000), Ananthakrishnan et al. (2010), Vargas et al. (2012b), and Vargas et al. (2012a). Finally, Zheng et al. (2003) performed this analysis on sparse sets of 65 points that were manually extracted from 3D MRI scans.

On the whole, we see that there are still some open issues: Previous work focused only on 2D data or sparse 3D data to analyze the anatomical and speech related variations. This sparse data representation might not be sufficient to capture the whole complex structure of the tongue. Initial work investigating these variations in 3D meshes obtained from MRI data of 9 speakers was presented with 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 used MRI data, which makes it infeasible for large collections of data. Here, work exists that aims at facilitating the 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).

In this paper, we present an extended version of our previous work (Hewer et al., 2014). Its contributions can be summarized as follows:

We propose a minimally supervised framework for extracting tongue meshes from 3D MRI data. It is minimally supervised in the sense that a user only has to annotate a few landmarks in the scan data, which significantly reduces the burden on the user compared to annotating the entire tongue surface. Originally, it combined an image segmentation technique and a template matching approach to achieve that goal. Here, we add an image denoising method to the framework in order to deal with possibly corrupt data. Moreover, we modify the template matching approach to better handle volumetric point clouds. Furthermore, we integrate a strategy for restoring missing tongue surface information that occurs due to contact between hard palate and tongue. This improvement increases the amount of tongue shape configurations we can register. Finally, the framework is augmented by making use of a bootstrapping strategy, which refines the quality of the obtained shape meshes.

All these modifications allowed us to register speech related tongue shapes of 11 speakers that we 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 examined the obtained model with respect to its compactness, generalization, and specificity properties. In the case of the specificity analysis, we investigated the surface parts of the tongue mesh that play an important role during human articulation. The results of our experiments motivated us to choose a model with 5 degrees of freedom for the anatomy and 4 for the speech related tongue pose.

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. The next section, section 4, is dedicated to describing how we estimate a tongue mesh from the surface information by using the template fitting. Here, 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. Furthermore, we conduct experiments to evaluate the compactness, generalization, and specificity properties of the acquired model. Finally, we conclude in section 7 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 are using a pure geometric representation 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 are treating 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 where we have the grid spacings h_x, h_y, and h_z. We say that $s(\mathbf{q})$ represents the measured hydrogen molecule density 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 hydrogen density values that scales them to a standard gray-scale. Here, we decided to use a standard visualization where bright values indicate a high density and dark values a low density.

Figure 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 data contains much more information than the vocal tract itself, 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

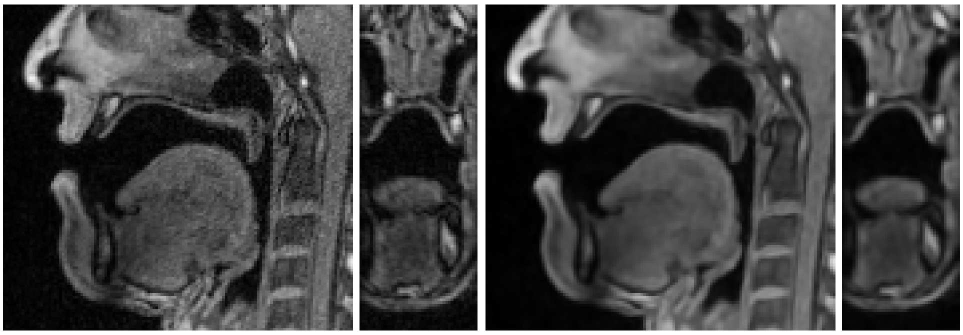

Figure 1: Raw MRI scan (**left**) and smoothed version (**right**). The left image of each pair shows a sagittal slice, the other one a coronal slice.

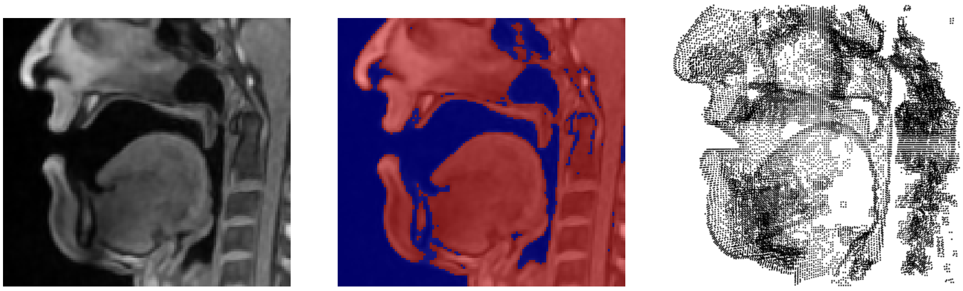

Figure 2: Extracting surface information from an MRI scan: Scan (**left**), segmentation (**center**), and resulting point cloud (**right**).

the data is degraded due to measurement noise. As a remedy, we apply a 3D variant of the edge-enhancing diffusion (Weickert, 1998) to the image. An example result of the approach can be inspected in Figure 1. 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 wanted 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 $\Omega_B = \Omega \setminus \Omega_O$ everything else. By inspecting the denoised data, we notice that tissue can be distinguished from non-tissue, such as air and bone for example, 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 by Otsu (1979) to perform this task as it is fully automatic. An example segmentation can be seen in Figure 2.

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 Figure 2.

3 Template Matching

Now, 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 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 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 wanted 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_{data} is minimized if applying A to the mesh moves it towards some points in the point cloud. The smoothness term E_{smooth} penalizes deformations that alter the original shape of the template. Finally, the landmark term E_{landmark} produces energy if correspondences between landmarks on the mesh and user-provided points are violated by the deformation.

As Equation 4 is not differentiable, it is usually optimized by minimizing a series of energies $E^t_{\text{Def}}(A^t)$ where $t \in [1,t_{\max}]$. We note that each energy uses adapted weights β^t and γ^t:

$$\beta^t = \beta - (t-1)\frac{\beta - \beta_{\min}}{t_{\max} - 1} \tag{5}$$

$$\gamma^t = \gamma - (t-1)\frac{\gamma - \gamma_{\min}}{t_{\max} - 1} \tag{6}$$

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

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.

We have modified this nearest neighbor heuristic somewhat: We 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

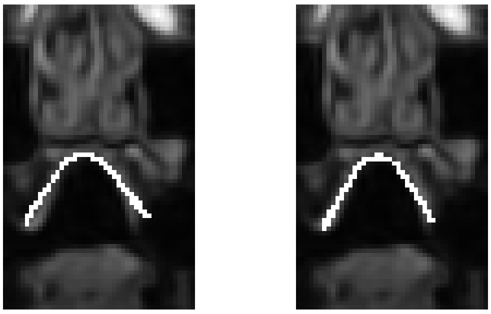

Figure 3: Comparison between nearest neighbor heuristics in the case of palate matching: **Left**: Standard heuristic causes template to get stuck at unrelated tissue. **Right**: Adapted heuristic moves template to the right position.

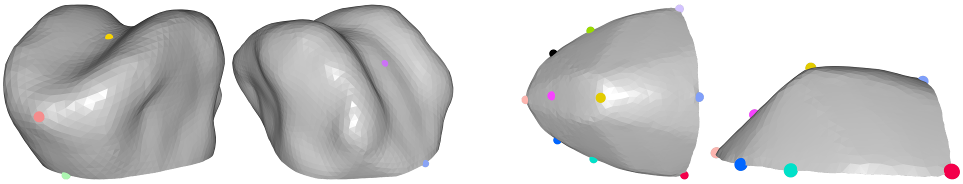

Figure 4: Used templates with landmarks of the tongue (**left**) and hard palate (**right**).

in the volumetric cloud during the optimization. An example showing the benefits of this modified heuristic can be inspected in Figure 3. Here, we note that we are showing the projection of the matched template on the scan data for the sake of visibility and interpretability.

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 a medical imaging software. Afterwards, we made the templates symmetric to remove this particular bias towards the original speaker.

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. We note that the tongue template is missing the sublingual part of the tongue that is negligible for speech production.

Both templates can be inspected in Figure 4 together with the landmarks used.

4 Tongue and Palate Shape Estimation

We first estimate the palate shape for each MRI scan. This shape information is needed in some cases to restore tongue surface information that is missing due to contacts between tongue and palate.

First, we select a scan for each speaker where the hard palate is clearly visible and perform template matching. We note that, in general, using a

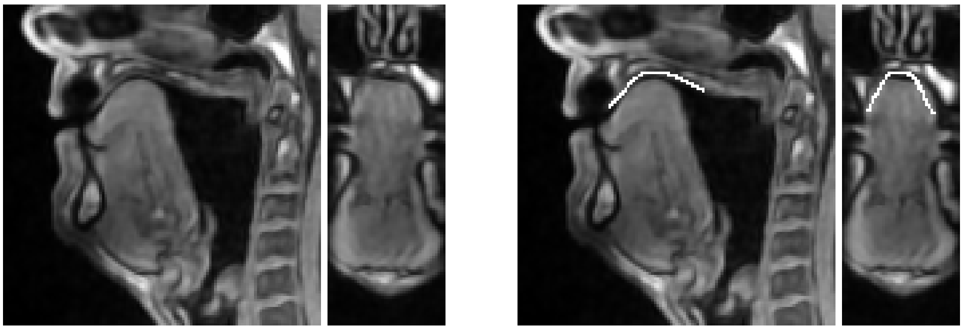

Figure 5: Palatal contact during the pronunciation of [i] (**left**) and result of restoring hard palate surface information (**right**).

single template might produce suboptimal 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. This model is then fitted to each point cloud and the results are afterwards used as the initialization for the template matching.

After we acquired the hard palate mesh for each considered speaker, we want to align this mesh to each scan of the corresponding speaker. This procedure serves the purpose of restoring tongue surface information that is missing due to contacts between tongue and palate as shown in Figure 5.

Here, 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 color information for this task. To this end, we define the color profile set $E(M,f) \subset \mathbb{R}^{\ell}$ of a mesh M in a scan f. A profile $\mathbf{e}^i(M,f) \in E(M,f)$ is a vector such that its entries are given by

$$\mathbf{e}^i_j(M,f) = f(\mathbf{v}_i + j\; d\; \mathbf{n}_i) \tag{7}$$

where $\mathbf{v}_i$ is a mesh vertex, $\mathbf{n}_i$ its corresponding normal, and d the chosen sampling distance. We see that we start above the palate surface in order to avoid taking samples in the possible contact area between tongue and palate.

Then, 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:

$$E_{\text{palate}}(A) = \sum_{i \in J(V)} \text{NCC}\left(\mathbf{e}^i(M,f), \mathbf{e}^i(A(M),g)\right) \tag{8}$$

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 NCC as similarity measure because it is known to be robust against noise and brightness differences. Furthermore, the NCC between color profiles was already successfully used in a nearest neighbor heuristic for template matching (Harandi et al., 2014). A result of this alignment approach can be seen in Figure 5.

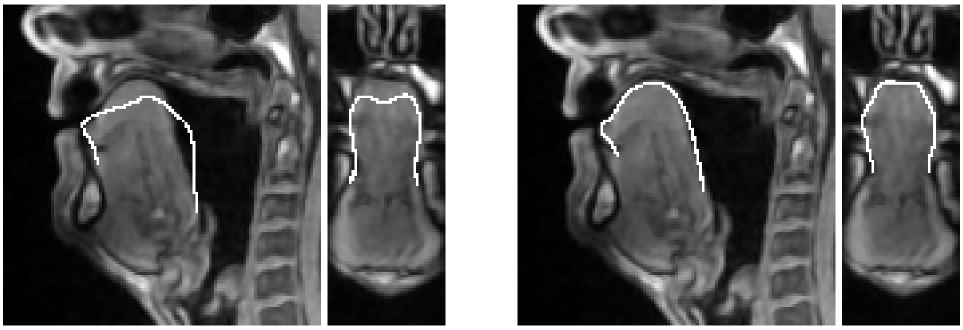

Figure 6: Effect of the bootstrapping strategy. **Left**: Initial template matching has trouble correctly registering the tongue shape. **Right**: Bootstrapping improves the result.

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 a 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 iteration as a statistical prior that is described in the next section. Effects of this bootstrapping operation can be seen in Figure 6.

5 Multilinear Tongue Model

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

$$M : S \times P \to \mathbb{M} \tag{9}$$

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 shape of the tongue. The set $P \subseteq \mathbb{R}^{\tilde{n}}$ contains coordinates $\mathbf{p}$ that determine the shape for a specific speech related tongue pose. Here, 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 Equation 9 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 destroy critical information, e.g., related to 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. Here, 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 motions or position differences without destroying any speech or anatomy specific information.

Finally, we have to ensure that for each speaker the meshes for all selected poses are available. Here, 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. We note that there exist more sophisticated methods to restore missing information, like for example HALRTC (Liu et al., 2013). In our case, however, this averaging approach was sufficient.

5.2 Deriving the Model

In order to derive our desired function in Equation 9, we need to analyze the anatomical and speech related variations separately. In several works (Harshman et al., 1977; Hoole et al., 2000, 2003; Ananthakrishnan et al., 2010; Vargas et al., 2012b,a; Zheng et al., 2003), the PARAFAC method (Harshman, 1970) was used to perform this analysis. This method, also known as CANDECOMP, 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, literature reports issues with this method: Hoole et al. (2000) found that it might be difficult to find reliable solutions. Vargas et al. (2012a) pointed out that the PARAFAC decomposition requires a lot of 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) that 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 (Vargas et al., 2012b).

Considering the issues of PARAFAC, we decided to use the Tucker decomposition to analyze our data. Here, 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 amount 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:

$$A = C \times_1 U_1 \times_2 U_2 \tag{10}$$

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 n-th 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 Equation 9: 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:

$$v(\mathbf{s}, \mathbf{p}) = \mu + C \times_1 \mathbf{s} \times_2 \mathbf{p} \tag{11}$$

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

$$M(\mathbf{s}, \mathbf{p}) := (V(\mathbf{s}, \mathbf{p}), F) \tag{12}$$

where F is the face set of our original template. We remark that the dimensionality of the speaker and pose subspace 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 $p \in \mathbf{P}$ that best describe the speaker anatomy and tongue pose that is represented in the data. To this end, we minimize the following energy:

$$E_{\text{Fit}}(\mathbf{s}, \mathbf{p}) = \alpha \, E_{\text{data}}(\mathbf{s}, \mathbf{p}) + \gamma \, E_{\text{landmark}}(\mathbf{s}, \mathbf{p}) \tag{13}$$

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, we note that if the correct neighbors are known, they can be set directly and only one energy has to minimized in that case.

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

$$[m_i - h\ \sigma_i, m_i + h\ \sigma_i] \tag{14}$$

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

We note that the above energy can also be used to fit a PCA model: In this case, the energy depends only on one parameter.

6 Evaluation

Our next goal is to apply the described framework to MRI data and evaluate the quality of the obtained tongue model.

6.1 Used Data

We use two datasets to derive our model: The dataset of Adam Baker (Baker, 2011) and the full dataset of the Ultrax project (Ult, 2014), which provides us with data of 12 speakers in total.

The Ultrax project consists of static MRI scans of 11 adult speakers 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 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 depict different articulatory configurations.

The data was recorded at the Clinical Research Imaging Centre in Edinburgh using a Siemens Verio 3T scanner where they 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. Here, we have as grid spacings $h_x = h_y = 1.1875\,\text{mm}$ and $h_z = 1.2\,\text{mm}$.

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, we note that 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. This means that we arrive at a tensor $A \in \mathbb{R}^{11\times 13\times 9300}$ where the dimension of the vertex mode is related to the vertex count of the tongue template we are using.

6.2 Applied Settings

For this data, the following settings were applied in our framework to extract the mesh collection:

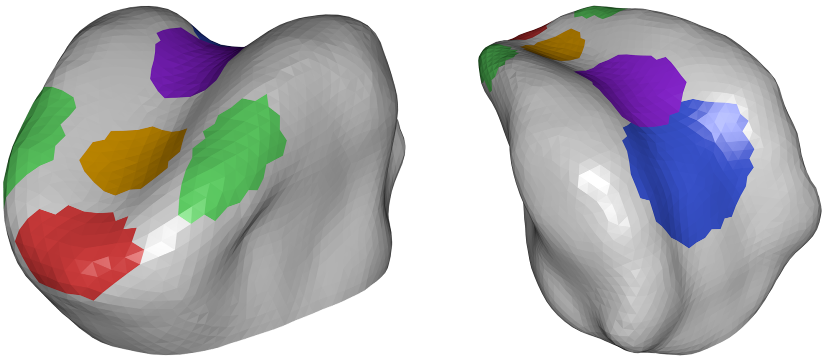

Figure 7: Speech related regions of the tongue surface: Tongue tip (red), tongue blade (brown), tongue back (violet), tongue dorsum (blue), and the lateral regions (green).

In the case of template matching, we used $\alpha = 1$, $\beta = 10$, $\beta_{\min} = 6$, 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 $\gamma_{\min} = 0$ to damp the effects of falsely placed landmarks. We used $\gamma_{\min} = 10$ 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 degrees. 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 $d = 1\,\text{mm}$.

In the bootstrapping strategy, we applied iterations until a satisfactory visual result was obtained: We used 1 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 that is not an anatomical expert.

6.3 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.

Compactness investigates how much the individual components of $\mathbf{s}$ and $\mathbf{p}$ contribute to the description of the used training data. In Figure 8, we see that using $\widetilde{m} = 5$ is sufficient to represent 91 percent of data variability. Approximately the same holds for $\widetilde{n} = 4$.

Generalization measures how well the model can represent data that was not part of the training. To evaluate the speaker generalization, we designed the following experiment: The data of each speaker was once excluded from

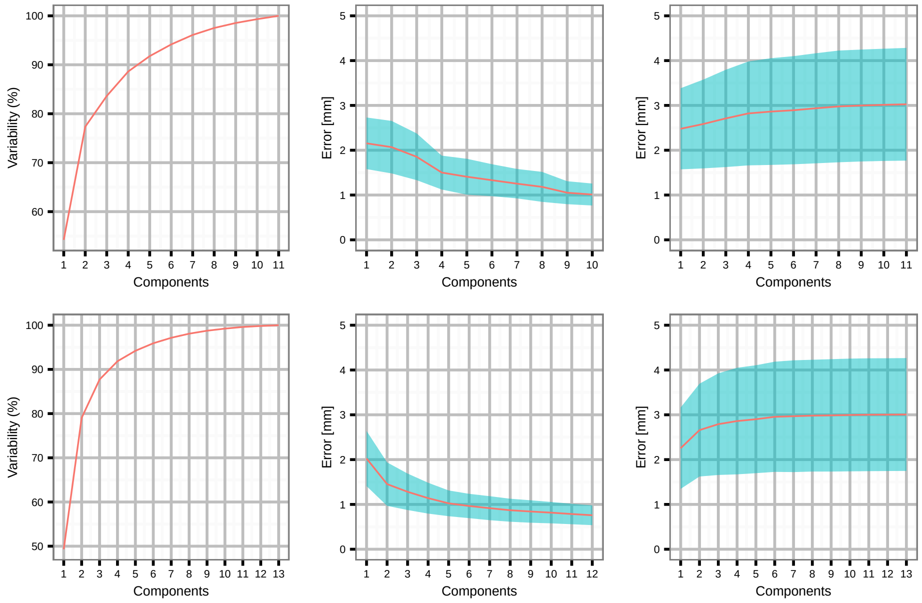

Figure 8: Compactness (**left**), generalization (**center**), and specificity (**right**) of the model for the speaker (**top**) and pose subspace (**bottom**). For the generalization and specificity, we visualize the mean (line) and the standard deviation (ribbon) of the experiments.

the training set. The derived model was then used to register this excluded data where we measured the average Euclidean distance between the registered mesh and the original one. 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 the same approach. In this case, the dimensionality of the speaker subspace was fixed to $\widetilde{m} = 5$. The results of these experiments are depicted in Figure 8. During this evaluation, we used the scale factor $h = 2$ in the model fitting optimization.

The specificity tries to assess how much the generated tongue shapes of the model differ from the original training data. In particular, we wanted to investigate how large these differences were for the regions of the tongue mesh that are speech related. Figure 7 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 tongue shape. The first experiment investigated the specificity of the speaker subspace. Here, 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 Figure 8.

Finally, we wanted to find out how much the tongue shapes belonging to specific phones differ from the corresponding ones generated by the model. Here, 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 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, we used in the distance evaluation and comparison parts of the tongue that are considered critical for this specific phone, *cf.* 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 the 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 Figure 9.

In all specificity experiments, we generated 10^6 samples.

The performed experiments provide an interesting insight into the model properties. The results of the generalization experiments show that only a few components of $\mathbf{p}$ and $\mathbf{s}$ are needed to reliably register unseen data. In particular, for $\mathbf{p}$, 3 components are enough to reach an average error that is slightly above the measurement precision of the MRI scan data. For $\mathbf{s}$, 7 components are needed to reach this kind of precision. Furthermore, we observe that a high number of components leads to errors below the measurement precision of the scan data, which can be considered as overfitting. Here, 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 phones [ʌ, ɔ], [e, i], or [e, ɛ] are similar

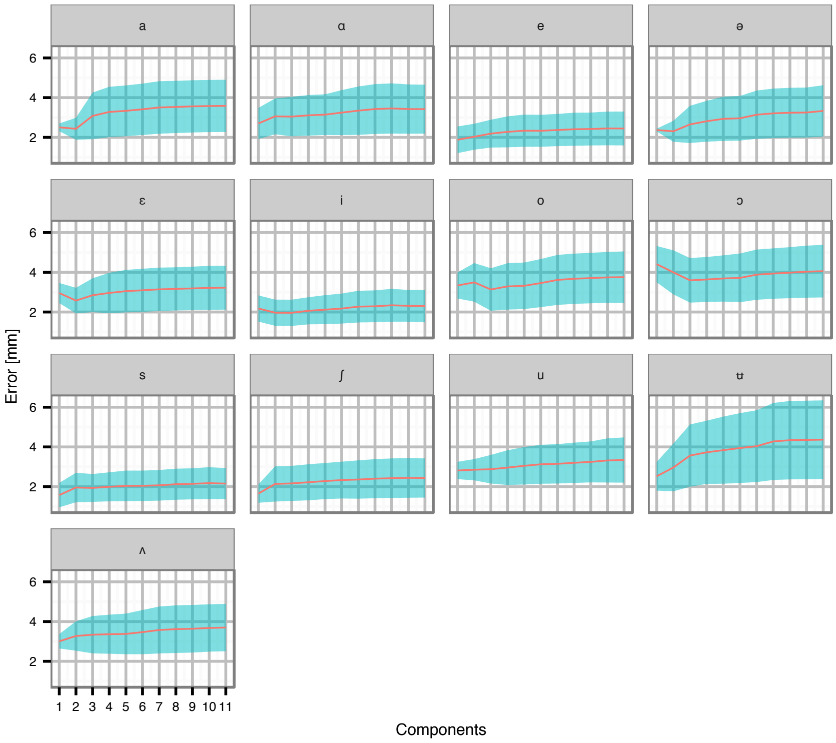

Figure 9: Specificity results for the individual fixed phone experiments. Plots show mean (line) and standard deviation (ribbon).

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.

Moreover, 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 that is then integrated into the model. From this observation, we draw the conclusion that the multilinear model might be used to detect such inconsistencies by performing the fixed phone experiments.

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.

7 Conclusion

In this work, we presented a multilinear tongue model that was derived from volumetric MRI scans in a minimally supervised way. In particular, we saw during the experiments that a model with a low dimensionality can reliably register unknown data with an acceptable precision.

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 [ɡ, k] that are known for having a contact with the soft palate. Here, we have to address the issue of recovering the surface of the soft palate in the corresponding scans that can also 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 is no longer sufficient. Furthermore, using more data also increases the risk of encountering falsely labeled or corrupt scans.

Our hypothesis that the multilinear model could be used to find inconsistencies in phone production could be tested in the future by choosing the phones in the training data as follows: One set should consist of phones that show little variance among speakers. The other set should contain phones that show a high variance among speakers because different strategies are available to produce them, for example [l, θ] (Keating, 2014). If the hypothesis were true, the fixed phone specificity experiments would show good results for the first set and bad results for the second one. In this case, this experiment could also be used as a heuristic to cluster speakers according to the articulation strategy they use. However, we note that the described experiment would require a dataset with recordings of the phones needed.

Moreover, we want to explore whether the derived model can be used to extract realistic 3D tongue motions from real-time 2D MRI data that was recorded in the mid-sagittal plane. We think that the acquired results can help to understand how the typical transitions between phones appear in the pose subspace and how they are affected by the anatomy of the speaker. Ultimately, this could lead to another multilinear model that could be used to generate these transitions.

Acknowledgments

This study uses data from work supported by EPSRC Healthcare Partnerships [grant number EP/I027696/1]. The work itself was funded by the German Research Foundation [grant number EXC 284].

References

Ultrax: Real-time tongue tracking for speech therapy using ultrasound, 2014. URL `http://www.ultrax-speech.org/`.

Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Transactions on Graphics*, 22(3):587–594, 2003. doi:10.1145/1201775.882311.

Gopal Ananthakrishnan, Pierre Badin, Julián Andrés Valdés Vargas, and Olov Engwall. Predicting unseen articulations from multi-speaker articulatory models. In *Interspeech*, pages 1588–1591, 2010. URL `http://www.isca-speech.org/archive/interspeech_2010/i10_1588.html`.

Pierre Badin and Antoine Serrurier. Three-dimensional linear modeling of tongue: Articulatory data and models. In *7th International Seminar on Speech Production (ISSP)*, pages 395–402, 2006.

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

Rémi Blandin, Marc Arnela, Rafael Laboissière, Xavier Pelorson, Oriol Guasch, Annemie Van Hirtum, and Xavier Laval. Effects of higher order propagation modes in vocal tract like geometries. *Journal of the Acoustical Society of America*, 137(2):832–843, February 2015. doi:10.1121/1.4906166.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

Timo Bolkart and Stefanie Wuhrer. 3D faces in motion: Fully automatic registration and statistical analysis. *Computer Vision and Image Understanding*, 131:100–115, 2015. doi:10.1016/j.cviu.2014.06.013.

Mario Botsch, Leif Kobbelt, Mark Pauly, Pierre Alliez, and Bruno Levy. *Polygon Mesh Processing*. A K Peters/CRC Press, 2010.

Vin De Silva and Lek-Heng Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127, 2008. doi:10.1137/06066518X.

Ian L Dryden and Kanti V Mardia. *Statistical Shape Analysis*. Wiley, 1998.

Olov Engwall. A 3D tongue model based on MRI data. In *6th International Conference on Spoken Language Processing (ICSLP)*, pages 901–904, 2000. URL `http://www.isca-speech.org/archive/icslp_2000/i00_3901.html`.

Olov Engwall. Can audio-visual instructions help learners improve their articulation? - an ultrasound study of short term changes. In *Interspeech*, pages 2631–2634, 2008. URL `http://www.isca-speech.org/archive/interspeech_2008/i08_2631.html`.

Abdulkadir Eryildirim and Marie-Odile Berger. A guided approach for automatic segmentation and modeling of the vocal tract in MRI images. In *19th European Signal Processing Conference (EUSIPCO)*, pages 61–65, 2011. URL `http://www.eurasip.org/Proceedings/Eusipco/Eusipco2011/papers/1569425007.pdf`.

Qiang Fang, Yun Chen, Haibo Wang, Jianguo Wei, Jianrong Wang, Xiyu Wu, and Aijun Li. An improved 3D geometric tongue model. *Interspeech*, pages 1104–1107, 2016. doi:10.21437/Interspeech.2016-901.

Negar M. Harandi, Rafeef Abugharbieh, and Sidney Fels. 3D segmentation of the tongue in MRI: a minimally interactive model-based approach. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 2014. doi:10.1080/21681163.2013.864958.

Richard Harshman, Peter Ladefoged, and Louis Goldstein. Factor analysis of tongue shapes. *Journal of the Acoustical Society of America*, 62(3): 693–707, 1977. doi:10.1121/1.381581.

Richard A Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16, 1970. URL `http://escholarship.org/uc/item/0410x385`.

Alexander Hewer, Ingmar Steiner, and Stefanie Wuhrer. A hybrid approach to 3D tongue modeling from vocal tract MRI using unsupervised image segmentation and mesh deformation. In *Interspeech*, pages 418–421, 2014. URL `http://www.isca-speech.org/archive/interspeech_2014/i14_0418.html`.

Alexander Hewer, Ingmar Steiner, Timo Bolkart, Stefanie Wuhrer, and Korin Richmond. A statistical shape space model of the palate surface trained on 3D MRI scans of the vocal tract. In *18th International Congress of Phonetic Sciences (ICPhS)*, 2015. URL `http://icphs2015.info/pdfs/Papers/ICPHS0724.pdf`.

Phil Hoole, Axel Wismüller, Gerda Leinsinger, Christian Kroos, Anja Geumann, and Michiko Inoue. Analysis of tongue configuration in multi-speaker, multi-volume MRI data. *5th Seminar on Speech Production*, pages 157–160, 2000.

Phil Hoole, Andreas Zierdt, and Christian Geng. Beyond 2D in articulatory data acquisition and analysis. In *15th Internation Congress of Phonetic Sciences (ICPhS)*, pages 265–268, 2003.

Philip JB Jackson and Veena D Singampalli. Statistical identification of articulation constraints in the production of speech. *Speech Communication*, 51(8):695–710, 2009. doi:10.1016/j.specom.2009.03.007.

Patricia A. Keating. Coronal places of articulation. In Carole Paradis and Jean-François Prunet, editors, *Phonetics and Phonology, Volume 2: The Special Status of Coronals*, chapter 2, pages 29–48. Academic Press, 2014.

Henk AL Kiers and Wim P Krijnen. An efficient algorithm for PARAFAC of three-way data with large numbers of observation units. *Psychometrika*, 56(1):147–152, 1991. doi:10.1007/BF02294592.

Yoon-Chul Kim, Shrikanth S. Narayanan, and Krishna S. Nayak. Accelerated 3D upper airway MRI using compressed sensing. *Magnetic Resonance in Medicine*, 61(6):1434–1440, 2009. doi:10.1002/mrm.21953.

Peter Ladefoged. *A Course in Phonetics*. Harcourt Brace Jovanovich, second edition, 1982.

Junghoon Lee, Jonghye Woo, Fangxu Xing, Emi Z. Murano, Maureen Stone, and Jerry L Prince. Semi-automatic segmentation of the tongue for 3D motion analysis with dynamic MRI. In *IEEE 10th International Symposium on Biomedical Imaging (ISBI)*, pages 1465–1468, 2013. doi:10.1109/ISBI.2013.6556811.

Hao Li, Bart Adams, Leonidas J. Guibas, and Mark Pauly. Robust single-view geometry and motion reconstruction. *ACM Transactions on Graphics*, 28(5):175:1–175:10, 2009. doi:10.1145/1618452.1618521.

Sajan G Lingala, Asterios Toutios, Johannes Toger, Yongwan Lim, Yinghua Zhu, Yoon-Chul Kim, Colin Vaz, Shrikanth S. Narayanan, and Krishna S. Nayak. State-of-the-art MRI protocol for comprehensive assessment of vocal tract structure and function. In *Interspeech*, 2016. doi:10.21437/Interspeech.2016-559.

Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013. doi:10.1109/TPAMI.2012.39.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi:10.1109/TSMC.1979.4310076.

Ting Peng, Erwan Kerrien, and Marie-Odile Berger. A shape-based framework to segmentation of tongue contours from MRI data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 662–665, 2010. doi:10.1109/ICASSP.2010.5495123.

Zeynab Raeesy, Sylvia Rueda, Jayaram K Udupa, and John Coleman. Automatic segmentation of vocal tract MR images. In *IEEE 10th International Symposium on Biomedical Imaging (ISBI)*, pages 1328–1331, 2013. doi:10.1109/ISBI.2013.6556777.

Martin A. Styner, Kumar T. Rajamani, Lutz-Peter Nolte, Gabriel Zsemlye, Gábor Székely, Christopher J. Taylor, and Rhodri H. Davies. Evaluation of 3D correspondence methods for model building. In *Information Processing in Medical Imaging*, pages 63–75. Springer, 2003. doi:10.1007/978-3-540-45087-0_6.

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

Julián Andrés Valdés Vargas, Pierre Badin, G Ananthakrishnan, and Laurent Lamalle. Normalisation articulatoire du locuteur par méthodes de décomposition tri-linéaire basées sur des données IRM. In *29e Journées d'Études sur la Parole (JEP)*, pages 529–536, 2012a. URL `http://www.aclweb.org/anthology/F12-1067`.

Julián Andrés Valdés Vargas, Pierre Badin, and Laurent Lamalle. Articulatory speaker normalisation based on MRI-data using three-way linear decomposition methods. In *Interspeech*, 2012b. URL `http://www.isca-speech.org/archive/interspeech_2012/i12_2186.html`.

Joachim Weickert. *Anisotropic Diffusion in Image Processing*. Teubner, 1998.

Yanli Zheng, Mark Hasegawa-Johnson, and Shamala Pizza. Analysis of the three-dimensional tongue shape using a three-index factor analysis model. *Journal of the Acoustical Society of America*, 113(1):478–486, 2003. doi:10.1121/1.1520538.