



---



# **Registration and Statistical Analysis of the Tongue Shape during Speech Production**

---



A dissertation submitted towards the degree of  
**Doctor of Engineering**  
of the Faculty of Mathematics and Computer Science  
of Saarland University

by

**Alexander Hewer**

Saarbrücken, 2019

**Date of Defense – Datum des Kolloquiums**

May 27, 2020

**Dean – Dekan**

Prof. Dr. Thomas Schuster

**Examination Board – Prüfungsausschuss**

Prof. Dr. Mark Groves  
(Chairman – Vorsitzender)

Dr. Ingmar Steiner  
(Thesis supervisor – Betreuer)

Prof. Dr. Joachim Weickert  
(Reviewer – Gutachter)

Dr. Dominik Schillo  
(Academic staff – Akademischer Mitarbeiter)



*To my parents*



# Abstract

This thesis analyzes the human tongue shape during speech production. First, a semi-supervised approach is derived for estimating the tongue shape from volumetric magnetic resonance imaging data of the human vocal tract. Results of this extraction are used to derive parametric tongue models. Next, a framework is presented for registering sparse motion capture data of the tongue by means of such a model. This method allows to generate full three-dimensional animations of the tongue. Finally, a multimodal and statistical text-to-speech system is developed that is able to synthesize audio and synchronized tongue motion from text.



# Zusammenfassung

Diese Dissertation beschäftigt sich mit der Analyse der menschlichen Zungenform während der Sprachproduktion. Zunächst wird ein semi-überwachtes Verfahren vorgestellt, mit dessen Hilfe sich Zungenformen von volumetrischen Magnetresonanztomographie-Aufnahmen des menschlichen Vokaltrakts schätzen lassen. Die Ergebnisse dieses Extraktionsverfahrens werden genutzt, um ein parametrisches Zungenmodell zu konstruieren. Danach wird eine Methode hergeleitet, die ein solches Modell nutzt, um spärliche Bewegungsaufnahmen der Zunge zu registrieren. Dieser Ansatz erlaubt es, dreidimensionale Animationen der Zunge zu erstellen. Zuletzt wird ein multimodales und statistisches Text-to-Speech-System entwickelt, das in der Lage ist, Audio und die dazu synchrone Zungenbewegung zu synthetisieren.



# Acknowledgments

Reading this dissertation sparks a lot of fond memories in me that are related to my time as a PhD student. Each section, each paragraph, and each sentence reminds me of a nice activity, a funny occasion, or great people who accompanied and supported me. Thus, to me, this thesis is more than just a summary of scientific findings: it is a diary, a collection of beautiful reminiscences of my time at Saarland university. Therefore, I wish to thank all the people who contributed to this great time.

I want to express my deepest gratitude to my family for all their support. My parents always encouraged and supported me – thank you for making all this possible. My brother Benedikt provided me with helpful advice and always believed in me. My brother Christian and his wife Sandra supported me and always had great barbecue ideas. My niece Ella and my nephew Niklas often created nice diversions for me. Our family dog Plato gave me the opportunity to go on long walks, which cleared my mind.

I am deeply indebted to my thesis advisor Ingmar Steiner. He gave me the opportunity to work on this exciting project and established a very nice and productive research environment, which also included game nights and other great social activities. I enjoyed the discussions with him and value his vast knowledge in various scientific fields. I am also immensely obliged to him for introducing me to many helpful technologies, which ultimately encouraged me to apply an agile and reproducible way of research. He also made sure to get me into contact with other researchers, which led to interesting collaborations.

I also want to thank Jessica for all her love and support. I am also grateful to her family for always making me feel welcome.

I also want to thank my friends for teaching me that there is more to life than just research. Patrick and Sarah Hilt always kept in touch and supported me. Dominik and Denise Reinert always had an open ear and provided helpful advice. Markus L’Hoste often created funny situations. Maximilian Erlacher, Dominik Schillo, and Josef Nguyen often watched anime together with me. Rami Ahmad, Nils Gutheil, and Stefan Schröder often invited me to play various video games with them. Daniel Kraemer joined me on various training sessions and always encouraged me to become better. Felix Leid, Laura Heine, Andreas Widenka, Pascal Kattler, Jens Horn, and Michael Jacobs also contributed to the great time I had – thank you.

I also want to thank a lot of people that I met at the department of computational linguistics. Nikolina Mitev, Christine Ankener, Mirjana Sekicki, and Maria Staudte were probably the best floor neighbors I can imagine. Bernd Möbius, Jürgen Trouvain, Jeanin Jügler, Iona Gessinger, and Iliana Simova contributed to the nice atmosphere at the department.

I also want to thank my colleagues from the Multimodal Speech Processing Group who

over the time became close friends. I am grateful to Eran Raveh for our great adventure – the road trip through Estonia and Latvia. I also enjoyed the various game nights we had. I want to thank Sébastien Le Maguer for introducing me to Emacs and various styles of metal music. I also had a lot of fun looking after his rabbits, which taught me a lot about vegetables. I want to express my gratitude to Arif Khan for all the tasty food he brought to the office. I also enjoyed the nice discussions we had. I am also grateful to all our student group members. Among others, Maureen Tanuadji, Kristy James, Tristan Hamilton, Paul Musset, Pradipta Deb, Atta-Ur-Rehman Shah, Moitree Basu, and Insa Kröger made our group a friendly community.

I also want to thank the numerous people I had the pleasure to meet and work with. Angelika Braun, Peter Knopp, Sebastian Musche, Fabian Tomaschek, and Konstantin Sering did a great job in recording EMA data for us. Shrikanth Narayanan and Asterios Toutios invited me for a research stay at the SPAN group of the University of Southern California. I enjoyed the discussions with them and I am thankful for the knowledge they shared. Together with Tanner Sorensen and Zisis Skordilis, they provided me also with access to the USC dataset. Fabrizio Nunnari and Alexis Heloir often had time for interesting discussions. Pierre Badin invited me to a stay at the GIPSA lab in Grenoble and shared his vast knowledge. Korin Richmond provided access to the Ultrax dataset and often collaborated with us on different projects. Stefanie Wuhrer and Timo Bolkart were always close collaborators. I am very grateful for the great discussions with them and their valuable feedback.

I also want to express my thanks to the people who helped me out on various organizational issues. The staff at DFKI took care of my travel planning and of my hardware equipment. The staff at the cluster office of MMCI always supported me when I needed help. The Dean’s office and the graduate school provided much needed help regarding the regulations of the PhD process. Christoph Clodo and his amazing systems group made working at the department a breeze.

I want to thank Joachim Weickert for teaching me a lot about image processing and for being part of my examination board. I also want to thank Mark Groves for training me in partial differential equations and for agreeing to be the chairman of my examination board.

I am also thankful to the staff of the Uni Fit for helping me to stay in shape.

I also want to express my appreciation towards all people whose articulatory data I could use during my research. Here, I also want to thank Adam Baker who allowed me to use his data.

Finally, I also want to express my gratitude towards all the reviewers and proofreaders for their constructive feedback.



# Contents

<b>List of Figures</b>	<b>xv</b>
<b>List of Tables</b>	<b>xxi</b>
<b>List of Acronyms</b>	<b>xxiii</b>
<b>1. Introduction</b>	<b>1</b>
1.1. Motivation . . . . .	1
1.2. Areas of application . . . . .	1
1.3. Magnetic resonance imaging . . . . .	2
1.4. Related work . . . . .	3
1.5. Contributions . . . . .	5
1.6. Thesis overview . . . . .	6
<b>2. Extracting articulator shape information from MRI data</b>	<b>7</b>
2.1. Introduction . . . . .	7
2.1.1. Motivation . . . . .	7
2.1.2. Related work . . . . .	7
2.1.3. Contribution . . . . .	8
2.1.4. Overview . . . . .	9
2.2. Data representation . . . . .	9
2.3. Datasets . . . . .	11
2.3.1. Ultrax dataset . . . . .	12
2.3.2. Baker dataset . . . . .	12
2.3.3. USC dataset . . . . .	12
2.3.4. Discussion . . . . .	13
2.4. Observations . . . . .	14
2.5. Dealing with noise . . . . .	16
2.5.1. Gaussian smoothing . . . . .	16
2.5.2. Median filtering . . . . .	16
2.5.3. Surface-enhancing diffusion filtering . . . . .	18
2.5.4. Discussion . . . . .	20
2.6. Detecting the spatial support . . . . .	20
2.6.1. Thresholding . . . . .	22
2.6.2. Otsu’s method . . . . .	22
2.6.3. Chan-Vese segmentation . . . . .	23
2.6.4. Graph cut . . . . .	23
2.6.5. Discussion . . . . .	27

2.7.	Estimating the shape . . . . .	27
2.7.1.	Surface point extraction . . . . .	27
2.7.2.	Observations . . . . .	28
2.7.3.	Visualizing meshes . . . . .	29
2.7.4.	Poisson reconstruction . . . . .	30
2.7.5.	Template matching . . . . .	31
2.8.	Experiments . . . . .	39
2.8.1.	Settings . . . . .	39
2.8.2.	Evaluation . . . . .	40
2.8.3.	Results . . . . .	40
2.9.	Conclusion . . . . .	46
<b>3.</b>	<b>Background on statistical shape analysis</b>	<b>47</b>
3.1.	Introduction . . . . .	47
3.2.	Vector representation of meshes . . . . .	47
3.3.	Linear modeling . . . . .	48
3.4.	Multilinear modeling . . . . .	49
3.4.1.	Tensor algebra . . . . .	50
3.4.2.	Tensor decompositions . . . . .	50
3.4.3.	Multilinear tongue model . . . . .	52
3.5.	Model fitting . . . . .	53
3.6.	Conclusion . . . . .	54
<b>4.</b>	<b>Deriving statistical tongue models from MRI datasets</b>	<b>55</b>
4.1.	Introduction . . . . .	55
4.1.1.	Motivation . . . . .	55
4.1.2.	Contribution . . . . .	55
4.1.3.	Overview . . . . .	56
4.2.	Palate reconstruction . . . . .	56
4.3.	Removing unrelated point cloud data . . . . .	59
4.4.	Pose normalization . . . . .	60
4.5.	Bootstrapping strategy . . . . .	62
4.6.	Reconstruction of missing shapes . . . . .	63
4.7.	Validation of the subjective evaluation . . . . .	65
4.8.	Registration experiments . . . . .	67
4.8.1.	Settings . . . . .	67
4.8.2.	Observed improvements . . . . .	67
4.8.3.	Results . . . . .	67
4.9.	Model creation and evaluation . . . . .	74
4.9.1.	Compactness . . . . .	74
4.9.2.	Generalization . . . . .	76
4.9.3.	Specificity . . . . .	77
4.9.4.	Fixed phone specificity . . . . .	78
4.9.5.	Comparison between Ultrax and USC model . . . . .	78

4.9.6. Final model . . . . .	81
4.10. Conclusion . . . . .	81
<b>5. Registering sparse motion capture data</b>	<b>83</b>
5.1. Introduction . . . . .	83
5.1.1. Motivation . . . . .	83
5.1.2. Related work . . . . .	86
5.1.3. Contribution . . . . .	87
5.1.4. Overview . . . . .	87
5.2. Data Representation . . . . .	88
5.3. Preprocessing . . . . .	88
5.3.1. Denoising . . . . .	88
5.3.2. Reconstruction of missing data . . . . .	89
5.3.3. Removing head motion . . . . .	89
5.3.4. Mapping the data into a canonical coordinate system . . . . .	89
5.3.5. Rotating the data . . . . .	91
5.3.6. Mapping to origin of model . . . . .	91
5.3.7. Simplification of the data . . . . .	92
5.3.8. Comparison to related work . . . . .	92
5.4. Registering EMA data . . . . .	93
5.5. Finding correspondences between EMA coils and model vertices . . . . .	94
5.5.1. Randomized approach . . . . .	95
5.5.2. Combinatorial approach . . . . .	95
5.6. Estimating the speaker anatomy . . . . .	96
5.7. Experiments . . . . .	97
5.7.1. Settings . . . . .	97
5.7.2. The EMA subset of mngu0 . . . . .	98
5.7.3. Tübingen dataset . . . . .	113
5.7.4. Trier dataset . . . . .	119
5.8. Conclusion . . . . .	122
<b>6. Multimodal speech synthesis</b>	<b>125</b>
6.1. Introduction . . . . .	125
6.1.1. Motivation . . . . .	125
6.1.2. Related work . . . . .	126
6.1.3. Contribution . . . . .	126
6.1.4. Overview . . . . .	127
6.2. Adapting the HTS framework . . . . .	127
6.2.1. Standard approach . . . . .	127
6.2.2. Multimodal extension . . . . .	128
6.2.3. Separating the articulatory model from the acoustical model . . . . .	128
6.3. Experiments . . . . .	130
6.3.1. Overview . . . . .	130
6.3.2. Setup . . . . .	130

## Contents

6.3.3. Database . . . . .	130
6.3.4. Used tongue model . . . . .	131
6.3.5. EMA registration approach . . . . .	133
6.3.6. Acoustic Synthesis . . . . .	133
6.3.7. Combined Acoustic and EMA Synthesis . . . . .	135
6.3.8. EMA Synthesis . . . . .	137
6.3.9. Tongue-only EMA Synthesis . . . . .	138
6.3.10. Tongue model based tongue motion synthesis . . . . .	140
6.3.11. Comparison of the different systems . . . . .	142
6.3.12. Experiment using new model and registration . . . . .	142
6.4. Conclusion . . . . .	143
<b>7. Conclusion</b>	<b>145</b>
7.1. Summary . . . . .	145
7.2. Future work . . . . .	145
7.3. Source code . . . . .	146
<b>Appendices</b>	<b>147</b>
<b>A. IPA Chart</b>	<b>149</b>
<b>Bibliography</b>	<b>151</b>

# List of Figures

2.1. Simplified illustration showing orientation of the coordinate system used for the MRI scans. Figure adapted from Richmond, Hoole, et al. (2011). It is important to note that the origin of the shown coordinate system was chosen arbitrarily and in general does not correspond to the true origin of the considered MRI scans. . . . .	9
2.2. Different slice views of the same MRI scan: sagittal (left), coronal (center), and transverse slice (right). Colored lines indicate where the individual slices are located within the different visualizations. . . . .	11
2.3. Sagittal views of scans from the three datasets. Rows show example results of the shape extraction process. . . . .	15
2.4. Original version of scan and results for different smoothing filters. Sagittal and coronal views are shown. The individual filters used the following settings. Median filtering: $r = 2$ . Gaussian convolution: $\sigma = 2$ . Diffusion: $\sigma = 1$ , $\rho = 1$ , $\lambda = 0.1$ , evolution time $t = 2.4$ . . . . .	17
2.5. Segmentation results for supervised and unsupervised approaches based on thresholding. Used parameters for manual thresholding: 15 (Ultrax) and 38 (USC). . . . .	21
2.6. Results of the Chan-Vese approach for example scans. White circles in the sagittal slices illustrate the zero set of the initial level set. Used parameters: $\lambda_O = 1$ , $\lambda_B = 3$ , $\mu = 500$ , $\eta = 0$ . . . . .	24
2.7. Example run of the graph cut approach for a $2 \times 2$ image. . . . .	25
2.8. Example annotations and corresponding graph cut results for 2D images of MRI slices. . . . .	26
2.9. Example point clouds obtained from segmentations. The clouds were clipped and decimated in order to improve visibility. . . . .	28
2.10. Standard visualization of meshes and projection onto corresponding scans. . . . .	29
2.11. Results of the Poisson reconstruction for given point clouds. Meshes and their projection onto the corresponding scans are shown. Again, the point clouds are modified to improve visibility. The reconstruction itself used the unmodified point clouds. . . . .	30
2.12. Used templates with landmarks of the tongue (left) and hard palate (right). . . . .	31
2.13. Impact of using result of rigid alignment as initialization for template matching. . . . .	34
2.14. Comparison between the direct optimization of the rigid alignment energy and the two-step approach. Meshes and their projection onto the corresponding scans are shown. . . . .	36

List of Figures

2.15. 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).	36
2.16. 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).	37
2.17. Smoothing a mesh in order to remove high frequency noise.	38
2.18. Example results for the Baker dataset where the approach provided acceptable registrations.	41
2.19. Example results for the phones [ə] (top row), [ɔ:] (center row), and [ɑ:] (bottom row) in the USC dataset. Registrations are shown for the speakers F1, F7, and M5 to illustrate different articulation strategies for the same phone.	42
2.20. Example results for the phones [ɑ] (top row), [ʌ] (center row), and [ɔ] (bottom row) in the Ultrax dataset. Registrations are shown for the speakers 04MRIF, 08MRIM, and 14MRIF.	43
2.21. Examples scans where the proposed approach fails to register the tongue surface properly. Captions provide information about dataset, speaker name, and produced phone.	44
2.22. A palatal contact causes the tongue surface to become indistinguishable from surrounding tissue in the corresponding area, which causes the approach to fail. Without contact, the approach produces an acceptable registration. Both scans belong to the same speaker. Figures show scan with and without projection of result mesh.	45
4.1. Results for injecting the palate shape estimated from a source scan into a target scan. In the shown case, direct injection causes the information to be located at the wrong position. Aligning the palate by using the estimated head motion between both scans leads to a better result.	57
4.2. Sagittal (left) and coronal slice (right) illustrating an example region for estimating the head motion by means of the approach of Lucas and Kanade.	58
4.3. Proposed palate reconstruction strategy may help to avoid annotation pitfalls. While the palate surface is clearly visible on the left scan, it is hard to detect on the center scan. The palate reconstruction helps to fill in the missing information.	59
4.4. Restoring the missing palate boundary information improves the tongue matching result to some degree.	59
4.5. Removing points above the hard palate improves the registration result.	60

4.6.	Coordinate system used for the pose normalization. The image shows the mid-sagittal plane. The origin of the system is located near the front teeth, highlighted by a red dot; the horizontal and vertical axes represent the $y$ and $z$ dimensions, respectively, while the $x$ axis is perpendicular to the image plane. Observe that compared to Figure 2.1, the roles of the axes changed. . . . .	61
4.7.	Effect of bootstrapping. . . . .	62
4.8.	Results of the preference test for each considered scan. Note that not all phonemes are available in the data for speaker 01MRIM. The scans were grouped by speaker to improve the visualization. . . . .	66
4.9.	Examples for scans where the participants preferred the initial template matching result (top row) over the bootstrapping one (bottom row). . . .	66
4.10.	Comparison between original approach and extended version. The registration quality significantly improves in the case of the extended approach. Dataset, speaker name, and produced phone are provided for reference. . .	69
4.11.	Example results of the extended approach for selected phones in the Baker dataset. . . . .	70
4.12.	Example results for the phones [i] (top row), [s] (center row), and [ʃ] (bottom row) in the Ultrax dataset. Registrations are shown for the speakers 03MRIF, 09MRIM, and 11MRIM. . . . .	71
4.13.	Example results for the phones [ɹ] (top row), [ɹ̥] (center row), and [θ] (bottom row) in the USC dataset. Registrations are shown for the speakers F5, M1, and M2. . . . .	72
4.14.	Missing information in the scan of speaker 11MRIM for phone [i]. The estimated tongue shape (shown as projection) may be seen as reconstruction of information in this case. For visualization purposes, a coronal (left) and a transverse slice (right) are provided. . . . .	73
4.15.	Compactness (left), generalization (center), and specificity (right) of the evaluated models. For the generalization and specificity, the mean (line) and the standard deviation (ribbon) are shown. The plots provide the results for the Baker (top row), Ultrax (center row), and USC dataset (bottom row). . . . .	75
4.16.	Speech related regions of the tongue surface: Tongue tip (red), tongue blade (brown), tongue back (violet), tongue dorsum (blue), and the lateral regions (green). . . . .	77
4.17.	Specificity results for the fixed phone experiments of the Ultrax (top) and USC dataset (bottom). Plots show mean (line) and standard deviation (ribbon). . . . .	79
4.18.	Comparison between Ultrax (top) and USC model (bottom). For the generalization and specificity, the mean (line) and the standard deviation (ribbon) of the experiments are shown. . . . .	80
5.1.	Examples for multimodal recording setups involving EMA. In both cases, the device including the transmitter coils is above the head of the subject. . . . .	85

List of Figures

5.2.	EMA sensor coils are glued to points of interest on the tongue. . . . .	88
5.3.	Example setup for recording the bite plane. EMA coils are placed on a triangle ruler (top). Subject puts ruler into mouth and bites on it (bottom). Images taken from Musche (2014). . . . .	90
5.4.	Sagittal view of an example visualization of an estimated palate trace (left). For reference, a sagittal slice of an MRI scan of the Baker dataset (right) is shown to illustrate the shape of a palate. . . . .	92
5.5.	Visualization of vertex subsets that are used in the combinatorial correspondence optimization approach: tongue tip (red), tongue blade (brown), tongue body (green), tongue back (violet), and tongue sagittal dorsum (blue). Due to overlapping vertex sets, multiple renderings of the tongue mesh are shown. . . . .	98
5.6.	Histogram of phone occurrence in the mngu0 dataset. The phone [ɔ̃] is omitted because it only occurs once. . . . .	99
5.7.	Illustration of tongue coil layout for the mngu0 dataset (left, adapted from Richmond, Hoole, et al., 2011) and rendering of combinatorial correspondence optimization result (right). Spheres on tongue mesh highlight the vertices that were found. . . . .	100
5.8.	Cumulative error for the two registrations of the mngu0 data. . . . .	101
5.9.	Tongue model parameter distribution for the two registrations of the mngu0 data. The violin plots show the density, with the mean and interquartile range marked by horizontal lines. . . . .	101
5.10.	Visual comparison between results obtained by full (top) and fixed speaker optimization (bottom). Spheres indicate EMA coil positions. The full optimization is continuously adapting the anatomy over time: the tongue shrinks or grows. The anatomical features remain consistent in the fixed speaker case. Registrations were obtained from utterance <i>s1_0002</i> . . . . .	103
5.11.	Trajectories of EMA coils and corresponding vertices of registered tongue model for utterance <i>s1_0711</i> . The top row shows the Euclidean distance between coil and corresponding vertex over time. The registration was obtained by optimizing only the tongue pose parameters and fixing the anatomy. . . . .	104
5.12.	Trajectories of EMA coils and corresponding vertices of registered tongue model for utterance <i>s1_0711</i> . The top row shows the Euclidean distance between coil and corresponding vertex over time. The registration was obtained by optimizing all parameters. . . . .	105
5.13.	Trajectories of tongue pose parameters over time for utterance <i>s1_0711</i> in the fixed speaker case. . . . .	106
5.14.	Trajectories of all tongue model parameters over time for utterance <i>s1_0711</i> in the full optimization case. . . . .	107
5.15.	Reconstructed tongue meshes from the obtained model parameters at time stamp 4.26 s for utterance <i>s1_0711</i> . Results for the full (left) and fixed speaker (right) registration are shown. Spheres representing the EMA coil positions are added for reference. . . . .	108



5.16. Cumulative error for the two registrations for the mngu0 data without silence intervals. . . . .	108
5.17. Tongue model parameter distribution for the two registrations for the mngu0 data without silence intervals. The violin plots show the density, with the mean and interquartile range marked by horizontal lines. . . . .	109
5.18. Trajectories of EMA coils and corresponding vertices of registered tongue model for diphone segment [l_k] in utterance <i>s1_1233</i> . The top row shows the Euclidean distance between coil and corresponding vertex over time. The registration was obtained by optimizing only the tongue pose parameters and fixing the anatomy. . . . .	110
5.19. Trajectories of tongue model parameters over time for diphone segment [l_k] in utterance <i>s1_1233</i> . The registration was obtained by optimizing only the tongue pose parameters and fixing the anatomy. . . . .	111
5.20. Reconstructed tongue meshes from the obtained model parameters at time stamp 5.745 s for utterance <i>s1_1233</i> . Results for the full (left) and fixed speaker (right) registration are shown. Spheres representing the EMA coil positions are added for reference. . . . .	111
5.21. Analysis of the tongue pose trajectories for the diphone [l_k]. Plots show the mean trajectory (line) and the standard deviation (ribbon). . . . .	112
5.22. Analysis of the tongue pose trajectories for the diphones [ð_ə] (top), [t_ə] (center), and [ə_n] (bottom). Plots show the mean trajectory (line) and the standard deviation (ribbon). . . . .	114
5.23. Photograph of tongue coil layout for subject sp02 of the Tübingen dataset is shown on the left. Combinatorial correspondence optimization result for mid-sagittal coils is provided for subjects sp02 (center) and sp03 (right). . . . .	115
5.24. Cumulative error for the two registrations of sp02 in the Tübingen dataset. . . . .	115
5.25. Tongue model parameter distribution for the two registrations of sp02 in the Tübingen dataset. The violin plots show the density, with the mean and interquartile range marked by horizontal lines. . . . .	116
5.26. Cumulative error for the two registrations of sp03 in the Tübingen dataset. . . . .	117
5.27. Tongue model parameter distribution for the two registrations of sp03 in the Tübingen dataset. The violin plots show the density, with the mean and interquartile range marked by horizontal lines. . . . .	117
5.28. Example frames of an animation showing subject sp02 created by fusing multimodal registration results of face and tongue. Frames belong to prompt 27 of the Tübingen dataset: “Did dad do academic bidding?” . . . . .	119
5.29. Photograph of tongue coil layout for subject VP08 of the Trier dataset (left) and rendering of combinatorial correspondence optimization result of mid-sagittal coils for subject VP05 (right). Spheres on tongue mesh highlight the vertices that were found. . . . .	120
5.30. Cumulative error for the two registrations of VP05 in the Trier dataset. . . . .	121
5.31. Tongue model parameter distribution for the two registrations of VP05 in the Trier dataset. The violin plots show the density, with the mean and interquartile range marked by horizontal lines. . . . .	122

List of Figures

6.1.	Diagrams of the different architectures used in the experiments. . . . .	129
6.2.	Full EMA coil layout of the used data of the mngu0 corpus. All coils are close to the mid-sagittal plane. The ref coil on the upper incisors forms the origin of the coordinate space. . . . .	131
6.3.	Distribution of phones across the training and test sets. The frequency of the silence intervals, denoted by <i>pau</i> , is also shown. . . . .	132
6.4.	Observed and predicted position trajectories (along the $x$ , $y$ , and $z$ axis), and Euclidean distance (top), for the tongue EMA coils (T1, T2, T3) for one test utterance, using combined acoustic and EMA synthesis with the HMM setup. The utterance <i>s1_0016</i> is shown: “Because these deer are gregarious, they go about in groups”. Based on the provided transcriptions, intervals containing silent (pause) and coronal and dorsal consonants have been highlighted. . . . .	136
6.5.	One test utterance produced using EMA-only synthesis; all other details are the same as in Figure 6.4. . . . .	138
6.6.	One test utterance produced using EMA-only synthesis restricted to the tongue coils; all other details are the same as in Figure 6.4. . . . .	139
6.7.	One test utterance produced using the tongue model parameters synthesis; all other details are the same as in Figure 6.4. . . . .	140
6.8.	Distributions of Euclidean distances between observed and predicted tongue EMA coil positions for each experimental TTS setup, split by phone class and tongue EMA coil. Results for the HMM (top) and DNN setup (bottom) are shown. . . . .	141

# List of Tables

1.1. Overview of several studies that have investigated shape variabilities of the vocal tract. The table lists the modality (or modalities) used, the analyzed data representation, and the number of subjects taking part in the corresponding study. . . . .	4
2.1. Used settings for the experiments. Parameter name, description, and value are provided. . . . .	39
4.1. Used settings for the experiments. Parameter name, description, and value are provided. . . . .	68
4.2. Summary of considered scans. Notes provide information about missing or ignored data in the datasets. . . . .	73
5.1. Comparison between different articulographs. The RMS values are provided by the manufacturers where no information is available on the AG500. . . . .	84
5.2. Selection of literature that applies EMA for analysis. . . . .	84
5.3. Selection of databases that provide EMA recordings. The individual entries provide information about recorded subjects, the used articulograph, and the used sample rate for the EMA data. Notes on the MOCHA TIMIT dataset: this dataset actually contains more speakers. However, in literature, only two are used. The quality of the other recordings of this dataset is unknown. . . . .	85
5.4. Used settings for the EMA registration experiments. Parameter name, description, and value are provided. . . . .	97
5.5. Statistics of tongue parameters and error for the two registrations of the mngu0 data. . . . .	102
5.6. Statistics of tongue parameters and error for the two registrations of the mngu0 data without silence intervals. . . . .	109
5.7. Statistics of tongue parameters and error for the two registrations of sp02 in the Tübingen dataset. . . . .	116
5.8. Statistics of tongue parameters and error for the two registrations of sp03 in the Tübingen dataset. . . . .	118
5.9. Statistics of tongue parameters and error for the two registrations of speaker VP05 in the Trier dataset. . . . .	121
6.1. EMA coil labels and locations in the used data of the mngu0 corpus. . . .	133

*List of Tables*

6.2.	Global evaluation measures for the acoustic synthesis baseline conditions. Results for the HMM and DNN setup are provided. . . . .	135
6.3.	Global evaluation measures for the combined acoustic and EMA synthesis.	136
6.4.	Global evaluation for the EMA-only synthesis. . . . .	138
6.5.	Global evaluation for the EMA-only synthesis restricted to the tongue coils.	139
6.6.	Global evaluation for the tongue model parameters synthesis. . . . .	140
6.7.	Comparison of results for old and new setup. . . . .	143

# List of Acronyms

<b>2D</b>	two-dimensional
<b>3D</b>	three-dimensional
<b>BAP</b>	aperiodicity per band
<b>CAPT</b>	computer-aided pronunciation training
<b>CBCT</b>	cone beam computed tomography
<b>CR</b>	cineradiography
<b>CT</b>	computed tomography
<b>DNN</b>	deep neural network
<b>DoF</b>	degrees of freedom
<b>EMA</b>	electromagnetic articulography
<b>EPG</b>	electropalatography
<b>F<sub>0</sub></b>	fundamental frequency
<b>fps</b>	frames per second
<b>HMM</b>	hidden Markov model
<b>HOSVD</b>	higher-order singular value decomposition
<b>HTS</b>	HMM/DNN-based speech synthesis system
<b>LCA</b>	linear component analysis
<b>MGC</b>	mel-generalized cepstral coefficients
<b>MLSA</b>	mel log spectrum approximation
<b>MRI</b>	magnetic resonance imaging
<b>NMR</b>	nuclear magnetic resonance
<b>PCA</b>	principal component analysis
<b>RMSE</b>	root mean square error

*List of Acronyms*

**rtMRI** real-time magnetic resonance imaging

**TTS** text-to-speech

**US** ultrasound

**UTI** ultrasound tongue imaging

**XRMB** X-ray microbeam

**ZNCC** zero normalized cross correlation

# 1. Introduction

## 1.1. Motivation

The human tongue plays an important role in everyday-life: people use it for eating, tasting, swallowing, or sometimes also for non-verbal communication. Moreover, as part of the human vocal tract, it interacts with other articulators, like, for example, the palate and the teeth, to produce speech. In this process, the whole vocal tract is able to assume a large quantity of shape configurations: the International Phonetic Alphabet<sup>1</sup> (International Phonetic Association, 2018) provides an overview on the different speech related sounds, referred to as *phones*, that can be created by humans. Therefore, it is of great interest in speech science to analyze the vocal tract during speech production to understand how articulators like the tongue move or change their shape to produce specific sounds.

Identifying the degrees of freedom (DoF) of shape changes the tongue can undergo during speech production, can be seen as one goal of such an analysis. After such DoF have been found, the obtained results can be used to build a tongue model. Basically, such a model takes as input some parameters and outputs a tongue shape. More specifically, it is desirable for such a model to have the following properties: first, it should be a statistical model, i.e., it can be used to measure the plausibility of a tongue shape, which helps to avoid generating unrealistic shapes. Second, its parameter set should distinguish between DoF that are related to the anatomy of a speaker and the ones that correspond to the speech related tongue pose. Separating the anatomy from the tongue pose 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, Lancia, et al., 2013; Yunusova, Rosenthal, et al., 2012). Third, the parameter set should be relatively small in order to limit the complexity of the model. Fourth, the model should be able to generate the whole three-dimensional (3D) surface of the tongue that is relevant for speech production. Lastly, it should use a shape representation that can easily be integrated into various applications.

## 1.2. Areas of application

A tongue model with the described properties can be used as prior knowledge for registering articulatory data. Here, it is particularly useful for reconstructing the full tongue shape from data that is incomplete or very sparse, like motion capture recordings. In

---

<sup>1</sup>For reference, it is also available in the appendix of this work (Appendix A).

## 1. Introduction

this regard, the model may be utilized to find a realistic tongue shape that is as close as possible to the data. Such a model can also be used to equip virtual avatars for multi-modal spoken interaction with a more natural animation of the tongue. In this context, it is very important 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, shape information provided by the model 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). Such a tutoring application may also give real-time feedback about the user’s current tongue shape by reconstructing it from motion capture data (Katz et al., 2014). A tongue model can also be employed in an articulatory speech synthesis framework to help approximate the vocal tract area function. It can also help to perform speaker normalization, that is, investigate only shape variations of an articulation that are independent of the speaker anatomy.

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 motion capture data of the tongue. In the case of estimating the tongue shape from motion capture data, using the wrong anatomy of the tongue may keep the model from registering the data correctly.

For completeness, it should be noted that another class of tongue models exists, 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 (Buchillard, Brix, et al., 2007), or for studying muscle activation during speech (Buchillard, Perrier, et al., 2009; Wu et al., 2014). However, for the presented target application areas, such a model may be regarded as too complex.

### 1.3. Magnetic resonance imaging

In order to derive DoF of the tongue that are speech related, data needs to be available that shows the 3D structure of the vocal tract during speech production. 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. Roughly speaking, this modality may be used to estimate the hydrogen density at specific spatial locations of a region of interest by using strong magnetic fields. Details about this



modality may be found, e.g., in Brown et al. (2014). This modality is considered to be non-invasive and non-hazardous to the subject (Schenck, 2000), and 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 et al., 1993; Shadle et al., 1999) and two-dimensional (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, Nayak, 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, Toutios, et al., 2016; Lingala, Zhu, et al., 2017) further reduced the acquisition time and improved the quality of obtained scans. For example, Lingala, Zhu, et al. (2017) reported rtMRI scanning at 83 fps for a single slice, or 27 fps for three slices. Recently, progress has been made towards 3D rtMRI (Y. Lim et al., 2019).

## 1.4. Related work

A sizable body of research has focused on analyzing the vocal tract shape during speech production using different modalities, including X-ray, X-ray microbeam (XRMB), electropalatography (EPG), cineradiography (CR), ultrasound (US), cone beam computed tomography (CBCT), real-time magnetic resonance imaging (rtMRI), or computed tomography (CT). Table 1.1 provides an overview of previous studies.

Even some of the earliest studies aimed at analyzing the anatomical and speech related shape differences 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, Alwan, et al. (1995) and Narayanan, Alwan, et al. (1997) analyzed shape variabilities using 3D MRI data. Analysis on 2D MRI was conducted by Hoole, Wismüller, et al. (2000), Ananthakrishnan et al. (2010), Vargas, Badin, and Lamalle (2012), and Vargas, Badin, Ananthakrishnan, et al. (2012). Zheng et al. (2003) performed this analysis on sparse sets of 65 points that were manually extracted from 3D MRI scans. Kaburagi (2015) used principal component analysis (PCA) to analyze the vocal tract area functions of ten speakers obtained from MRI. The work by Woo, Xing, et al. (2015) used dynamic MRI to build a spatio-temporal atlas of the vocal tract. Woo, J. Lee, et al. (2015) analyzed a high resolution atlas of the vocal tract using PCA. In the study by Stone, Woo, et al. (2018), the muscle architectures of different subjects were investigated. Speaker normalization was performed by Geng and Mooshammer (2009) and Serrurier et al. (2017).

## 1. Introduction

Work	Modality	Analyzed Data	Subjects
Mermelstein (1973)	X-ray	2D contours	1
Harshman et al. (1977)	X-ray	2D contours	5
Perkell and Nelson (1985)	XRMB	points	3
Baer et al. (1991)	MRI	vocal tract area functions and shapes	4
Stone and Lele (1992)	US	fitted polynomial functions	1
Narayanan, Alwan, et al. (1995)	MRI	shapes	4
Stone and Lundberg (1996)	3D US + EPG	interpolated meshes	1
Tiede, Yehia, et al. (1996)	MRI	cross-section shapes	1
Narayanan, Alwan, et al. (1997)	MRI + EPG	shapes	4
Badin, Bailly, Raybaudi, et al. (1998)	MRI + CR	meshes	1
Engwall and Badin (1999)	MRI	meshes + 2D contours	1
Engwall (2000a)	MRI	meshes	1
Hoole, Wismüller, 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, Bailly, Revéret, 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
Vargas, Badin, and Lamalle (2012)	MRI	2D contours	7
Toutios and Narayanan (2015)	rtMRI	2D contours	1
Kaburagi (2015)	MRI	vocal tract area functions	10
Woo, Xing, et al. (2015)	dynamic MRI	images	18
Woo, J. Lee, et al. (2015)	MRI	deformation fields	20
Fang et al. (2016)	MRI + CBCT	meshes	1
Serrurier et al. (2017)	MRI	2D contours	11
Stone, Woo, et al. (2018)	MRI	muscle architectures	14

Table 1.1.: Overview of several studies that have investigated shape variabilities of the vocal tract. The table lists the modality (or modalities) used, the analyzed data representation, and the number of subjects taking part in the corresponding study.

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, Woo, et al. (2017) used cine MRI to derive speaker-specific biomechanical models.

For the intended purposes, the anatomical and speech related variations of the 3D tongue surface have to be analyzed. Initial work investigating these variations obtained from MRI data of 9 speakers was presented by Hoole, Zierdt, and Geng (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; Labrunie et al., 2018; Somandepalli et al., 2017; Valliappan et al., 2018), produce only a low-level volume segmentation (J. Lee et al., 2013), or require an anatomical expert to provide the tongue templates (Harandi, Abugharbieh, et al., 2015).

## 1.5. Contributions

The contribution of this thesis may be summarized as follows. It proposes a semi-supervised framework for estimating the tongue shape from provided volumetric MRI datasets that is speaker and tongue pose independent, which eliminates the need for manually extracting the corresponding shapes. Furthermore, statistical tongue models are built and evaluated that fulfill the aforementioned properties:

- two sets of parameters: one for the anatomy, one for the tongue pose,
- relatively small number of parameters,
- ability to generate 3D shape of tongue, and
- usage of shape representation that can easily be integrated into applications.

Additionally, the tongue models are used to construct a semi-supervised and speaker-adaptive registration approach for sparse motion capture data of the tongue. Finally, the derived models are embedded into a multimodal and statistical text-to-speech (TTS) system that synthesizes speech with synchronized tongue motions. The thesis itself is based on the following publications:

Hewer, Alexander, Ingmar Steiner, Timo Bolkart, Stefanie Wuhler, and Korin Richmond (Aug. 2015). “A statistical shape space model of the palate surface trained on 3D MRI scans of the vocal tract”. In: *International Congress of Phonetic Sciences*. Glasgow, Scotland, pp. 0724.1–0724.5. URL: <https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0724.pdf>.

## 1. Introduction

- Hewer, Alexander, Ingmar Steiner, and Korin Richmond (Mar. 2019). “Analysis of coarticulation using EMA data with a statistical shape space model of the tongue”. In: *Conference on Electronic Speech Signal Processing*. Dresden, Germany, pp. 296–303. URL: [http://www.essv.de/pdf/2019\\_296\\_303.pdf](http://www.essv.de/pdf/2019_296_303.pdf).
- Hewer, Alexander, Ingmar Steiner, and Stefanie Wuhrer (Sept. 2014). “A hybrid approach to 3D tongue modeling from vocal tract MRI using unsupervised image segmentation and mesh deformation”. In: *Interspeech*. Singapore, pp. 418–421. URL: [http://www.isca-speech.org/archive/interspeech\\_2014/i14\\_0418.html](http://www.isca-speech.org/archive/interspeech_2014/i14_0418.html).
- Hewer, Alexander, Stefanie Wuhrer, Ingmar Steiner, and Korin Richmond (2016). “Tongue mesh extraction from 3D MRI data of the human vocal tract”. In: *Perspectives in Shape Analysis*. Springer, pp. 345–365. DOI: 10.1007/978-3-319-24726-7\_16.
- (Sept. 2018). “A multilinear tongue model derived from speech related MRI data of the human vocal tract”. In: *Computer Speech & Language* 51, pp. 68–92. DOI: 10.1016/j.csl.2018.02.001.
- James, Kristy, Alexander Hewer, Ingmar Steiner, and Stefanie Wuhrer (Sept. 2016). “A real-time framework for visual feedback of articulatory data using statistical shape models”. In: *Interspeech*. San Francisco, CA, USA, pp. 1569–1570. URL: [http://www.isca-speech.org/archive/Interspeech\\_2016/abstracts/2019.html](http://www.isca-speech.org/archive/Interspeech_2016/abstracts/2019.html).
- Le Maguer, Sébastien, Ingmar Steiner, and Alexander Hewer (Aug. 2017). “An HMM/DNN comparison for synchronized text-to-speech and tongue motion synthesis”. In: *Interspeech*. Stockholm, Sweden, pp. 239–243. DOI: 10.21437/Interspeech.2017-936.
- Steiner, Ingmar, Sébastien Le Maguer, and Alexander Hewer (Dec. 2017). “Synthesis of tongue motion and acoustics from text using a multimodal articulatory database”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12, pp. 2351–2361. DOI: 10.1109/TASLP.2017.2756818.

## 1.6. Thesis overview

The next chapter presents a basic semi-supervised method for estimating the tongue shape from MRI data. Additionally, this chapter introduces the MRI datasets that are used throughout this work. Chapter 3 provides some background on statistical shape analysis methodologies that are relevant for the rest of the thesis. Afterwards, the basic shape extraction framework is extended in Chapter 4 to increase the amount of vocal tract configurations the approach can handle. In particular, issues of the basic approach that were discovered in Chapter 2 and Chapter 3 are resolved. Chapter 4 also derives tongue models from the used datasets and evaluates them accordingly to find the best model. The following Chapter 5 uses the constructed model to register sparse motion capture data in order to generate realistic tongue motions. Chapter 6 describes the construction of a multimodal TTS system that synthesizes speech with synchronized tongue motion. Finally, the last chapter summarizes the thesis and provides an outlook for potential future work.

## 2. Extracting articulator shape information from MRI data

### 2.1. Introduction

#### 2.1.1. Motivation

Deriving a three-dimensional (3D) tongue model that provides access to the degrees of freedom (DoF) during speech production is the main goal of this work. In this regard, it is important to choose the right modality for analyzing the surface shape of the tongue during articulation. Previously, some modalities were mentioned that are or were used actively for investigating the behavior of the vocal tract. 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 electromagnetic articulography (EMA), it is able to provide dense volumetric measurements.

In order to analyze the tongue's shape, it is necessary to select a suitable shape representation. A polygon mesh is a good choice in this case because it 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). Furthermore, polygon meshes already have been successfully used for statistical shape analysis of, e.g., human bodies (Allen et al., 2003), faces (Banz and Vetter, 1999), or tongues (Badin and Serrurier, 2006).

#### 2.1.2. Related work

Now the question arises how such meshes that are needed for a statistical analysis can be extracted from MRI scans. Previous studies (Badin, Bailly, Revéret, et al., 2002; Badin and Serrurier, 2006; Badin, Elisei, et al., 2008; Engwall, 2003; Hoole, Zierdt, and Geng, 2003; Fang et al., 2016) obtained such meshes by manually annotating the MRI data. However, such a manual approach is tedious, very time-consuming, and requires annotators with an expertise in human anatomy. Additionally, the results of such a strategy might be hard to reproduce because different annotators might disagree on the shape of the tongue in the same MRI scan due to a personal bias or experience. Clearly, methods are needed that facilitate this process. They should at least make it semi-supervised and reproducible. This is also motivated by the fact that nowadays a lot of

## 2. Extracting articulator shape information from MRI data

MRI recordings of speech production can be obtained in a short period of time, which would make manual annotation highly infeasible.

Extracting and estimating the tongue shape from such data is an active field of research: Peng et al. (2010) employed an approach based on active contours (C. Li et al., 2007) to find the contour of the tongue in a two-dimensional (2D) mid-sagittal scan, using a previously trained shape model to control the evolution of the contour. Eryildirim and Berger (2011) extended this approach to align the contour’s end points to the corresponding extremities of the tongue. Raeesy et al. (2013) demonstrated that oriented active shape models (Jiamin Liu and Udupa, 2009) can be trained to reliably identify the boundary of the tongue in 2D MRI scans. The method by Labrunie et al. (2018) uses modified active shape models for segmenting mid-sagittal images obtained from real-time magnetic resonance imaging (rtMRI). In this context, approaches based on neural networks are also used for this purpose, like, e.g., the methods by Somandepalli et al. (2017) and Valliappan et al. (2018). These methods rely on manually preparing a training set and are limited to the 2D case. Another technique for 2D is the one by Su et al. (2018) that uses a snake model (Kass et al., 1988) that is geometrically constrained.

Studies focusing on the 3D shape also exist: J. Lee et al. (2013) presented a framework for minimally supervised tongue segmentation from 3D dynamic MRI. They used the random walker approach (Grady, 2006) as the base segmentation technique, which requires seeds manually provided by the user. This approach only provides access to a low-level volume segmentation which has to be further processed. Harandi, Abugharbieh, et al. (2015) employed a template-matching technique to generate a mesh representation of the tongue from 3D MRI scans. They used a mesh created by an expert from a source scan as their template, which is then deformed using color information to match a target scan. Specifically, the mesh points are moved in such a way that the color at the original point in the source scan is similar to the deformed point in the target scan. This approach is limited by requiring an expert to provide the templates.

### 2.1.3. Contribution

The contribution of this chapter may be summarized as follows: it describes a basic semi-supervised way for estimating the 3D tongue shape from volumetric MRI data. In particular, this approach uses image processing methods and template matching to extract polygon meshes from provided data. The method is semi-supervised in the sense that users only have to provide a small set of annotations on the scan and set a few parameters. In contrast to previous work, it is independent of the presence of an anatomical expert. Furthermore, training data is not required in the presented framework. This chapter is based on, and extends, the following papers:

Hewer, Alexander, Ingmar Steiner, and Stefanie Wuhler (Sept. 2014). “A hybrid approach to 3D tongue modeling from vocal tract MRI using unsupervised image segmentation and mesh deformation”. In: *Interspeech*. Singapore, pp. 418–421. URL: [http://www.isca-speech.org/archive/interspeech\\_2014/i14\\_0418.html](http://www.isca-speech.org/archive/interspeech_2014/i14_0418.html).

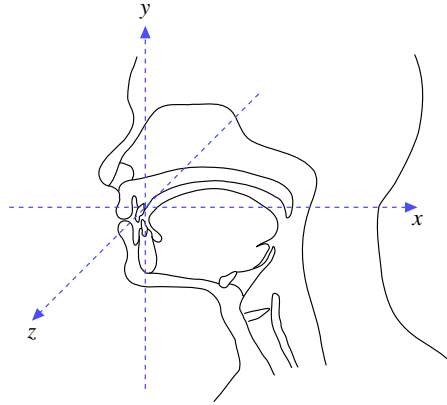


Figure 2.1.: Simplified illustration showing orientation of the coordinate system used for the MRI scans. Figure adapted from Richmond, Hoole, et al. (2011). It is important to note that the origin of the shown coordinate system was chosen arbitrarily and in general does not correspond to the true origin of the considered MRI scans.

Hewer, Alexander, Stefanie Wuhrer, Ingmar Steiner, and Korin Richmond (2016). “Tongue mesh extraction from 3D MRI data of the human vocal tract”. In: *Perspectives in Shape Analysis*. Springer, pp. 345–365. DOI: 10.1007/978-3-319-24726-7\_16.

#### 2.1.4. Overview

The chapter is organized as follows. First, it is discussed how an MRI scan may be represented as a 3D image. Afterwards, the three datasets of MRI scans used in this work are presented. Then, the approach is incrementally derived where each step is motivated by the current immediate results of the process. These steps may be summarized as follows: the quality of the scan is first enhanced by applying a filter. Then, the scan is segmented to gain access to the spatial support of the tongue. This segmentation is used to derive a surface point representation of the data. Finally, this representation is utilized to estimate a polygon mesh of the tongue shape. After the approach has been fully described, experiments are conducted to assess its performance. Finally, the conclusion section provides a summary and outlines needed extensions of the approach.

## 2.2. Data representation

For visualization purposes and for simplicity, an MRI scan is interpreted as a 3D image in this work, which also offers the advantage that image processing methods can be applied to the resulting image. Such an image is defined as follows:

$$f : \Omega \rightarrow I. \quad (2.1)$$

## 2. Extracting articulator shape information from MRI data

Here, the measured nuclear magnetic resonance (NMR)<sup>1</sup> at location  $\mathbf{x} \in \Omega$  is quantized and represented as the gray value  $f(\mathbf{x}) \in I$  where a standard visualization uses bright colors to indicate a high NMR, and dark colors for a low resonance measurement. The image domain  $\Omega \subset \mathbb{R}^3$  represents the area of interest that was measured in the scan acquisition. An illustration showing the orientation of the coordinate system with respect to the head can be inspected in Figure 2.1. The set  $I \subset \mathbb{R}$  is a range of all possible gray values that may occur during the quantization process where a common choice is the interval  $I = [0, 255]$ . As a convention, the following abbreviations are used in the remainder of this chapter if the meaning is clear from the context:

Original	Abbreviation	Meaning
$f(\mathbf{x})$	$f$	gray value of image $f$ at position $\mathbf{x}$
$\frac{\partial f(\mathbf{x})}{\partial a}$	$\partial_a f$ or $f_a$	partial derivative along axis $a$ of $f$ at $\mathbf{x}$
$\nabla f(\mathbf{x})$	$\nabla f$	gradient of $f$ at $\mathbf{x}$
$\text{div}(f(\mathbf{x}))$	$\text{div}(f)$	divergence of $f$ at position $\mathbf{x}$

For simplicity, the region  $\Omega$  is assumed to be continuous, bounded, rectangular, and connected throughout this work. Furthermore, the image obeys the following von Neumann boundary conditions on the boundary  $\partial\Omega$  of the image domain:

$$\frac{\partial f}{\partial x} = 0, \quad (2.2)$$

$$\frac{\partial f}{\partial y} = 0, \quad (2.3)$$

$$\frac{\partial f}{\partial z} = 0 \text{ on } \partial\Omega. \quad (2.4)$$

These boundary conditions serve the purpose of making the image differentiable on the boundary  $\partial\Omega$ .

Because such images are 3D, it is difficult to visualize them directly on 2D media like a computer screen. Therefore, it is common to only visualize 2D parts of it, so-called slices. Figure 2.2 shows different types of such slices. Basically, such slices show the gray values that belong to specific axis-aligned image planes:

Slice type	Image plane
sagittal	$xy$ -plane
coronal	$yz$ -plane
transverse	$xz$ -plane

By inspecting the examples in Figure 2.2, it becomes apparent that actual MRI scans are of discrete nature, i.e., there exist  $\mathbf{x} \in \Omega$  and  $\mathbf{y} \in \Omega$  such that information about the NMR is missing between those locations. In simplified terms, this means that the MRI acquisition process samples the continuous image domain  $\Omega$  at specific locations

<sup>1</sup>correlated with hydrogen molecule density, i.e., high for soft tissue, low for bone and air



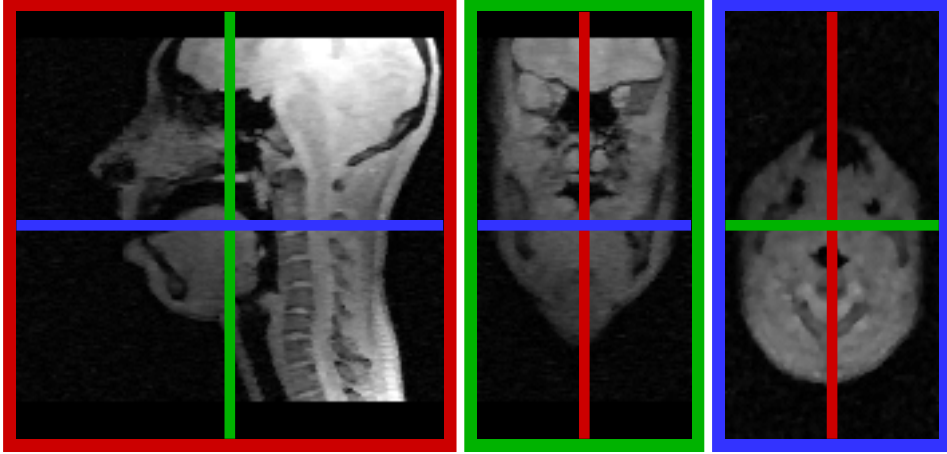


Figure 2.2.: Different slice views of the same MRI scan: sagittal (left), coronal (center), and transverse slice (right). Colored lines indicate where the individual slices are located within the different visualizations.

and produces a discrete 3D image. The sample positions are in general located on a regular grid with the axis-aligned grid spacings  $h_x, h_y$ , and  $h_z$ . Thus, it is possible to access the individual samples or voxels of the discrete image by using a unique index  $\mathbf{i} := (i, j, k)^\top \in \mathbb{O} \subset \mathbb{N}^3$ , such that the gray values of the discrete image are given by:

$$[f]_{\mathbf{i}} := [f]_{i,j,k} \approx f(\omega(i, j, k)), \quad (2.5)$$

where the mapping function  $\omega : \mathbb{O} \rightarrow \Omega$  is defined as

$$\omega(i, j, k) := (ih_x, jh_y, kh_z)^\top. \quad (2.6)$$

Here,  $\mathbb{O}$  may be seen as the discrete counterpart to  $\Omega$ . Having a relationship between the continuous image and its discrete version offers the advantage that image processing methods may be modeled in the continuous setting and then transferred to the discrete version.

In some situations, it might be necessary to gain access to gray values at an unsampled position  $\mathbf{x} \in \Omega$ . These values may be obtained by performing interpolation:

$$f(\mathbf{x}) \approx [f](x, y, z) := \text{interpolate}([f]_{x,y,z}). \quad (2.7)$$

In this work, a simple linear interpolation is used.

## 2.3. Datasets

This section serves the purpose of describing the different datasets that are used throughout this work. In summary, information is provided about

- the amount of recorded speakers,

## 2. *Extracting articulator shape information from MRI data*

- the used phonetic inventory,
- the used MRI scanner,
- and properties of the resulting MRI scans.

### 2.3.1. **Ultrax dataset**

The dataset of the Ultrax project (Richmond and Renals, 2012) 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, ε, a, α, ʌ, ɔ, o, u, ʊ, ə, s, ʃ]<sup>2</sup>. Thus, this dataset focuses on vowels and the two sibilants [s] and [ʃ].

The acquisition took place at the Clinical Research Imaging Centre in Edinburgh using a Siemens Verio 3T scanner; the scans were recorded 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  $320 \times 320$  pixels. The grid spacings are  $h_x = h_y = 1.1875$  mm and  $h_z = 1.2$  mm.

The labels for the individual subjects consist of two digits and the character sequence MRI $X$  where  $X$  is either F (female subject) or M (male subject).

Currently, this dataset is only available for internal use.

### 2.3.2. **Baker dataset**

The Baker dataset (A. Baker, 2011) is a collection of speech related volumetric MRI recordings of a single speaker. This data was recorded as part of the Ultrax project, but released separately. It consists of 25 scans of one male speaker that are speech related and represent different articulatory configurations. Its phone set is given by [α, i, u, ɪ, æ, ε, ʌ, ʊ, e, o, ɪ, l, w, m, n, ŋ, p, t, k, q, v, δ, z, ʒ, ʎ]. Thus, it contains more vocal tract configurations than the Ultrax data, especially with respect to consonants. However, it lacks scans for the phones [a, ɔ, ʊ, ə, s, ʃ].

According to the author, the speaker used creaky voice during the recordings to make his breath last longer and thus allow for the required acquisition time of about 20 s. The scans have the same properties as the other scans of the Ultrax project.

In contrast to the Ultrax dataset, this part is freely available. As a convention, the data of the Baker dataset is combined with the Ultrax dataset in the following, which increases the speaker amount of the dataset to 12. There, the speaker of the Baker dataset is referred to as 01MRIM.

### 2.3.3. **USC dataset**

The USC dataset (Sorensen et al., 2017) consists of volumetric MRI scans of sustained sound and rtMRI recordings with synchronized audio. The volumetric part of it contains

---

<sup>2</sup>Appendix A provides background information on these symbols.

vocal tract scans of 17 speakers. The original paper states that all recorded subjects were native speakers of American English and that English was the only language they could speak fluently. An acquisition of one scan took around 7s. In terms of phonetic inventory, the dataset provides a balance between vowels (13) and consonants (14): [ə, eɪ, æ, ɪ, ε, ɜ, ɪ, oʊ, uɪ, ɔɪ, ʌ, ɑ, ʊ, f, ʒ, h, l, m, n, ŋ, ɹ, s, ʃ, θ, ð, v, z]

In this case, the labels for the subjects consist of one character and one digit. The character is either F (female) or M (male).

This work also uses the pilot speaker of this dataset that was kindly provided by the authors, which results in a total amount of 18 subjects. This speaker is missing a scan for the phone [ə]. Moreover, the release of the database used in this work is lacking a recording of [ɑ] for speaker F5.

The recordings were made by using a GE Healthcare 3.0 T Signa Excite HD MRI scanner. The resulting scans consist of 80 sagittal slices with a thickness of 1.5625 mm where each slice has a size of  $160 \times 160$  pixels. The grid spacings are  $h_x = h_y = h_z = 1.5625$  mm.<sup>3</sup>

This dataset excluding the pilot speaker is freely available for research purposes.

### 2.3.4. Discussion

Two aspects merit mentioning for these MRI datasets. First of all, it is necessary to be aware of the following fact: literature (Engwall and Badin, 1999; Badin, Borel, et al., 2000; Kitamura et al., 2005; Engwall, 2006) has found indicators that static MRI recordings of sustained vocal tract configurations may show unnatural articulation. This might be caused by the supine position of the subject during recording (Tiede, Masaki, Wakumoto, et al., 1997; Tiede, Masaki, and Vatikiotis-Bateson, 2000; Kitamura et al., 2005; Stone, Stock, et al., 2007; Steiner and Ouni, 2011; Steiner, Knopp, et al., 2014) and the long acquisition time (Engwall and Badin, 1999). In this regard, Engwall (2000b) also found that static articulations used for volumetric MRI recordings were hyperarticulated. However, it is important to say that such recordings still provide access to human tongue shapes. Thus, it is worthwhile to analyze such data to estimate the DoF of the tongue, even if the articulation is unnatural.

Second, it is important to note that the used corpora of MRI data can be regarded as small, which means that only a few subjects and vocal tract configurations were recorded. For example, the largest dataset contains scans for 18 speakers. In comparison, databases of 3D face scans (Yin, Wei, et al., 2006; Yin, Chen, et al., 2008; Savran et al., 2008), for example, often offer on average recordings of 100 subjects. Thus, it might be argued that this data is insufficient for reliably deriving the DoF of the human tongue shape. 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

---

<sup>3</sup>The spacings in the original paper are incorrect (A. Toutios, personal communication).

## 2. *Extracting articulator shape information from MRI data*

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, it may be stated that even limited datasets are a valuable resource and can lead to useful results.

Finally, these datasets only provide access to static information of the vocal tract configuration. Thus, insight is missing about the natural transitions between phones. In this regard, 3D rtMRI might be helpful to address this issue in the future.

### 2.4. Observations

Figure 2.3 shows example scans from the three datasets. Here, it can be seen that the scans contain much more information than the vocal tract itself. Cropping the individual images to a region of interest showing the vocal tract leads to an observation: the data is suffering from degradations like noise and vignetting artifacts caused by the placement of the coils needed for the MRI acquisition. Thus, the MRI scanning and reconstruction process might fail to estimate the correct values, therefore the discrete image should be assumed to contain only approximations of the original values.

However, these images also lead to another interesting observation: the spatial support of the tongue region can be visually identified and thus the shape may be estimated from such MRI scans. This means that it is possible to annotate the scan manually and then extract the tongue shape from it, which has been successfully done by, e.g., Badin, Bailly, Revéret, et al. (2002) and Engwall (2003). Like previously stated, manual annotation is tedious, time-consuming, and moreover the results are hard to reproduce because experts performing the annotation might introduce a personal bias. Thus, it is worthwhile to automate the process as much as possible. The different quality of the datasets also makes it necessary to make the approach flexible, such that it can easily be adapted. In simple terms, the process presented here is split into three main steps:

1. Preprocess scans
  - a) crop to region of interest containing the tongue
  - b) enhance quality of scan
2. Identify spatial support of tongue and related tissue
3. Use found spatial support to estimate tongue shape

In this context, the preprocessing serves two purposes: on the one hand, it prepares the data for the following steps. On the other hand, it aims at improving the quality of the scans to make visual inspection easier. For the sake of brevity, the description of the cropping step is omitted here. Example results for the individual steps are shown in Figure 2.3.

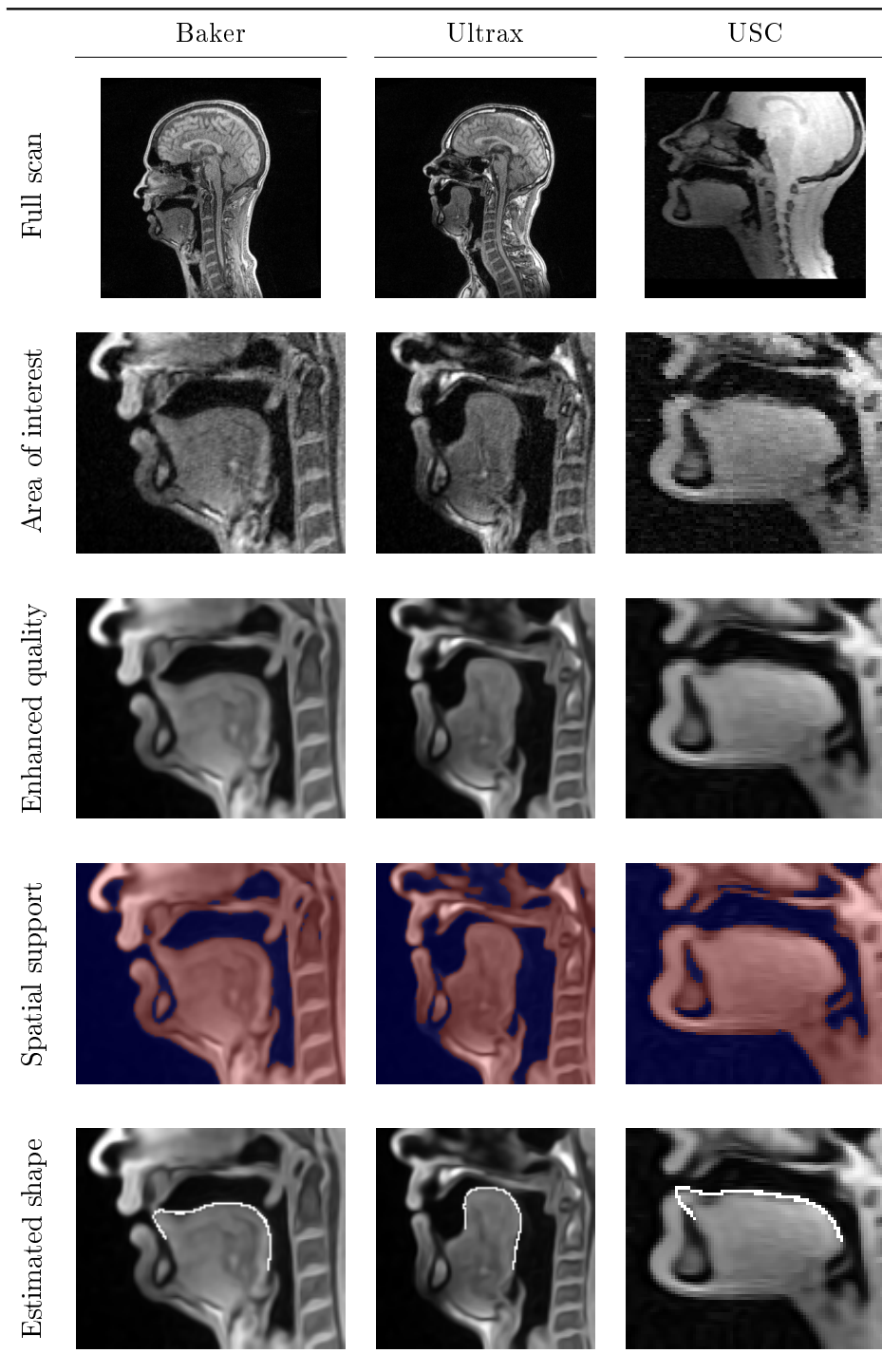


Figure 2.3.: Sagittal views of scans from the three datasets. Rows show example results of the shape extraction process.

## 2.5. Dealing with noise

As previously stated, available MRI data often suffers from degradation: for example, high frequency measurement noise, small gaps in the contours, or vignetting-like effects make it difficult to identify the shape of the tongue. As a remedy, image processing methods can be used to improve the quality of the images before further processing can take place. In this field, a lot of different filters have been proposed to deal with this kind of degradation. This section is dedicated to presenting and discussing some means of denoising MRI data. The reader is encouraged to have a look at Aubert and Kornprobst (2006), Gonzalez and Woods (2017), and Szeliski (2010) for a bigger selection of available filters and more theoretical background.

### 2.5.1. Gaussian smoothing

A popular choice for removing high-frequency noise in images is a Gaussian filter. In this approach, the image is improved by convolving it with a 3D Gaussian kernel. The continuous version of this operation is given below:

$$f_\sigma := (K_\sigma * f)(\mathbf{x}) := \int_{\mathbb{R}^3} K_\sigma(\mathbf{x} - \mathbf{t}) f(\mathbf{t}) d\mathbf{t}. \quad (2.8)$$

$K_\sigma : \mathbb{R}^3 \rightarrow \mathbb{R}$  is the Gaussian kernel of the operation with standard deviation  $\sigma$  and mean  $\mathbf{0}$ :

$$K_\sigma(\mathbf{x}) = \frac{1}{\sqrt{(2\pi\sigma^2)^3}} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{x}\|^2\right). \quad (2.9)$$

In Figure 2.4, it can be seen that this kind of filter attenuates measurement noise. However, at the same time other vital information may be destroyed during the process: the filter may blur away important boundary information that is needed for identifying the tongue shape. This is due to the fact that this filter tries to smooth in any direction without taking any structural information into account.

### 2.5.2. Median filtering

Another tool for dealing with this kind of noise is a median filter. This is a filter that is most suited for dealing with salt-and-pepper noise. Such a filter operates on the image by using a so-called structuring element. An example for such a structuring element is a box that is defined by a center point  $\mathbf{x} \in \mathbb{O}$  and a radius  $r > 0$ :

$$B_r(\mathbf{x}) := \{\mathbf{y} \mid \|\mathbf{x} - \mathbf{y}\|_\infty \leq r\}. \quad (2.10)$$

In this case,  $\|\cdot\|_\infty$  represents the infinity norm. The median filter centers such a structuring element at each voxel of the image and replaces the gray value at this voxel with the median gray value of voxels located in this box, which leads to the filtered image  $g$ :

$$[g]_{\mathbf{x}} = \text{median}(\{[f]_{\mathbf{y}} \mid \mathbf{y} \in B_r(\mathbf{x})\}). \quad (2.11)$$

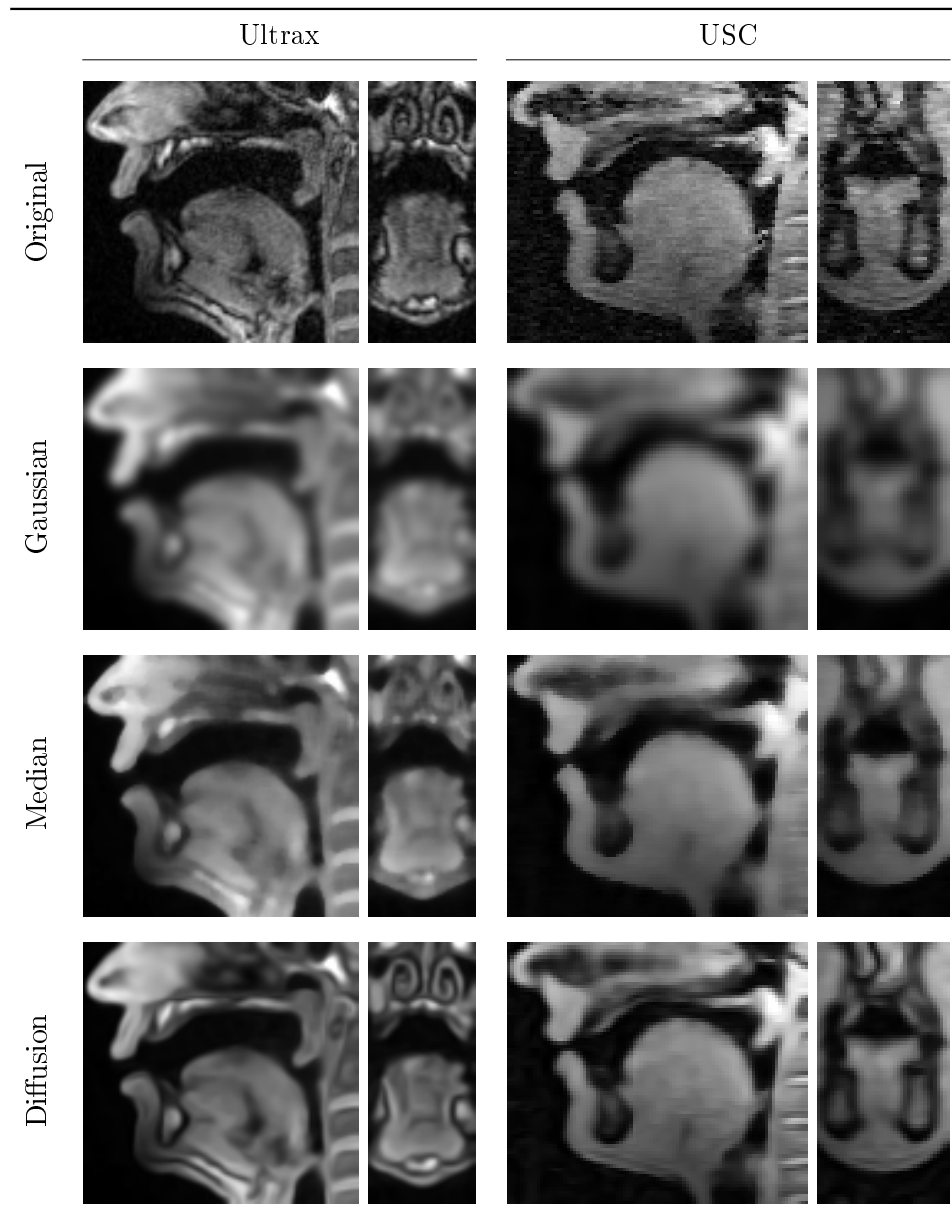


Figure 2.4.: Original version of scan and results for different smoothing filters. Sagittal and coronal views are shown. The individual filters used the following settings. Median filtering:  $r = 2$ . Gaussian convolution:  $\sigma = 2$ . Diffusion:  $\sigma = 1$ ,  $\rho = 1$ ,  $\lambda = 0.1$ , evolution time  $t = 2.4$ .

## 2. Extracting articulator shape information from MRI data

In Figure 2.4, the result of such a filtering can be seen. Compared to the Gaussian approach, boundaries are better preserved. However, the filter seems to erode the image at several locations. This can be seen, for example, in the hard palate area for the scan of the USC dataset. Here, the edge seems to disappear. Again, the filter operates on the image data without taking structural information into account. It is important to note here that this filter produces acceptable results for the Ultrax example scan: the noise is removed and boundaries are preserved.

### 2.5.3. Surface-enhancing diffusion filtering

The previous filters had one drawback: they ignored structural information during the smoothing process. Thus, a filter would be preferable that takes such information into account. In particular, it should remove noise, but preserve or even enhance important structures such as boundaries.

In the case of two dimensions, the edge-enhancing diffusion approach by Weickert (1998) provides these wanted features. This approach smooths along edge-like structures and inside regions where clear coherent structural features are present. In addition to that, it enhances the contrast across edges. Thus, it might be applied to the individual sagittal slices of a scan to obtain a denoised version of it. However, in this case, only the edges in the slices itself would be enhanced. The scan itself is a 3D structure where instead of edges surface-like structures occur that should be enhanced. These observations motivate to extend the approach to the 3D case. In the following, only the modeling ideas and the modeling itself are discussed. For the theoretical background, the reader is encouraged to consult the original work.

#### Diffusion process

The approach itself is modeled as a partial differential equation:

$$\partial_t u(\mathbf{x}, t) = \operatorname{div}_s (D(u(\mathbf{x}, t)) \nabla_s u(\mathbf{x}, t)). \quad (2.12)$$

The function  $u : \Omega \times T \rightarrow I$  describes how a scan is evolving over time with  $T := [0, \infty)$  being the time interval of the process. The state of the function at time 0 corresponds to the original image  $f$  that should be smoothed:

$$u(\mathbf{x}, 0) = f(\mathbf{x}). \quad (2.13)$$

The filtered version of the image can then be obtained by selecting the state of  $u$  at a specific time  $t$ . In the equation, the operators  $\operatorname{div}_s$  and  $\nabla_s$  only use the spatial coordinates  $\mathbf{x}$  of the function:

$$\operatorname{div}_s(u) := u_x + u_y + u_z, \quad (2.14)$$

$$\nabla_s(u) := (u_x, u_y, u_z)^\top. \quad (2.15)$$



### Diffusion tensor

The matrix  $D(u(\mathbf{x}, t))$  is called the diffusion tensor of the process. This tensor governs the smoothing directions with associated smoothing strength at the given position. It has to be constructed in such a way that the process has the wanted properties. Therefore, it should smooth along surface-like structures and inside coherent regions. Additionally, in the first case, it should enhance the contrast along the normal of the surface, which implies that the smoothing strength in this direction depends on the contrast.

In a first step, the required directions are estimated by means of the structure tensor (Förstner and Gülch, 1987):

$$J_{\sigma, \rho}(u, \mathbf{x}, t) := K_{\rho} * \left( \nabla_s (K_{\sigma} *_s u(\mathbf{x}, t)) \nabla_s (K_{\sigma} *_s u(\mathbf{x}, t))^{\top} \right). \quad (2.16)$$

Again, the notation  $*_s$  indicates that the convolution only affects the spatial components. In the computation, the image  $u$  at time  $t$  is first presmoothed by applying a Gaussian convolution with standard deviation  $\sigma$ , which makes the structure tensor robust against noise. Furthermore, the information in the matrices  $\nabla_s(\cdot) \nabla_s(\cdot)^{\top}$  in the local neighborhood is combined by convolving them entry-wise with a Gaussian kernel with standard deviation  $\rho$ . The result of this operation is a matrix that describes the structure in the spatial neighborhood around the given point: according to the terminology of Weickert (1998), the largest eigenvalue  $\lambda_1$  belongs to the direction of the highest gray value fluctuations, while the remaining values  $\lambda_2$  and  $\lambda_3$  can be thought of as the preferred local orientations, the coherence directions. As a consequence of generalizing the approach to three dimensions, two coherence directions are present instead of one.

By setting  $\lambda_2 = 1$  and  $\lambda_3 = 1$ , the diffusion process will always smooth along the coherence directions with a strength of 1. Now the question arises how to modify the largest eigenvalue  $\lambda_1$ : in coherent regions, it should be near to 1 in order to allow the process to smooth in all directions. However, in regions with a surface-like structure, the smoothing should be attenuated along the normal of this structure. To this end, the Perona-Malik diffusivity (Perona and Malik, 1990)  $\psi_{\lambda} : \mathbb{R} \rightarrow (0, 1]$  can be used that is defined as follows:

$$\psi_{\lambda}(x) := \frac{1}{1 + x^2/\lambda^2}. \quad (2.17)$$

The value  $\lambda > 0$  is called the contrast parameter of the diffusivity. Roughly speaking, it determines which surface-like structures should be preserved. For example, a high value requires a high contrast along the corresponding direction for the structure to be preserved.

Thus, the final diffusion tensor is given by:

$$D(u(\mathbf{x}, t)) := \psi_{\lambda}(J_{\sigma, \rho}(u, \mathbf{x}, t)). \quad (2.18)$$

In this context,  $\psi_{\lambda}(\cdot)$  is defined to be only applied to the largest eigenvalue of the corresponding matrix.

## 2. *Extracting articulator shape information from MRI data*

### Results

Results of this filter can be inspected in Figure 2.4. This time, the filtered images are denoised and important structural information is still present. Additionally, small gaps in the surface structures were filled by the approach. However, the better performance of the method comes with a drawback: instead of a single parameter like in the other two filters, this time, the 4 parameters  $\sigma$ ,  $\rho$ ,  $\lambda$ , and  $t$  have to be selected.

#### 2.5.4. Discussion

In this section, a few filters were presented and applied to scans of the used datasets. Example results revealed important observations: whereas the Gaussian filtering was a suboptimal choice in all cases, the median filtering was an appropriate choice for the scans of the Ultrax dataset. However, the median filter had issues with the scan of the USC data. This implies that a specific filter might provide good results for one dataset and fail for another one. Thus, it is important to visually inspect material of the dataset and investigate how well filters are performing. In this work, the surface-enhancing diffusion filter was performing favorably for all datasets. In particular, it preserved important boundary information and was able to close small gaps in the boundary. Therefore, this filter is used for the remainder of this work for preprocessing MRI data.

## 2.6. Detecting the spatial support

The initial step of the proposed process provides access to denoised scans that are cropped to a region of interest showing the vocal tract. Inspecting such scans in Figure 2.5 reveals an interesting observation: the spatial support of tissue can easily be distinguished from the support of other non-tissue objects by using color information. As the tongue consists of tissue, its spatial support can also be detected in this way. Tissue appears brighter than other material, which is due to the fact that tissue contains more hydrogen compared to, e.g., bones or air, which results in a higher NMR value. Thus, the goal of the current step is concerned with finding a partition  $\Omega = \Omega_O \cup \Omega_B$ , such that

- $\Omega_O$  describes the region of the tissue in the scan and
- $\Omega_B = \Omega \setminus \Omega_O$  consists of everything else in the scan, like, for example, bones or air.

In the field of image processing, such a process is called segmentation. In literature, several methods have been proposed to solve this problem. Again, Aubert and Kornprobst (2006), Gonzalez and Woods (2017), and Szeliski (2010) may be inspected for getting a small overview of available methods. In this section, several segmentation techniques are investigated in order to find out if they can be used to segment scans of the different MRI datasets. For visualization purposes, the region  $\Omega_O$  will be colored in red and  $\Omega_B$  in blue.

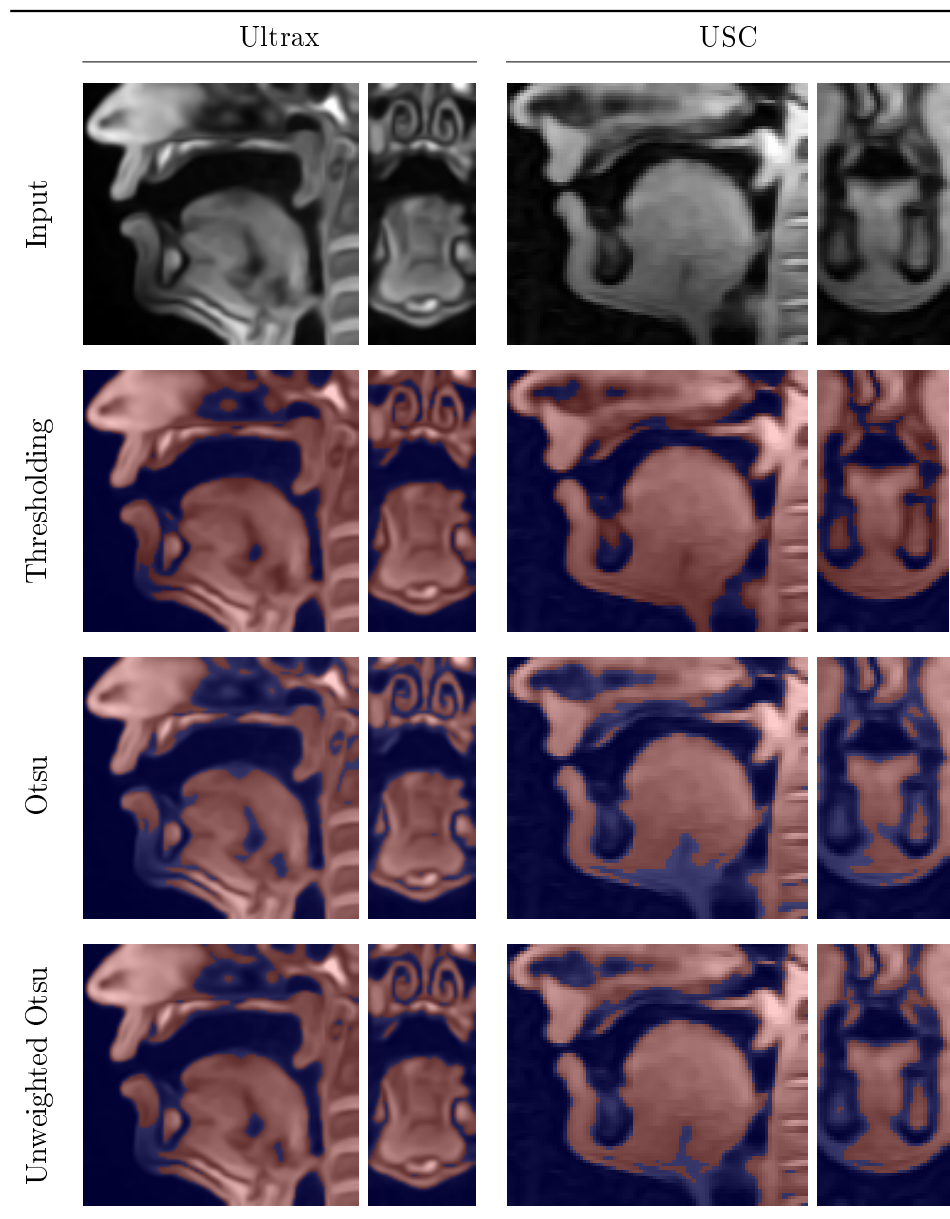


Figure 2.5.: Segmentation results for supervised and unsupervised approaches based on thresholding. Used parameters for manual thresholding: 15 (Ultrax) and 38 (USC).

## 2. Extracting articulator shape information from MRI data

### 2.6.1. Thresholding

One of the most basic segmentation methods is thresholding. Given a threshold parameter  $t \in \mathbb{R}$  and the image  $f$ , the regions  $\Omega_O$  and  $\Omega_B$  are derived as follows:

$$\Omega_O(t) := \{ \mathbf{x} \mid f(\mathbf{x}) \geq t \}, \quad (2.19)$$

$$\Omega_B(t) := \{ \mathbf{x} \mid f(\mathbf{x}) < t \}. \quad (2.20)$$

This is the explicit version of thresholding where the threshold parameter  $t$  has to be provided. Figure 2.5 shows example results of this method. In general, this approach performs well for all datasets although there are some small gaps in the object region. However, the threshold parameter has to be manually tuned to the corresponding scan.

### 2.6.2. Otsu's method

The method by Otsu (1979) tries to automatically derive the thresholding parameter from the data itself, which makes the resulting segmentation strategy fully automatic. Basically, it optimizes the thresholding parameter  $t$ , such that the following energy is minimized:

$$E_{\text{Otsu}}(t) = \Pr(\Omega_O(t)) \int_{\mathbf{x} \in \Omega_O(t)} (f(\mathbf{x}) - \mu_O(t))^2 d\mathbf{x} + \Pr(\Omega_B(t)) \int_{\mathbf{x} \in \Omega_B(t)} (f(\mathbf{x}) - \mu_B(t))^2 d\mathbf{x}. \quad (2.21)$$

The quantities  $\mu_O(t)$  and  $\mu_B(t)$  represent the mean gray values in the regions  $\Omega_O(t)$  and  $\Omega_B(t)$ , respectively. The values  $\Pr(\Omega_O(t))$  and  $\Pr(\Omega_B(t))$  denote the probability of a voxel belonging to  $\Omega_O(t)$  or  $\Omega_B(t)$ , respectively. Thus, it finds the threshold  $t$  that minimizes the gray value variance in each of the resulting regions.

In the original version, both variances are weighted according to the class probability, which in turn assigns much importance to the larger region. This autoscaling can be deactivated by reformulating the energy as follows:

$$E_{\text{unweighted Otsu}}(t) = \int_{\mathbf{x} \in \Omega_O(t)} (f(\mathbf{x}) - \mu_O(t))^2 d\mathbf{x} + \int_{\mathbf{x} \in \Omega_B(t)} (f(\mathbf{x}) - \mu_B(t))^2 d\mathbf{x}. \quad (2.22)$$

In this version, the variance in each region is penalized in the same way.

Results from both variants can be seen in Figure 2.5. It becomes clear that the original method by Otsu fails for both datasets because gaps are visible in the object region. Reasons for this behavior can be found in literature, e.g., Kittler and Illingworth (1985) or S. U. Lee et al. (1990): the bad performance may be attributed to the histogram of both scans that are degraded due to vignetting effects. Due to the vignetting effect, the gray value variance in the optimal region  $\Omega_O$  becomes too large and the method fails to find a suitable segmentation. However, the unweighted variant of the approach appears to provide acceptable results where the amount of gaps is reduced.

### 2.6.3. Chan-Vese segmentation

So far, methods relied on using a threshold parameter to separate the regions from each other. Another class of techniques uses a level set to derive the segmentation. Roughly speaking, a level set is a function  $l : \Omega \rightarrow \mathbb{R}$ , such that  $\Omega_O$  and  $\Omega_B$  can be derived as follows:

$$\Omega_O(l) := \{ \mathbf{x} \mid l(\mathbf{x}) \geq 0 \}, \quad (2.23)$$

$$\Omega_B(l) := \{ \mathbf{x} \mid l(\mathbf{x}) < 0 \}. \quad (2.24)$$

The zero set of  $l$  can then be seen as a surface  $C(l)$  that separates  $\Omega_O(l)$  from  $\Omega_B(l)$ . Such a level set is, for example, given by the following function:

$$L_{\mathbf{c},r}(\mathbf{x}) = 1 - \frac{\|\mathbf{x} - \mathbf{c}\|_2}{r}. \quad (2.25)$$

According to this level set, all points inside and on the surface of the sphere with radius  $r$  and center  $\mathbf{c}$  are part of  $\Omega_O$ . All other points are part of the background region.

One method working with level sets is the approach by Chan and Vese (2001). The original method is intended for 2D images, but can easily be extended to the 3D case. The wanted segmentation can be obtained by minimizing the following energy:

$$E_{\text{Chan-Vese}}(l) = \mu \text{area}(C(l)) + \eta \text{volume}(\Omega_O(l)) + \lambda_O \int_{\mathbf{x} \in \Omega_O(l)} (f(\mathbf{x}) - \mu_O(l))^2 + \lambda_B \int_{\mathbf{x} \in \Omega_B(l)} (f(\mathbf{x}) - \mu_B(l))^2, \quad (2.26)$$

where  $\text{area}(C(l))$  refers to the surface area of  $C(l)$  and  $\text{volume}(\Omega_O(l))$  to the volume of  $\Omega_O$ . It is important to note that this approach requires an initialization where a simple one suffices, like the one in Equation (2.25). Like Otsu's method, this approach tries to find a segmentation by minimizing the variance inside each of the regions. However, by using a level set instead of a threshold parameter, it offers the following advantage: the smoothness of the boundary between  $\Omega_O(l)$  and  $\Omega_B(l)$  can be adjusted by choosing an appropriate value for  $\mu$ , where high values increase the smoothness of the resulting segmentation boundary.

Example results can be seen in Figure 2.6. For both datasets, the approach produces acceptable results. Unfortunately, this method requires the user to specify four parameters, which limits its usefulness as an unsupervised method. Moreover, it needs a level set to be provided as an initialization.

### 2.6.4. Graph cut

A third class of methods tries to solve the segmentation problem by using a graph representation of the discrete image. Such a method is the graph cut approach by Boykov and Funka-Lea (2006). A basic form of this technique tries to maximize the energy

$$E_{\text{graph cut}}(\mathbb{O}_O, \mathbb{O}_B) := \sum_{\mathbf{i} \in \mathbb{O}_O} \sum_{\mathbf{j} \in \mathcal{N}(\mathbf{i}, \mathbb{O}_O)} g([f]_{\mathbf{i}}, [f]_{\mathbf{j}}) + \sum_{\mathbf{i} \in \mathbb{O}_B} \sum_{\mathbf{j} \in \mathcal{N}(\mathbf{i}, \mathbb{O}_B)} g([f]_{\mathbf{i}}, [f]_{\mathbf{j}}). \quad (2.27)$$

## 2. Extracting articulator shape information from MRI data

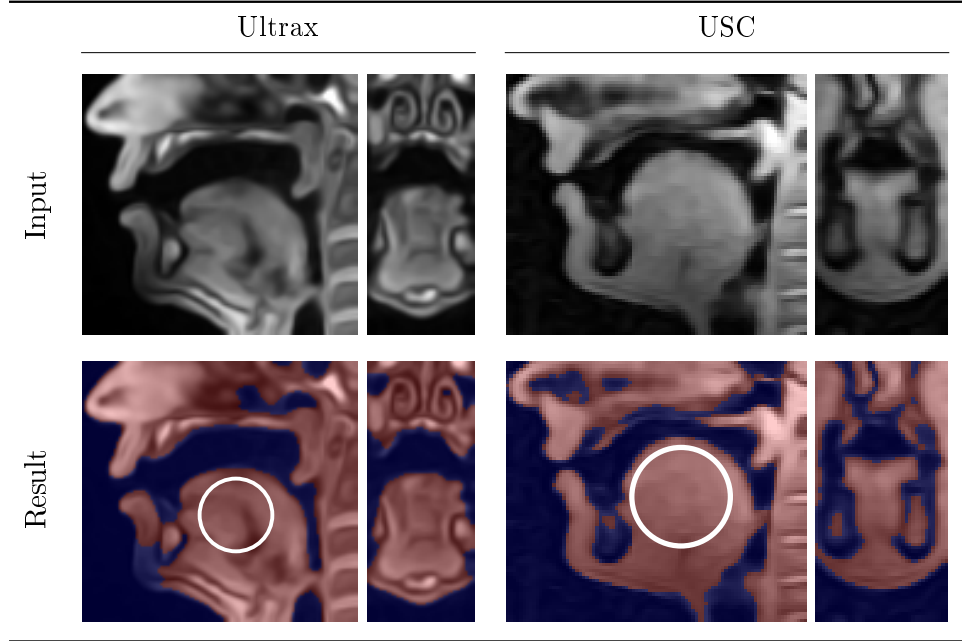


Figure 2.6.: Results of the Chan-Vese approach for example scans. White circles in the sagittal slices illustrate the zero set of the initial level set. Used parameters:  $\lambda_O = 1$ ,  $\lambda_B = 3$ ,  $\mu = 500$ ,  $\eta = 0$ .

Here, the partition  $\mathbb{O} = \mathbb{O}_O \cup \mathbb{O}_B$  is the discrete counterpart of  $\Omega = \Omega_O \cup \Omega_B$ . The set  $\mathcal{N}(\mathbf{i}, K)$  represents the direct neighbors of  $\mathbf{i}$  contained in  $K$ :

$$\mathcal{N}(\mathbf{x}, K) := \{\mathbf{j} \mid \|\mathbf{i} - \mathbf{j}\|_1 = 1 \wedge \mathbf{j} \in K\}. \quad (2.28)$$

with  $\|\cdot\|_1$  denoting the Manhattan metric. Finally,  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$  is a similarity measure between two gray values, like for example:

$$g(a, b) := e^{-|a-b|}. \quad (2.29)$$

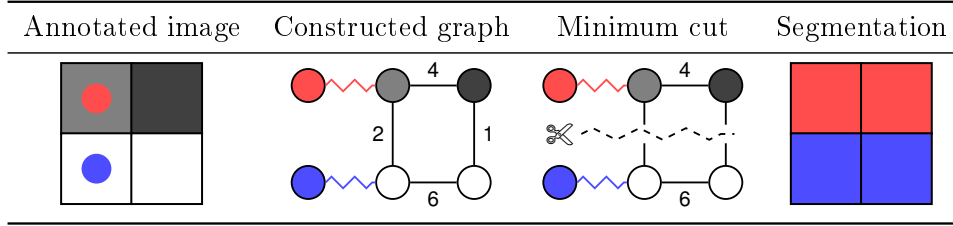
Thus, high values for  $g$  represent a high similarity while low values indicate a low similarity. This means that the graph cut approach wants to maximize the sum of neighbor similarities in each region.

The energy in Equation (2.27) may be rewritten:

$$E_{\text{graph cut}}(\mathbb{O}_O, \mathbb{O}_B) = \sum_{\mathbf{i} \in \mathbb{O}_O} \sum_{\mathbf{j} \in \mathcal{N}(\mathbf{i}, \mathbb{O}_O)} g([f]_{\mathbf{i}}, [f]_{\mathbf{j}}) + \sum_{\mathbf{i} \in \mathbb{O}_B} \sum_{\mathbf{j} \in \mathcal{N}(\mathbf{i}, \mathbb{O}_B)} g([f]_{\mathbf{i}}, [f]_{\mathbf{j}}) \quad (2.30)$$

$$= \sum_{\mathbf{i} \in \mathbb{O}} \sum_{\mathbf{j} \in \mathcal{N}(\mathbf{i}, \mathbb{O})} g([f]_{\mathbf{i}}, [f]_{\mathbf{j}}) \quad (2.31)$$

$$- \left( \sum_{\mathbf{i} \in \mathbb{O}_O} \sum_{\mathbf{j} \in \mathcal{N}(\mathbf{i}, \mathbb{O}_B)} g([f]_{\mathbf{i}}, [f]_{\mathbf{j}}) + \sum_{\mathbf{i} \in \mathbb{O}_B} \sum_{\mathbf{j} \in \mathcal{N}(\mathbf{i}, \mathbb{O}_O)} g([f]_{\mathbf{i}}, [f]_{\mathbf{j}}) \right). \quad (2.32)$$

Figure 2.7.: Example run of the graph cut approach for a  $2 \times 2$  image.

Therefore, the original energy can be maximized by minimizing the sum of neighbor similarities across the boundary between both regions:

$$E_{\text{graph cut}}^*(\mathbb{O}_O, \mathbb{O}_B) := \sum_{\mathbf{i} \in \mathbb{O}_O} \sum_{\mathbf{j} \in \mathcal{N}(\mathbf{i}, \mathbb{O}_B)} g([f]_{\mathbf{i}}, [f]_{\mathbf{j}}) + \sum_{\mathbf{i} \in \mathbb{O}_B} \sum_{\mathbf{j} \in \mathcal{N}(\mathbf{i}, \mathbb{O}_O)} g([f]_{\mathbf{i}}, [f]_{\mathbf{j}}). \quad (2.33)$$

This energy can be minimized by constructing a graph representation of the image and computing the minimum cut of the resulting graph. To this end, it requires also annotations by the user that serve the purpose of providing example locations in  $\mathbb{O}_O$  and  $\mathbb{O}_B$ . An example run of the algorithm is illustrated in Figure 2.7.

Roughly speaking, the graph is constructed as follows: first, a node for each point  $\mathbf{i} \in \mathbb{O}$  is created. Additionally, two terminal nodes are present: one representing the object region and the one referring to the background region. Non-terminal nodes are connected to other nodes via an edge that correspond to the direct neighbors of the original points. These edges are assigned a weight that represents the gray similarity between the corresponding image points. Finally, nodes that have been annotated by the user are connected to one of the two terminal nodes.

In order to obtain the wanted segmentation, the minimum cut of the graph is computed. A cut of a graph is a partition into two sets: one set is connected only to the object terminal via some path, and the other set to the background terminal. A cut is performed by removing a suitable set of edges from the graph. The minimum cut of a graph is the cut where the sum of participating edge weights is minimal among all possible cuts. According to Equation (2.33), the minimal cut is the wanted minimizer.

Thus, all nodes are now either connected to the object terminal node or the background one, which can be used to derive the segmentation.

Example results for 2D images of MRI slices can be inspected in Figure 2.8. In both cases, the strategy provides acceptable results. In contrast to other methods, the graph cut approach offers the option to select local object regions: the identified object region in the example results consists largely of the tongue region. Although only 2D results are shown, it is also possible to apply the strategy to 3D data. In this case, annotating the image might become more demanding. Moreover, this requirement for annotations makes the approach suboptimal for an unsupervised process. In this regard, it is also important to store the annotations in a suitable format in order to make the process reproducible.

2. *Extracting articulator shape information from MRI data*

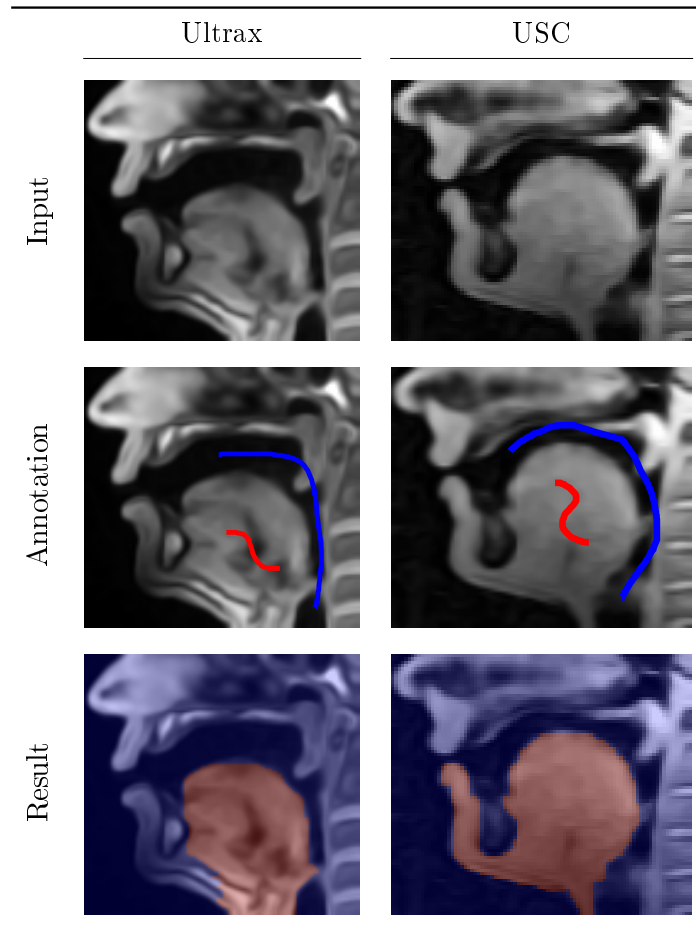


Figure 2.8.: Example annotations and corresponding graph cut results for 2D images of MRI slices.



### 2.6.5. Discussion

In this section, several segmentation techniques were investigated. Like in the case of removing noise from the scan data, it is apparent that the choice for the best segmentation technique depends on the used dataset: if the dataset is small, methods like the graph cut approach may be used to annotate each scan individually. In cases where vignetting effects in the scans are negligible, Otsu's method or its unweighted variant may be used to automatically segment the dataset. The Chan-Vese approach is a good choice if the same parameters work well for each scan of the underlying dataset. The same observation holds true for basic thresholding: if the same threshold can be used for each scan, then this strategy is well-suited for segmenting the data in a minimally supervised way. Thus, it is important to inspect the data and select an appropriate segmentation strategy.

In the remainder of this work, the unweighted variant of Otsu's method is used for the Ultrax and Baker datasets because it led to satisfying results. For the USC data, basic thresholding posed a reliable choice.

## 2.7. Estimating the shape

The previous step provides access to the spatial support of the tissue region in the MRI scan. Now the question arises how this information can be used to estimate the shape of the tongue, for example.

### 2.7.1. Surface point extraction

The obtained partition  $\Omega = \Omega_O \cup \Omega_B$  can be used to extract surface information of the tissue region. Here, the set of surface locations may be defined by means of the discrete image as follows:

$$P := \{ \omega(\mathbf{i}) \mid \mathbf{i} \in \mathbb{O}_O \wedge \mathcal{N}(\mathbf{i}, \mathbb{O}_B) \neq \emptyset \}. \quad (2.34)$$

This means that a location  $\omega(\mathbf{i})$  is classified as a surface location of the tissue region if it belongs to the region  $\Omega_O$  and at least one of its direct neighbors is located in  $\Omega_B$ .

The obtained partition can also be used to derive surface normals at these points. To this end, a new binary image  $g : \Omega \rightarrow [0, 255]$  is constructed with

$$g(\mathbf{x}) = \begin{cases} 255 & \mathbf{x} \in \Omega_O, \\ 0 & \mathbf{x} \in \Omega_B. \end{cases} \quad (2.35)$$

Now, the structure tensors are computed at the surface points  $\mathbf{p} \in P$ :

$$J_{\sigma, \rho}(\mathbf{p}) = K_\rho * \left[ \nabla g_\sigma(\mathbf{p}) \nabla g_\sigma(\mathbf{p})^\top \right]. \quad (2.36)$$

Finally, the eigenvector corresponding to the largest eigenvalue of this structure tensor is used as the surface normal. By default, these normals are modified in such a way that they point to the outside of the tissue region.

The result of this whole process is a point cloud where examples can be seen in Figure 2.9. This basically means that the image representation of the scan was turned into

## 2. Extracting articulator shape information from MRI data

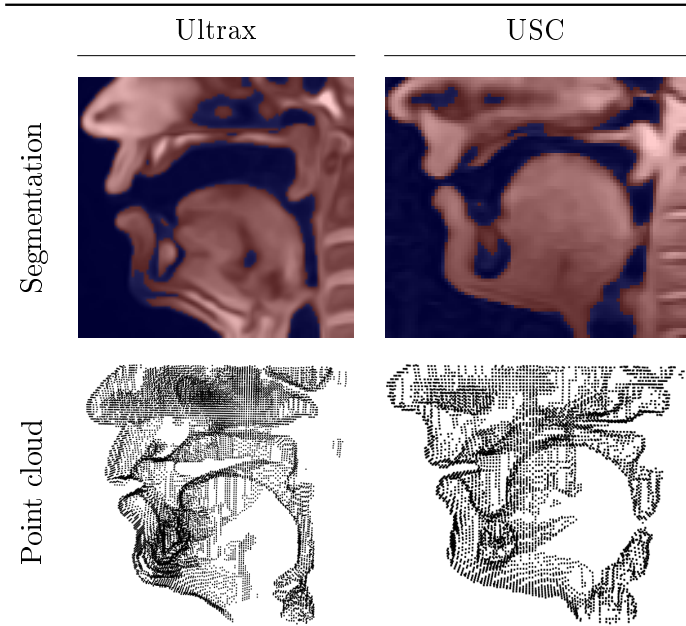


Figure 2.9.: Example point clouds obtained from segmentations. The clouds were clipped and decimated in order to improve visibility.

a fully geometric one such that a second-order surface approximation is now available for the tissue region. A purely geometric representation offers the advantage that is easy to add information: for example, missing information in one point cloud may be reconstructed by adding the corresponding points from a different scan where the needed information is present.

### 2.7.2. Observations

The previous steps turned a 3D image representation of an MRI scan into a geometric representation consisting of the surface points of the tissue region. However, this extracted surface information is currently ill-suited for either representing the shape of an articulator or performing a shape analysis. This is due to the fact that  $P$  is only a loose collection of points. On the one hand, such point clouds lack any coherent surface information because there are gaps between the individual points due to the discrete nature of the original scans. Additionally, huge holes might also be present because one tissue region might touch another, which causes the surface points to be missing in the resulting point cloud. On the other hand, the cloud contains much more information than just the desired object that should be analyzed. Hence, the following tasks should be addressed: first, the unknown subset of points in  $P$  implicitly representing the shape of the wanted articulator has to be identified. Second, this subset has to be used to derive a shape representation of the corresponding articulator.

Like stated earlier, a polygon mesh  $M := (V, F)$  is a useful shape representation. The

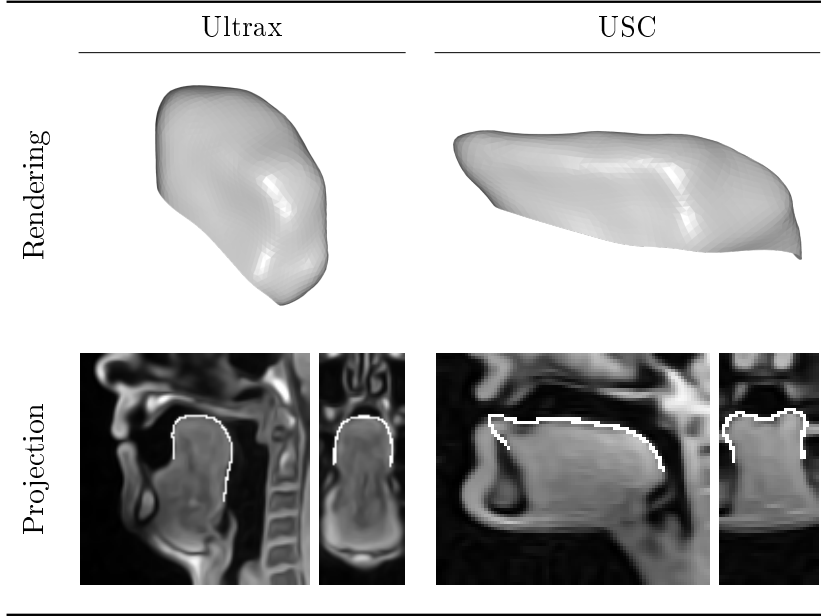


Figure 2.10.: Standard visualization of meshes and projection onto corresponding scans.

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 the mesh. A face  $f \subset F$  is an ordered list of vertices. Connecting the vertices in this list with lines in the order in that they occur and adding a line between the first and the last element creates a polygon. Thus, a face can be seen as a surface patch, such that the whole surface representation is obtained by stitching those surface patches together.

The above representation is called a face-vertex mesh. This data structure can be used to derive the edges that connect the individual vertices to each other. Thus, the edge set  $E(M)$  of a mesh  $M = (V, F)$  may be defined as follows:

$$E(M) = \{(\mathbf{v}_i, \mathbf{v}_j) \mid \mathbf{v}_i \text{ and } \mathbf{v}_j \text{ are directly connected as part of a face } f \in F\}. \quad (2.37)$$

### 2.7.3. Visualizing meshes

The meshes occurring in this work describe 3D objects. This means, a suitable visualization is required for this data structure in order to inspect them on a 2D medium. A standard visualization of meshes can be seen in Figure 2.10 where methods of computer graphics are used to render the 3D objects. Here, the question arises if such a visualization is suitable for evaluating how well an extracted mesh approximates the surface of the corresponding articulator. Such a manual evaluation is needed because in general MRI datasets are lacking a ground truth for the contained articulator shapes. Of course, it is possible to visualize the mesh and the corresponding point cloud in the same image. In this case, the resulting image would be difficult to interpret because surface points of the wanted articulator could be located below the mesh surface and thus be hidden from

## 2. Extracting articulator shape information from MRI data

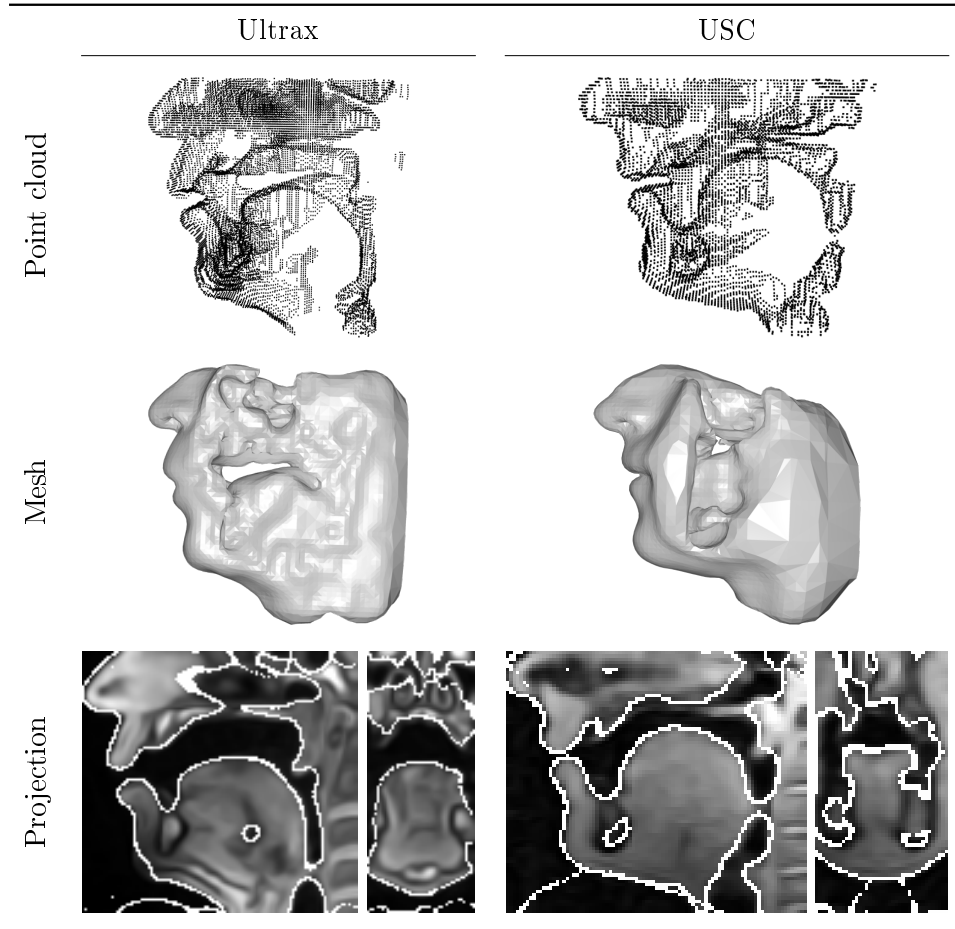


Figure 2.11.: Results of the Poisson reconstruction for given point clouds. Meshes and their projection onto the corresponding scans are shown. Again, the point clouds are modified to improve visibility. The reconstruction itself used the unmodified point clouds.

view.

To this end, another means of visualization is needed that brings the extracted mesh into a clear context with the MRI scan it was extracted from. In this work, the projection of a mesh  $M$  onto an MRI scan is also used as visualization. This means that all voxels of the scan image containing a part of the mesh surface are colored in white. The coordinates of the corresponding voxels are obtained by sampling points on the edges of the mesh. Examples using this type of visualization can be inspected in Figure 2.10.

### 2.7.4. Poisson reconstruction

The Poisson reconstruction (Kazhdan et al., 2006) is a method for estimating a surface mesh purely from data of a given point cloud. Example results of this method are shown

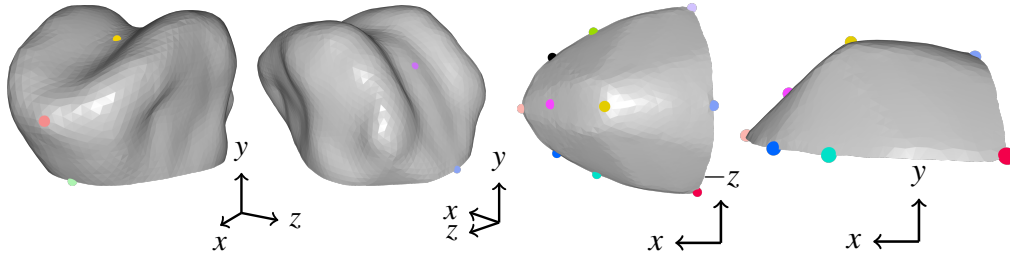


Figure 2.12.: Used templates with landmarks of the tongue (left) and hard palate (right).

in Figure 2.11. The reconstruction was obtained by using the *MeshLab* tool (Cignoni et al., 2008) and applying the default parameters. By inspecting the results, it becomes clear that this method reconstructs a mesh approximating all the surface points contained in the cloud, which poses an issue because only the shape of the wanted articulator should be estimated. Of course, this problem could be solved by manually removing unneeded regions. Another issue is given by the fact that the meshes obtained from this method are missing semantic information. This means that for one reconstruction the vertex  $\mathbf{v}_i$  might belong to the tongue tip while in the reconstruction of another scan this specific vertex might belong to a different region. For analyzing the shape variations, however, this semantic information is needed. It allows for example to analyze how the tongue tip is changing depending on the produced sound.

### 2.7.5. Template matching

#### Motivation

A class of methods that extract a mesh from a collection of points and provide semantic information afterwards are template matching methods. These methods use a single template that is deformed to match the corresponding points. Such an approach can be described as follows: given a template mesh  $M = (V, F)$  that resembles the desired object and a point cloud  $P$ , it finds a set  $A := \{A_i\}$  where  $A_i : \mathbb{R}^3 \rightarrow \mathbb{R}^3$  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  $P$ . Thus, a template matching approach provided with a template shaped like a tongue, for example, tries to identify the subset of points resembling a tongue-like shape and deforms the provided template accordingly to match these points. Template matching strategies can also make use of so-called landmark information: in this case, a user provides desired correspondences between a few mesh vertices  $\mathbf{v}_i$  and locations  $\mathbf{p}_i \in \Omega$ , which help to guide the template matching approach.

#### Template meshes

In this work, two articulators are of interest: the tongue and the hard palate. Templates and used landmarks for these two parts of the vocal tract are shown in Figure 2.12. Both templates were extracted from MRI data by means of medical imaging software (Rosset

## 2. Extracting articulator shape information from MRI data

et al., 2004). Afterwards, the templates were made symmetric to remove this particular bias towards the original speaker by mirroring the respective mesh at a selected center plane.

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. Here, the tongue template is lacking 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.

### Used approach

This work uses the template matching approach by Wuhrer et al. (2015). In this strategy, the following type of rigid body motion is used:

$$A_i := M_{\text{translate}}(\mathbf{v}_i)M_{\text{rotate}}(\mathbf{a}_i, \xi_i)M_{\text{translate}}(\mathbf{t}_i)M_{\text{translate}}(-\mathbf{v}_i). \quad (2.38)$$

Thus, first the vertex is mapped to a local coordinate system centered at  $\mathbf{v}_i$  by  $M_{\text{translate}}(\mathbf{v}_i)$ . Afterwards, it is translated by the vector  $\mathbf{t}_i$ . The next transformation is given by  $M_{\text{rotate}}(\mathbf{a}_i, \xi_i)$  that rotates the result around the axis  $\mathbf{a}_i$  by the angle  $\xi_i$ . Finally, it is translated back into the global coordinate system. Using an axis-angle representation of the rotation serves the purpose of making the rotation angle comparable among neighboring vertices.

Therefore, for each vertex  $\mathbf{v}_i$ , the translation vector  $\mathbf{t}_i$ , the rotation axis  $\mathbf{a}_i$ , and the associated rotation angle  $\xi_i$  have to be estimated by the template matching approach. Here, this is accomplished by minimizing the following energy:

$$E_{\text{def}}(A) = \alpha E_{\text{data}}(A) + \beta E_{\text{smooth}}(A) + \gamma E_{\text{landmark}}(A). \quad (2.39)$$

The data term

$$E_{\text{data}}(A) := \frac{1}{|V^L|} \sum_{\mathbf{v}_i \in V^L} \left\| A_i(\mathbf{v}_i) - \arg \min_{\mathbf{p}_j \in P} \|A_i(\mathbf{v}_i) - \mathbf{p}_j\| \right\|^2 \quad (2.40)$$

measures the distance between the deformed vertices  $A_i(\mathbf{v}_i)$  and their nearest neighbors  $\mathbf{p}_j$  in the point cloud  $P$ . 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. Excluding landmark vertices from the data term serves the purpose of preserving the user-wanted correspondences. This term is weighted by  $\alpha > 0$ .

The smoothness term with a weight  $\beta > 0$

$$E_{\text{smooth}}(A) := \frac{1}{|V|} \sum_{v_i \in V} \left( \frac{1}{|\mathcal{N}_2(v_i)|} \sum_{v_j \in \mathcal{N}_2(v_i)} \|A_i - A_j\|^2 \right) \quad (2.41)$$

evaluates the differences between the rigid body motion  $A_i$  at vertex  $\mathbf{v}_i$  and the motions  $A_j$  in its geodesic neighborhood  $\mathcal{N}_2(\mathbf{v}_i)$ . This neighborhood is defined as:

$$\mathcal{N}_2(\mathbf{v}) := \{ \mathbf{w} \mid \text{shortestPath}(\mathbf{v}, \mathbf{w}) \leq 2 \text{res}(M) \}. \quad (2.42)$$

Roughly speaking, it consists of vertices  $\mathbf{w}$  whose distance to  $\mathbf{v}$  along the edges of the mesh is smaller or equal to  $2 \text{res}(M)$ . Here,  $\text{res}(M)$  is the resolution of the mesh  $M$ , which corresponds to the average edge length. On the whole, this means that the smoothness term penalizes deformations that alter the original shape of the template. It is important to note that the smoothness term only measures the differences among the translation vectors  $\mathbf{t}_i$  and rotation angles  $\xi_i$ .

Finally, the landmark term

$$E_{\text{landmark}}(A) := \frac{1}{|L|} \sum_{(\mathbf{v}_i, \mathbf{p}_i) \in L} \|A_i(\mathbf{v}_i) - \mathbf{p}_i\|^2 \quad (2.43)$$

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. These landmarks are selected manually and therefore might be missing from the actual generated point cloud. This term is weighted by  $\gamma \geq 0$ .

All terms are normalized with respect to the amount of vertices that are participating in the respective term. This serves the purpose of making the terms comparable to each other. As a convention, the weight  $\alpha$  is always set 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. In the beginning of the optimization, the individual rotation axes are initialized to the normals at the corresponding vertices. The remaining components are chosen in such a way that they describe the identity transformation.

As the energy in (2.39) is not differentiable due to the data term, it is usually optimized by minimizing a series of energies  $E_{\text{def}}^t(A^t)$  where  $t \in [1, t_{\text{max}}]$ . Each energy uses adapted weights  $\beta^t$  and  $\gamma^t$ :

$$\beta^t = \beta - (t - 1) \frac{\beta - \beta_{\text{min}}}{t_{\text{max}} - 1}, \quad (2.44)$$

$$\gamma^t = \gamma - (t - 1) \frac{\gamma - \gamma_{\text{min}}}{t_{\text{max}} - 1}, \quad (2.45)$$

where  $\beta_{\text{min}}$  and  $\gamma_{\text{min}}$  are set by the user. First, large weights for the smoothness term and the landmark term cause the template to move to the desired location. In each iteration, they may become smaller, which allows the approach to adapt the template more and more to local structures in the point cloud. For finding the nearest neighbors in the current energy, the transformations found as minimizer of the previous energy are used. As transformations for the first energy, the identity transformation is applied. In the following, the amount of energies used is referred to as the amount of optimization steps. For optimizing such energies, a quasi-Newton approach is used (D. C. Liu and Nocedal, 1989). The nearest neighbors are estimated with the ANN library (Mount and Arya, 2010).

### Modified nearest neighbor heuristic

Originally, a standard heuristic (Allen et al., 2003; H. Li et al., 2009) was used by the approach to distinguish valid data observations from invalid ones in the optimization of

## 2. Extracting articulator shape information from MRI data

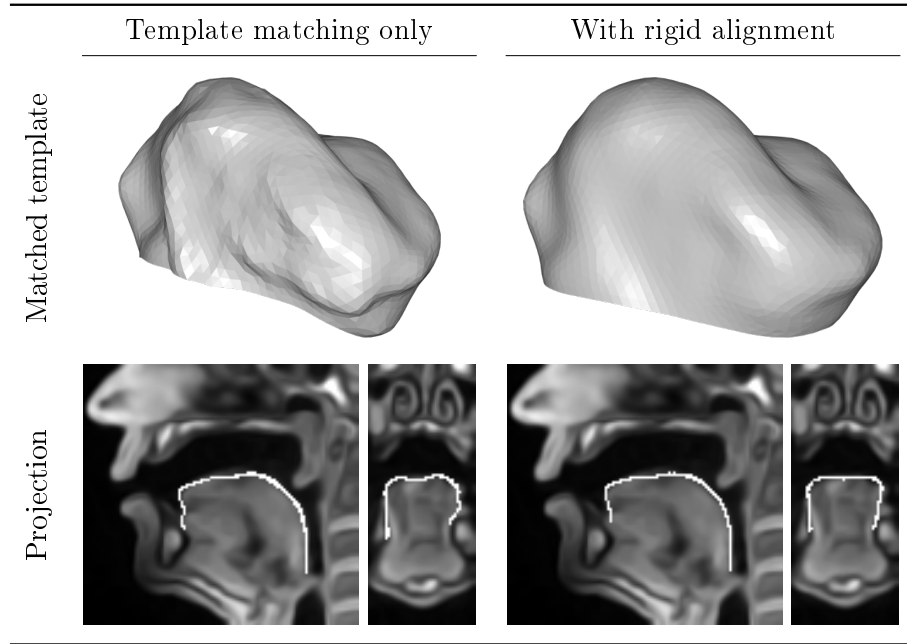


Figure 2.13.: Impact of using result of rigid alignment as initialization for template matching.

$E_{\text{data}}$ . In particular,  $\mathbf{p}$  is assumed to be a valid data point candidate for a deformed vertex  $A_i(\mathbf{v}_i)$  if the Euclidean distance between both is below a given threshold and if the orientation of their normals is similar. For evaluating the similarity of the normals, the angle between them is computed. If the angle is above a prescribed threshold  $\theta$ , the corresponding nearest neighbor candidate is ignored.

Due to the volumetric nature of the processed data, it might happen that the template matching gets stuck at unrelated points during the optimization. In order to mitigate this issue, the nearest neighbor heuristic may be modified somewhat: now all valid data point candidates are collected within a fixed radius and then the best candidate is selected that lies below the current mesh surface. If such a candidate is missing below the surface, the best one above it will be selected.

### Rigid alignment

Applying the template matching directly to the data and the template mesh may lead to suboptimal results, as it can be seen in Figure 2.13. While the approach manages to move the template towards the tongue surface, the mesh looks very deformed at the side. This may be related to the fact that the template might initially be located outside the corresponding point cloud, which implies that the process starts with a bad initialization. Thus, it is worthwhile to first modify the template in order to have a good initialization. A good initialization means in this case that the position, orientation, and scale are already very close to the present shape of the corresponding object in the point cloud.



For this purpose, a rigid alignment of the template can be performed by minimizing the following energy:

$$E_{\text{rigid}}(A) = E_{\text{data}}(A) + E_{\text{landmark}}(A) \quad (2.46)$$

where the data and landmark term have a similar structure to the template matching case. The transformation is given by:

$$A = M_{\text{translate}}(\mathbf{o})M_{\text{scale}}(\mathbf{x}, s_x)M_{\text{scale}}(\mathbf{y}, s_y)M_{\text{scale}}(\mathbf{z}, s_z) \\ M_{\text{rotate}}(\mathbf{x}, \alpha)M_{\text{rotate}}(\mathbf{y}, \beta)M_{\text{rotate}}(\mathbf{z}, \gamma)M_{\text{translate}}(\mathbf{t})M_{\text{translate}}(-\mathbf{o}). \quad (2.47)$$

In contrast to the template matching, only one global transformation has to be estimated this time. This transformation is then applied to each vertex of the mesh. Moreover, this transformation is slightly different from the rigid body motion that is used for the template matching itself. First of all, the local coordinate system is centered at  $\mathbf{o}$ , the current center of the mesh. Next, instead of using a custom rotation axis  $\mathbf{a}$ , it rotates around the principal axes  $\mathbf{x}$ ,  $\mathbf{y}$ , and  $\mathbf{z}$ . In terms of optimization, this reduces the components to be found from 4 (rotation axis and rotation angle) to 3 (three rotation angles) and removes the requirement of providing an initialization for the rotation axis. Moreover, this transformation consists of a scaling. In particular, a non-uniform scaling is used in order to account for the fact that an articulator might differ in scale across the different axes. However, using this type of scaling has a disadvantage: if the current nearest neighbors in the energy optimization fail to describe a volume, the approach will be missing enough information to properly optimize  $s_x$ ,  $s_y$ , and  $s_z$ . In particular, multiple minima may exist in such cases. An example of this problem can be seen in Figure 2.14. As the template may be far away from the point cloud, only the landmarks are used to estimate the transformation parameters at the beginning of the optimization. However, these landmarks may only describe a plane in the case of the tongue, which causes the template to nearly collapse to a 2D object. In fact, a full collapse is avoided by a constraint that enforces  $s_x$ ,  $s_y$ , and  $s_z$  to remain positive during the optimization. As the sides of the template surface are now too far away from points in the cloud, the nearest neighbor heuristic fails to find any suitable target points for the corresponding vertices on the mesh.

In order to avoid such situations, the optimization may be performed in two steps: first, only the orientation and the position are optimized by using the landmark information, which causes the template to be near data points including the sides of the mesh surface. The second step uses both, the landmark information and the point cloud in the optimization. This time, the scaling components are also estimated. The result in Figure 2.14 shows that this two step strategy helps to stabilize the rigid alignment: the collapsing is avoided and the template is properly aligned.

Finally, Figure 2.13 shows that using a rigid alignment before starting the template matching leads to the desired result: the template is close to the data without showing the artifacts that were observed before.

2. *Extracting articulator shape information from MRI data*

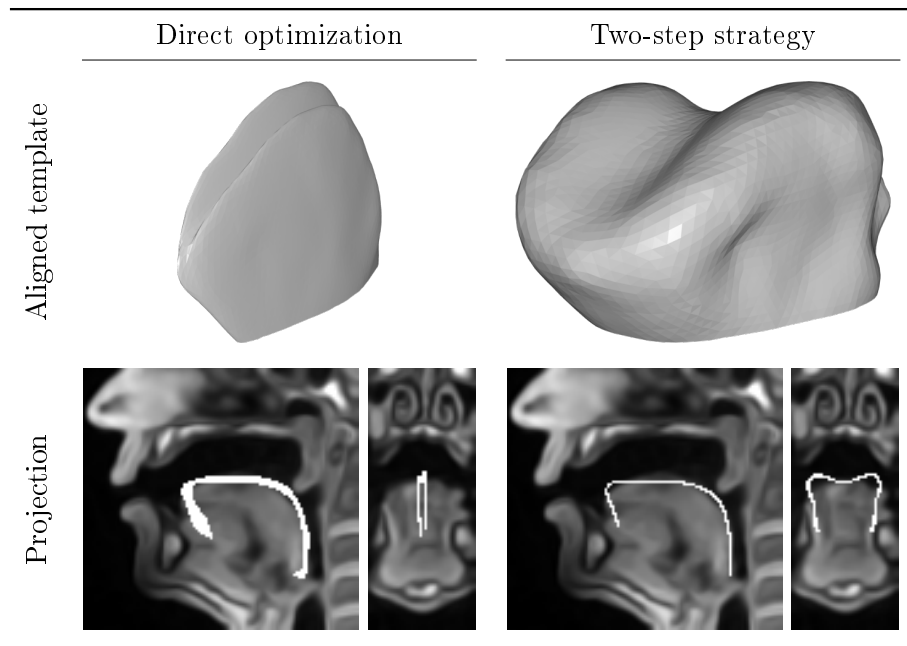


Figure 2.14.: Comparison between the direct optimization of the rigid alignment energy and the two-step approach. Meshes and their projection onto the corresponding scans are shown.

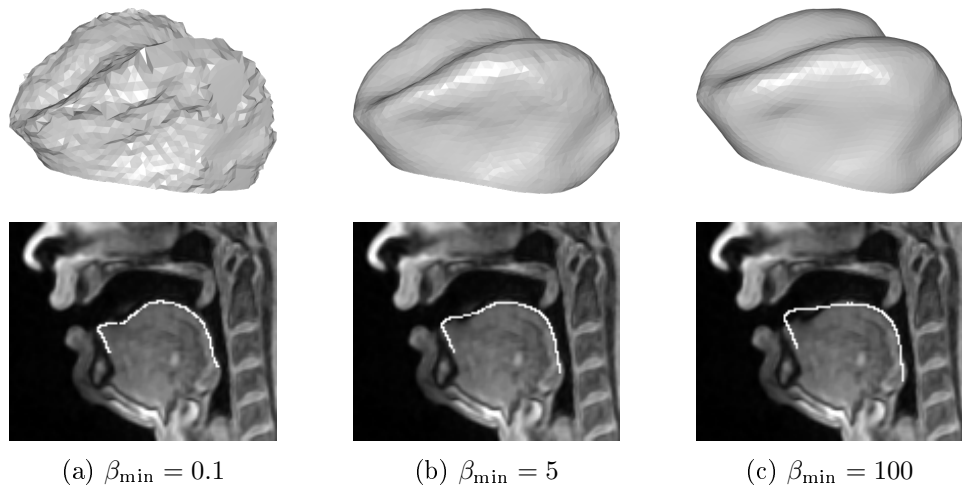


Figure 2.15.: Effect of  $\beta_{\min}$  on the resulting mesh. A low value (a) leads to an overfitting of the data and a very noisy mesh, whereas a high value causes underfitting and produces a very smooth result (c). Choosing an appropriate value provides a good compromise between data nearness and mesh quality (b).

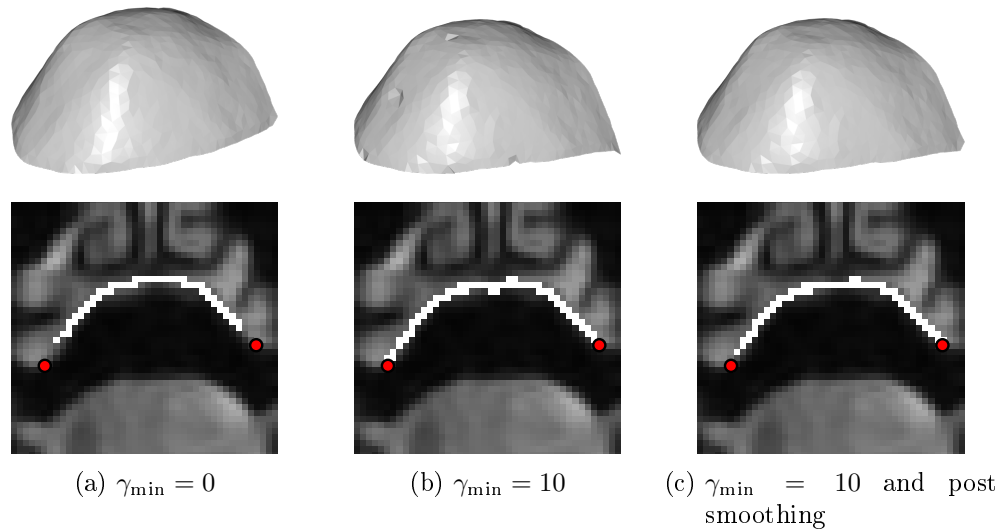


Figure 2.16.: Effect of  $\gamma_{\min}$  on the resulting mesh. Setting it to 0 prevents the template matching from reaching the provided landmarks shown as red dots (a). Using the value 10 aligns the template to the wanted positions, but leads to spike-like artifacts (b). Applying a smoothing afterwards removes these spikes while keeping the template close to the landmarks (c).

### Effect of weights

The weights of the template matching approach 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. Figure 2.15 shows results for different values of  $\beta_{\min}$ . Here, it becomes apparent why it makes sense to also inspect the rendering of the mesh: while the projection of the result shows the data nearness, it fails to reveal the overfitting artifacts on the mesh surface.

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 Figure 2.16, the effect of  $\gamma_{\min}$  on the mesh can be inspected.

### Postprocessing the obtained meshes

In order to mitigate such effects of wrongly placed landmarks, a smoothing operation is applied after the template matching: the measured rigid body motion  $A_i$  is replaced at the corresponding vertex with the average of the rigid body motions in the neighborhood around this vertex. Figure 2.16 shows how this filtering can help to remove spike-like artifacts. It is important to avoid noise or artifacts on the mesh because otherwise this

## 2. Extracting articulator shape information from MRI data

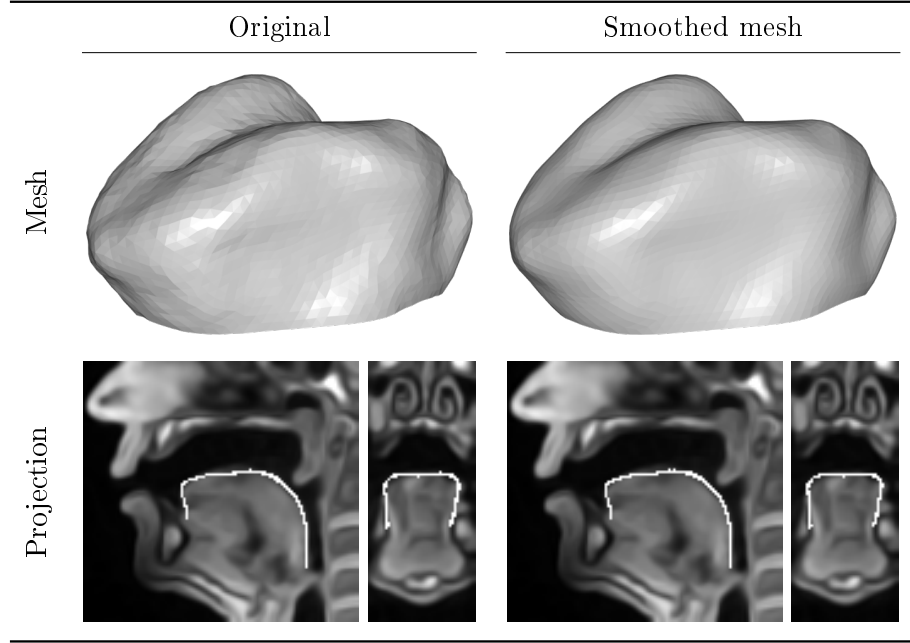


Figure 2.17.: Smoothing a mesh in order to remove high frequency noise.

might be incorporated into a statistical shape model.

So far, only a local smoothing was applied at the landmarks positions. In Figure 2.17, it can be seen that results of the template meshing may be degraded due to global high frequency noise. Thus, it is worthwhile to apply a second post processing step to the matched template:

$$\mathbf{v}^{n+1} = \begin{cases} \frac{1}{|\mathcal{N}_1(\mathbf{v}^n)| + 1} \left( \sum_{\mathbf{w}^n \in \mathcal{N}_1(\mathbf{v}^n)} (\mathbf{w}^n) + \mathbf{v}^n \right) & \text{for inner vertex} \\ \frac{1}{|\mathcal{N}_1(\mathbf{v}^n) \cap \partial V| + 1} \left( \sum_{\mathbf{w}^n \in \mathcal{N}_1(\mathbf{v}^n) \cap \partial V} (\mathbf{w}^n) + \mathbf{v}^n \right) & \text{for boundary vertex.} \end{cases} \quad (2.48)$$

This post processing step is a variant of Laplacian smoothing (Herrmann, 1976; Sorkine et al., 2004; Hansen et al., 2005) and iteratively modifies the vertex positions by averaging the positional data in the 1-ring neighborhood. The 1-ring neighbors of a vertex  $\mathbf{v}$  of a mesh are defined as follows:

$$\mathcal{N}_1(\mathbf{v}) := \{ \mathbf{w} \mid (\mathbf{v}, \mathbf{w}) \in E \}. \quad (2.49)$$

Basically, these are the vertices that are directly connected to the corresponding vertex  $\mathbf{v}$  via an edge. The initial values  $\mathbf{v}^0$  of this step correspond to the vertices of the mesh

Surface-enhancing diffusion		
$\sigma$	standard deviation for presmoothing the image	1
$\rho$	standard deviation for combining structural information	1
$\lambda$	contrast parameter of Perona-Malik diffusivity	0.1
$t$	evolution time	2.4
Point cloud extraction		
$\sigma$	standard deviation for presmoothing the image	1
$\rho$	standard deviation for combining structural information	1
Template matching		
$\beta_{\max}$	initial weight for smoothness term	10
$\beta_{\min}$	weight for smoothness term on last iteration	6
$\gamma_{\max}$	initial weight for landmark term	0.1
$\gamma_{\min}$	weight for landmark term on last iteration	0
$n_{\text{match}}$	amount of optimization steps	100
$r_{\text{search}}$	search radius for nearest neighbor heuristic	5 mm
$\theta_{\text{search}}$	angle threshold for nearest neighbor heuristic	60 °
$n_{\text{postsmooth}}$	post smoothing iterations	1

Table 2.1.: Used settings for the experiments. Parameter name, description, and value are provided.

after the landmark smoothing has been applied. Here, a special treatment is given to the vertices  $\mathbf{v}$  that are part of the mesh boundary  $\partial V$ . The boundary of a mesh may be defined as the set of vertices that are part of an edge that is generated by exactly one face. In this case, only neighbors are used in the averaging that are also part of the boundary, which avoids a shrinking of the boundary. Results can be inspected in Figure 2.17. On the whole, this smoothing removes high frequency noise from the resulting mesh.

## 2.8. Experiments

In the previous sections, a framework was derived for extracting articulator shape information from given MRI scans. It is now of interest to evaluate this framework and investigate to which degree it is able to estimate the wanted shapes. To this end, it was applied to the Baker, Ultrax, and USC datasets in order to extract the tongue shape from the individual scans.

### 2.8.1. Settings

In order to assess if the approach is independent of speaker and produced phone, all scans were processed by using the same settings, which are summarized in Table 2.1. These settings were manually selected in order to obtain acceptable results, which implies that it is possible that parameter combinations exist that improve the results. Furthermore, the required landmarks for the tongue were distributed on the corresponding scans by a user who is not an anatomical expert.

## 2. *Extracting articulator shape information from MRI data*

In order to segment the USC data, thresholding was used. The corresponding parameter was manually selected by visually inspecting the scan data. For the Ultrax and the Baker datasets, the unweighted variant of Otsu’s method was used as segmentation technique.

### 2.8.2. Evaluation

Like previously stated, the used MRI datasets are missing a ground truth solution of the correct shape of the articulators. This is due to the fact that these datasets consist of real-world data. The lack of a ground truth makes a quantitative analysis of the results very hard. There exists the possibility of manually annotating the MRI scans to create a reference solution. Again, it should be stressed that 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, a qualitative analysis was undertaken: the results were inspected manually as a post-hoc analysis in order to assess their quality. In particular, the projections of the registered tongue meshes were used to check if the mesh surface was close to the true tongue contour and if they show any anomaly.

### 2.8.3. Results

At first sight, the approach seems to be promising: several results that can be inspected in Figures 2.18 to 2.20 appear to be very close to the shape of the tongue. However, the approach fails in several cases. Some of those are shown in Figure 2.21.

The results where the approach succeeded provide an interesting observation: despite producing the same phone, the speakers might apply a different articulatory strategy for this purpose, which might be related to the specific anatomy of the speaker or to personal preference. An example is the phone [a] in the results for the Ultrax data in Figure 2.20. All three speakers use a different vocal tract configuration for the same phone. This observation justifies the original plan to derive a tongue model where anatomy and tongue pose parameters are separated from each other.

The following reasons for the bad performance of the strategy may be identified: first, the method leads to suboptimal results if parts of the tongue surface are missing. This can be seen in Figure 2.21a. In particular, the tongue touches the hard palate in this MRI scan, which causes the boundary of the tongue surface to disappear in the corresponding contact area. A more detailed visualization of this situation is shown in Figure 2.22. Thus, it might be necessary to find a means of restoring this missing surface information in order to solve this issue.

Another observation implies that the template matching can also get stuck at locations that are unrelated to the tongue, which is, for example, the case in Figures 2.21b and 2.21d. This issue may be explained by acknowledging the fact that specific structures in the mouth or combinations thereof can actually resemble the used template, which causes the template matching approach to register these structures. Another explanation could be that the search radius selected for the modified nearest neighbor heuristic was

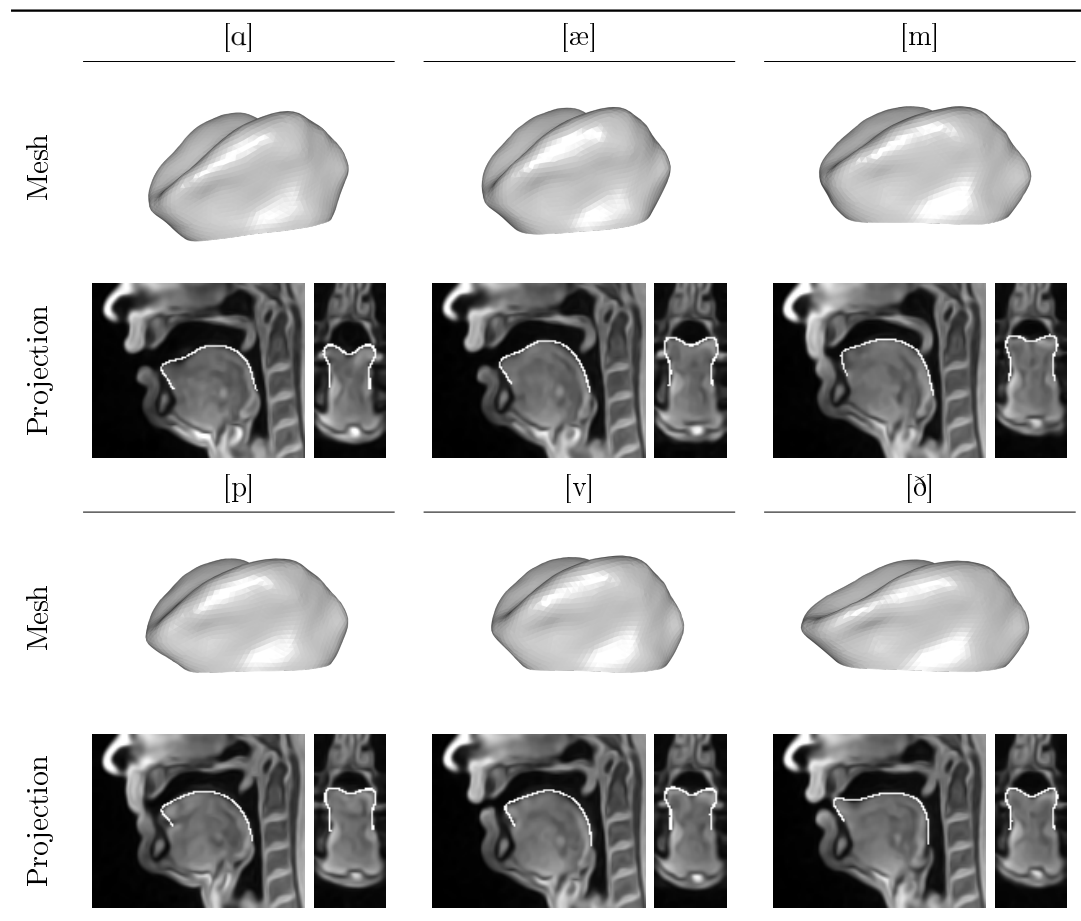


Figure 2.18.: Example results for the Baker dataset where the approach provided acceptable registrations.

## 2. Extracting articulator shape information from MRI data

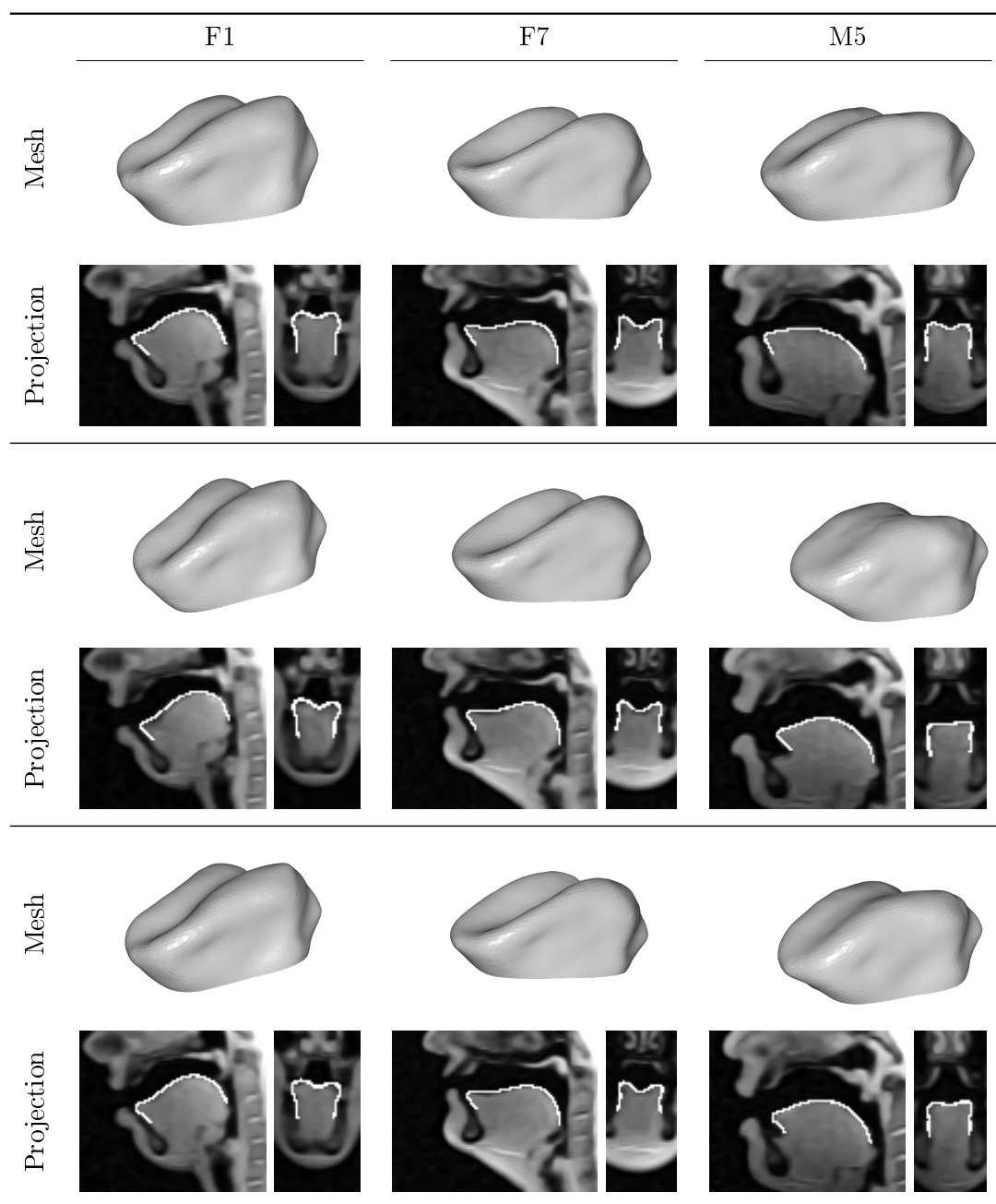


Figure 2.19.: Example results for the phones [ə] (top row), [ɔ:] (center row), and [ɑ:] (bottom row) in the USC dataset. Registrations are shown for the speakers F1, F7, and M5 to illustrate different articulation strategies for the same phone.



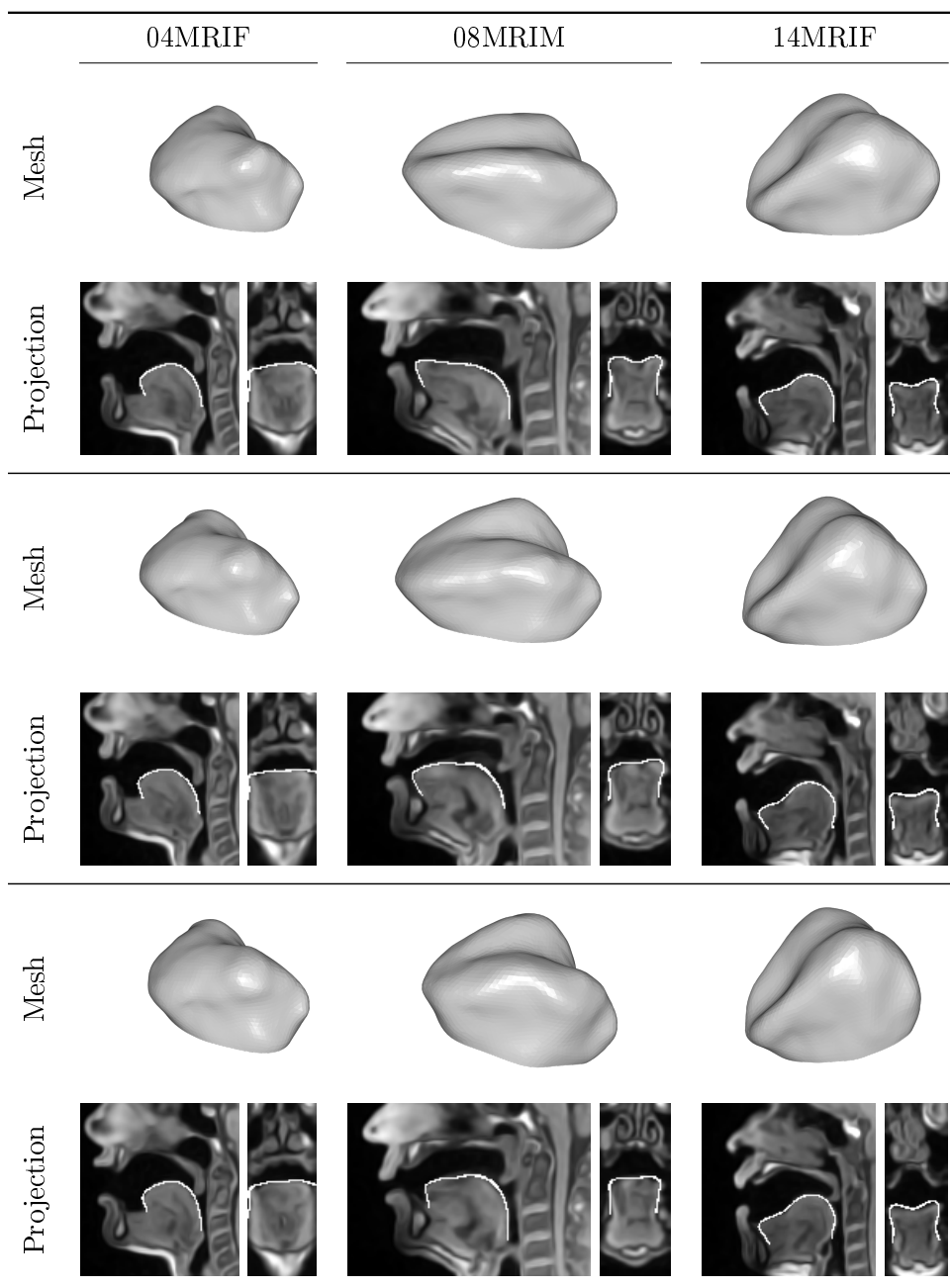


Figure 2.20.: Example results for the phones [a] (top row), [ʌ] (center row), and [ɔ] (bottom row) in the Ultrax dataset. Registrations are shown for the speakers 04MRIF, 08MRIM, and 14MRIF.

2. *Extracting articulator shape information from MRI data*

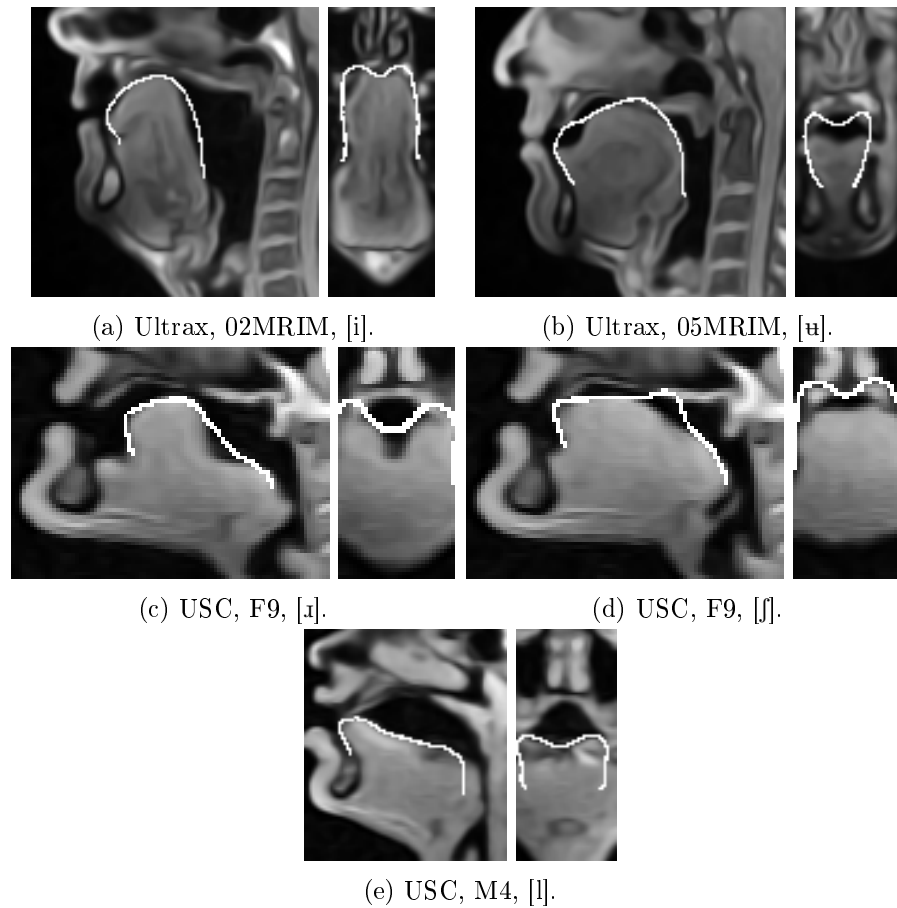


Figure 2.21.: Examples scans where the proposed approach fails to register the tongue surface properly. Captions provide information about dataset, speaker name, and produced phone.

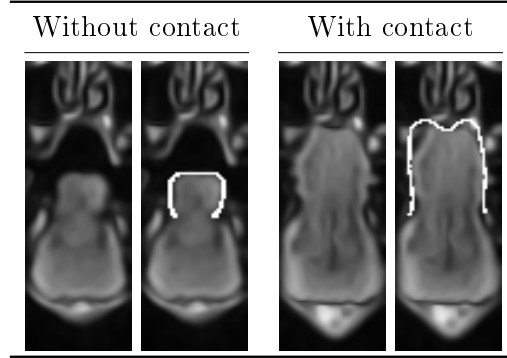


Figure 2.22.: A palatal contact causes the tongue surface to become indistinguishable from surrounding tissue in the corresponding area, which causes the approach to fail. Without contact, the approach produces an acceptable registration. Both scans belong to the same speaker. Figures show scan with and without projection of result mesh.

too low. Several ideas on how to address this issue come to mind: increasing the search radius of the heuristic could help, but would also increase the running time of the approach. Another solution consists of using more landmarks in order to force the template matching to move the mesh to the correct position. However, this would increase the annotation time and also put more burden on the user. As an alternative, the issue could be solved by automatically detecting and removing these structures from the point cloud.

In other cases that are shown in Figures 2.21c and 2.21e, the approach only registered a part of the tongue surface. At first sight, this behavior appears to be abnormal because the needed surface information is present. At second sight, however, the performance may be explained as follows: the template matching only uses one specific instance of all possible tongue shapes to register the point cloud. This means that if the shape to be registered differs too much from the used template, the smoothness assumption of the template matching keeps the approach from estimating the shape correctly. Reducing the importance of the smoothness assumption could solve this issue, but could lead to overfitting effects. A more promising remedy would be using a template that is more similar to the data to be registered.

Additionally, it is currently unclear if the applied subjective evaluation of the results may be considered as valid. In this context, it is important to consult experts in the field.

Due to the above issues of the approach, it is omitted here to evaluate if the proposed framework is speaker and tongue pose independent. This evaluation is postponed until Chapter 4 where the issues of the approach are addressed.

## 2.9. Conclusion

In this chapter, a basic framework was built for extracting articulator meshes from provided MRI scans in a semi-supervised way. In particular, image processing techniques involving denoising and segmentation methods were combined with a template matching approach to achieve this goal. Here, the image processing methods can be chosen by the user to adapt the approach to the dataset that should be processed. For evaluating its performance, the derived framework was used to estimate the tongue shape from scans of different datasets. During this evaluation, the following observations were made: the proposed strategy can already estimate tongue shapes from data, but it fails to find a proper registration in certain cases. Thus, it might be worthwhile to improve the strategy in order to gain access to these missing shapes because they might provide information about important DoFs of the tongue.

In summary, the following modifications are needed:

- Reconstruct missing surface information in a palatal contact area
- Remove unrelated structures from the point cloud
- Automatically provide adapted templates for the template matching approach

Furthermore, the subjective evaluation of the obtained meshes has to be validated.

## 3. Background on statistical shape analysis

### 3.1. Introduction

The previous chapter presented an initial approach for estimating the tongue shape from speech related magnetic resonance imaging (MRI) data, which provided access to tongue meshes. It is now of interest to analyze the obtained meshes in order to derive generative shape models. Such a model uses parameters that describe the shape that should be generated. The main goal of this work is to acquire information about the degrees of freedom (DoF) of the tongue shape during the production of speech. Thus, the analysis should discover the DoF present in the data and allow these detected DoF to be mapped to the parameters of the model. Ideally, the parameter set should be small and therefore only the most important DoF should be represented by the model. Furthermore, the model should also allow to estimate the plausibility of a created shape. Basically, these properties imply that a simplification of the acquired data is desired that also preserves the structure of the data.

To this end, two methods are presented: the linear approach based on principal component analysis (PCA) and the multilinear one based on tensor decomposition. This chapter is largely based on Bolkart (2016, Sections 3.2, 3.4, and 4.2) and accordingly adapts the corresponding notation.

### 3.2. Vector representation of meshes

Before turning to the actual modeling ideas, it is important to explain how meshes  $M = (V, F)$  can be represented as a vector. As only the vertex positions are changed by the template matching step of the basic approach, the face configuration  $F$  of the mesh remains the same. This means that only the vertex information  $V$  is of interest. The vertex set  $V := \{\mathbf{v}_i\}$  of a mesh that consists of  $n$  vertices may be represented as  $\mathbf{x} := (v_1^x, v_1^y, v_1^z, v_2^x, v_2^y, v_2^z, \dots, v_n^x, v_n^y, v_n^z)^\top$ , where

$$\mathbf{v}_i := \begin{pmatrix} v_i^x \\ v_i^y \\ v_i^z \end{pmatrix}. \quad (3.1)$$

This means that the positional data of a mesh is serialized and stored in a single vector. Basically, this procedure can also be inverted to reconstruct a mesh from such a vector.

### 3.3. Linear modeling

In cases, where the shape differences of acquired data can be attributed to a single source, linear modeling approaches are a suitable choice. Essentially, such models provide only one set of parameters for generating shapes. For example, tongue meshes obtained from a single speaker database can be analyzed by such a method, like data of the Baker dataset. Here, only tongue pose related differences can be expected. Another example would be shape information about the hard palate where only anatomical factors influence its shape.

A standard method for performing this kind of analysis is given by PCA (Dryden and Mardia, 1998, Chapter 5). Given a collection of vector representations of  $k$  tongue meshes  $\{\mathbf{x}_i\}$  that have dimension  $n$ , it basically fits an  $m$ -dimensional hyperellipsoid to the data, such that its orthogonal axes  $\{\mathbf{a}_j\}$  correspond to the most dominant directions of variance. Here, the collection of mesh representations has been processed accordingly, such that only shape differences are present that are of interest. The ellipsoid is centered at  $\bar{\mathbf{x}} = \frac{1}{k} \sum_{i=1}^k \mathbf{x}_i$ . In mathematical terms, the axes are then computed by maximizing  $\sum_{j=1}^m \sum_{i=1}^k ((\mathbf{x}_i - \bar{\mathbf{x}})^\top \mathbf{a}_j)^2$ . These axes are referred to as the principal components or directions of the data. Basically, these principal components can be used to project the original data into a subspace. As the projection is orthogonal, the structure of the data is preserved. Essentially, each point in this subspace can be thought of as a projection of a shape. In fact, the projection can be inverted and a mesh reconstructed from the result. Thus, this type of subspace can be called a shape space. It is important to keep in mind that these directions are purely derived from the data and therefore may make an intuitive interpretation of their meaning difficult. Therefore, an analysis of these components is omitted in this work.

The desired principal components may be obtained as follows. First, a  $k \times n$  matrix  $\mathbf{X}$  is constructed such that its rows correspond to the centered vectors  $\mathbf{x}_i - \bar{\mathbf{x}}$ . In this matrix representation, the rows correspond to the different shapes while the columns refer to the individual coordinates of the mesh vertices. Then,  $\mathbf{X}$  is used to compute the covariance matrix of the data:

$$\mathbf{D} = \frac{1}{k} \mathbf{X}^\top \mathbf{X}. \quad (3.2)$$

Afterwards, the principal components are obtained by computing the eigenvectors associated to the first  $m$  eigenvalues  $\lambda_j$  in descending order of the covariance matrix. An important observation is now that PCA performs a dimensionality reduction, i.e., a simplification of the data, if  $m < n$ . As a convention, this reduction is often called shape space truncation in this work. Of course, such a truncation of the space can lead to information loss. Assuming that the original data is distributed according to a multivariate Gaussian distribution, this information loss can explicitly be computed. In this specific case, the eigenvalues  $\lambda_j$  represent the dispersion or variability of the data along axis  $\mathbf{a}_j$ . Thus, the acquired PCA shape space represents

$$100 \frac{\sum_{j=1}^m \lambda_j}{\sum_{i=1}^n \lambda_i} \% \quad (3.3)$$

of the data. This measure is called the compactness of the shape space.

Again, under the assumption that the data has a multivariate Gaussian distribution, the PCA analysis result can be used to build a generative statistical model. Basically, the parameters  $\mathbf{p} \in \mathbb{R}^m$  representing a shape in the subspace are turned into a vector representation  $\mathbf{y} \in \mathbb{R}^n$  of a mesh as follows:

$$\mathbf{y} = \bar{\mathbf{x}} + \mathbf{U}\mathbf{p}. \quad (3.4)$$

The columns of the matrix  $\mathbf{U} \in \mathbb{R}^{n \times m}$  consist of the computed principal components.

Accordingly, the created vector  $\mathbf{y}$  can be used to construct a mesh. Formally, this can be expressed as a function  $f: \mathbb{R}^m \rightarrow \mathcal{M}$  where  $\mathcal{M}$  is the set of meshes that can be generated by the model. Essentially, all meshes  $M \in \mathcal{M}$  share the same face set and differ only in the positions of their vertices. In the case of the tongue, the parameters  $\mathbf{p}$  can be seen as the DoF of the tongue shape that can be identified in the data and  $\mathcal{M}$  corresponds to the set of tongue meshes.

PCA or variants of it were already used for analyzing the shape of the tongue: Maeda (1990) applied it to investigate two-dimensional (2D) contours of the tongue. Guided PCA (Badin, Elisei, et al., 2008; Fang et al., 2016) was used for analyzing these variations in three-dimensional (3D) meshes. Another linear and related method is given by linear component analysis (LCA) that was used by Engwall (2003). In contrast to PCA, the guided variant of PCA and the LCA impose a meaning on the parameters. This is done, e.g., to make the extracted parameters more interpretable in terms of control (Badin, Elisei, et al., 2008). However, PCA better explains the variance in the data. As this work is concerned with registration of data, PCA was chosen over guided PCA or LCA for analyzing shape differences and creating the associated generative models.

For completeness, it should be noted that the linear modeling approach may also be applied to mesh collections where shape differences are originating from multiple causes. However, the resulting model will only provide one set of parameters in this case.

### 3.4. Multilinear modeling

So far, a linear modeling approach was described for creating a generative model. Basically, the shapes that should be analyzed were arranged in a matrix where rows and columns had a semantic meaning: a row represented a specific shape while a column described a specific vertex coordinate component of the associated mesh. However, this modeling only allowed to extract one set of parameters describing the shape. This may be seen as a consequence for using a data representation that failed to preserve the cause for the shape differences. The reason for a shape difference of the tongue may be attributed to two different causes: in one case, the anatomical features of the tongue changed while in another case, the tongue assumed a different speech related shape. This work aims at deriving tongue shape models that separate the associated parameters into these two sets: the first set, the speaker parameters, should represent the anatomical features of the tongue. The second set, the tongue pose parameters, should describe the speech related tongue pose.

### 3. Background on statistical shape analysis

Thus, another representation of the data is needed that allows to preserve this kind of information. To this end, this section first introduces the concept of a tensor and accordingly provides the required background of tensor algebra. This background is developed so far that is possible to present tensor decomposition methods that might be useful for deriving generative models. Finally, the multilinear tongue model approach is presented. For more information, the reader can turn to De Lathauwer (1997) or Kolda and Bader (2009).

#### 3.4.1. Tensor algebra

Vectors and matrices are means of representing data: vectors are one-dimensional arrays containing data while matrices are 2D arrays. Tensors extend this paradigm to even higher dimensions. Thus, a tensor is an  $n$ -dimensional array where  $n \geq 3$ . In this context,  $n$  is called the order of the tensor. For the sake of simplicity, the following concepts are explained at the example of a tensor of third order, which is the only type of tensor needed in this work.

Formally, such a tensor is represented by  $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ . Its entries can be accessed by using three indices:  $\mathbf{x}_{i_1, i_2, i_3}$  where  $i_j \in \{1, \dots, d_j\}$ . Similar to the matrix case, indices are assigned a role:  $i_1$  is the mode-1 index,  $i_2$  the mode-2 index, and  $i_3$  the mode-3 one.

It is possible to extract the data along one mode from a tensor. Such data is called a mode fiber. Essentially, this extends the concept of rows and columns of the matrix case to the tensor case. A mode- $n$  fiber is obtained by fixing the indices for the other modes and only altering the index of the  $n$ -th mode, which results in a  $d_n$ -dimensional vector. Accordingly, there are  $d_2 d_3$  mode-1 fibers,  $d_1 d_3$  mode-2 fibers, and  $d_1 d_2$  mode-3 fibers.

The notation of fibers can be utilized to turn the tensor into a matrix, which is known as unfolding. The mode- $n$  unfolding leads to a matrix  $\mathbf{X}_{(n)} \in \mathbb{R}^{d_n \times \prod_{k \neq n} d_k}$  where the columns correspond to the mode- $n$  fibers of the original tensor. As the tensor is of third order, three unfolding variants are available. These operations can be reversed in order to turn a matrix with suitable dimensions into a tensor.

The last concept needed is the  $n$ -mode multiplication  $\times_n$  of the tensor  $\mathcal{X}$  with a matrix  $\mathbf{A} \in \mathbb{R}^{m \times d_n}$ . This type of multiplication is defined as follows. First, the tensor  $\mathcal{X}$  is unfolded along mode  $n$  resulting in  $\mathbf{X}_{(n)}$ . The matrix  $\mathbf{X}_{(n)}$  is then multiplied with  $\mathbf{A}$ , which gives  $\mathbf{Y}_{(n)} = \mathbf{A} \mathbf{X}_{(n)}$ . Finally, the matrix  $\mathbf{Y}_{(n)}$  is folded back into a tensor  $\mathcal{Y} = \mathcal{X} \times_n \mathbf{A}$ . For completeness, it should be noted that  $\mathbf{A}$  might also be vector.

#### 3.4.2. Tensor decompositions

The previous section introduced the concept of a tensor that can be used to represent data. It is now of interest to analyze such data in order to gain access to its underlying structure and eliminate unimportant information. In particular, results of such an analysis should also be usable to build a generative model. To this end, tensor decomposition approaches are helpful. In literature, several approaches are available to perform this task.



One example is the canonical polyadic (CP) decomposition (Hitchcock, 1927). It is often called the method of parallel factors (PARAFAC) (Harshman, 1970) or the canonical decomposition (CANDECOMP) (Carroll and Chang, 1970). In the case of the tongue, it was used by several studies (Harshman et al., 1977; Hoole, Wismüller, et al., 2000; Hoole, Zierdt, and Geng, 2003; Ananthakrishnan et al., 2010; Vargas, Badin, and Lamalle, 2012; Vargas, Badin, Ananthakrishnan, et al., 2012; Zheng et al., 2003). Essentially, 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, Wismüller, et al. (2000) found that it might be difficult to find reliable solutions; Vargas, Badin, Ananthakrishnan, et al. (2012) 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 L.-H. 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.

A second example is the Tucker decomposition (Tucker, 1966) that decomposes  $\mathcal{X}$  into a tensor  $\mathcal{C}$  and a product of three matrices  $\mathbf{U}_1$ ,  $\mathbf{U}_2$ , and  $\mathbf{U}_3$ :

$$\mathcal{X} = \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3. \quad (3.5)$$

Here,  $\mathcal{C}$  is called the core tensor of the decomposition. The matrices  $\mathbf{U}_1$ ,  $\mathbf{U}_2$ , and  $\mathbf{U}_3$  are the factor matrices. The columns of these matrices represent basic vectors for the respective modes while the core tensor links them together. Thus, the matrices actually represent subspaces. 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, Badin, and Lamalle, 2012). To avoid the issues with PARAFAC, the Tucker decomposition is chosen to analyze the data in this work.

There exist several ways for computing the Tucker decomposition. In this work, the Tucker2 decomposition is used that essentially sets  $\mathbf{U}_3$  to the identity matrix with suitable dimensions. Furthermore, the factor matrices are enforced to be orthogonal. These matrices are computed by applying the higher-order singular value decomposition (HOSVD) method (De Lathauwer, 1997). For the sake of completeness, it should be noted that other approaches exist for computing these matrices, like the higher-order orthogonal iteration (De Lathauwer, 1997) or the Newton-Grassmann optimization approach (Eldén and Savas, 2009).

The HOSVD may also be seen as an extension of the singular value decomposition to tensors. In order to compute  $\mathbf{U}_n$ , the following steps are performed: first,  $\mathcal{X}$  is unfolded along mode  $n$ , which provides access to the matrix  $\mathbf{X}_{(n)}$ . Afterwards,  $\mathbf{X}_{(n)}$  is decomposed by applying the singular value decomposition in order to obtain the desired  $\mathbf{U}_n$ :  $\mathbf{X}_{(n)} = \mathbf{U}_n \mathbf{\Sigma} \mathbf{V}^\top$ . Like in the PCA case, the matrix  $\mathbf{U}_n$  can be truncated for dimensionality reduction purposes by only considering the singular vectors associated to the first  $m_n$  singular values in descending order, which is called the truncated HOSVD. It is also possible to calculate the compactness of a mode by computing the eigenvalues

### 3. Background on statistical shape analysis

$\lambda_i$  of  $\frac{1}{d_n} \mathbf{X}_{(n)}^\top \mathbf{X}_{(n)}$  and evaluating:

$$100 \frac{\sum_{j=1}^{m_n} \lambda_j}{\sum_{i=1}^{d_n} \lambda_i} \% . \quad (3.6)$$

This quantity can then be used to estimate the information loss caused by an applied truncation of the mode.

#### 3.4.3. Multilinear tongue model

Basically, the above concepts are now used to derive a multilinear tongue model that can be controlled by two sets of parameters: one that modifies the anatomy of the tongue and another one that modifies its pose. Essentially, the same modeling approach is used that was already successfully applied to human faces (Vlasic et al., 2005; Bolkart, 2016).

In a first step, the tensor  $\mathcal{X}$  is constructed. To this end, a collection of tongue meshes is required. Formally, this mesh collection consists of tongue shapes of  $d_1$  different speakers. For each speaker, the shape information of  $d_2$  tongue poses is available where an individual pose is associated to a specific phone that was produced during acquisition. It is important to stress that the recorded phone set has to be the same for each speaker. Furthermore, it is assumed that the mesh collection was aligned accordingly, such that only anatomical and pose related shape differences remain. Like in the PCA case, the meshes are turned into vectors  $\mathbf{x}_i$  and afterwards centered by subtracting the mean  $\bar{\mathbf{x}}$  of the vector collection. Thus, the vector representations have dimension  $n$  like before. Then, they are arranged in a tensor as follows. The data that can be accessed with a fixed mode-1 index  $i_1$  belongs to one specific speaker, the data for a fixed mode-2 index  $i_2$  to one specific tongue pose, and the data for a fixed mode-3 index  $i_3$  to one specific coordinate component of the associated vertex. Accordingly, the first mode may be called the speaker mode of the tensor, the second mode the pose mode, and the third mode may be referred to as the vertex mode. Here, it can be seen that a tensor is a suitable way for representing shape data while still preserving the cause of potential shape differences.

The constructed tensor  $\mathcal{X}$  can then be decomposed by means of the described Tucker2 decomposition with orthogonality constraints:

$$\mathcal{X} \approx \mathcal{T} \times_1 \mathbf{S} \times_2 \mathbf{P}. \quad (3.7)$$

In this equation,  $\mathcal{T} \in \mathbb{R}^{m_s \times m_p \times n}$  is the computed multilinear model. The matrix  $\mathbf{S}$  consists of  $m_s$  columns. This matrix represents the speaker subspace of the model and each row denotes the parameters of a specific speaker in this subspace. A similar observation holds true for the matrix  $\mathbf{P}$  that consists of  $m_p$  columns. Here, each row corresponds to the parameters of a tongue pose in the associated subspace. Accordingly, the matrix  $\mathbf{P}$  represents the pose subspace of the model.

Under certain conditions,  $\mathcal{T}$  can be used to build a statistical generative model. In particular, the data in the speaker and pose subspace have to obey a multivariate Gaussian distribution. In this case, a tongue shape  $\mathbf{y}$  can be represented by the model as follows:

$$\mathbf{y} = \bar{\mathbf{x}} + \mathcal{T} \times_1 \mathbf{s} \times_2 \mathbf{p}. \quad (3.8)$$

The vectors  $\mathbf{s} \in \mathbb{S}$  and  $\mathbf{p} \in \mathbb{P}$  are assumed to describe the properties of the generated shape. The speaker parameters  $\mathbf{s}$  represent the anatomical features of the shape while the pose parameters  $\mathbf{p}$  correspond to the tongue pose. Accordingly, the spaces  $\mathbb{S} := \mathbb{R}^{m_s}$  and  $\mathbb{P} := \mathbb{R}^{m_p}$  represent the speaker parameter space and tongue pose parameter space of the model, respectively. As a convention, the pair  $(\mathbf{s}, \mathbf{p})$  is called the shape space coordinate in the following. Again, the vector  $\mathbf{y}$  can be turned into a mesh. Like in the PCA case, a function  $f : \mathbb{R}^{m_s} \times \mathbb{R}^{m_p} \rightarrow \mathcal{M}$  may be formulated that maps the parameters  $\mathbf{s}$  and  $\mathbf{p}$  to a mesh  $M \in \mathcal{M}$ .

### 3.5. Model fitting

The previous sections presented generative models that are able to create meshes from one set (PCA model) or two sets of parameters (multilinear model). Such models can be used to register data. Given correspondences between mesh vertices and target data points, the parameters of the models can be optimized such that the vertices of the generated mesh are close to the corresponding target points. This procedure may be compared to template matching where a template mesh is deformed to match the data. Using a generative model to register data offers the following advantage over template matching: on the one hand, instead of using one single instance of the shape of interest, such a model uses a whole space of shapes to fit the data. On the other hand, the parameters used by the model correspond to the DoF of the analyzed shape.

During the optimization of the model parameters, it is important to pay attention to how plausible the generated shapes are. To this end, it is worthwhile to attach some sort of probability measure to the parameters. This work follows the modeling described in Bolkart (2016, Section 4.2.1) that can be summarized as follows. Basically, the data is normalized in order to use a Gaussian distribution  $\mathcal{N}(\mu, \mathbf{I})$  as statistical prior for the individual subspaces. The mean  $\mu$  of the distribution depends on the associated subspace. In the case of the speaker subspace,  $\mu = \mu(\mathbf{s})$  is set to the mean of the original speaker parameters, i.e., the rows of  $\mathbf{S}$  in Equation (3.7). For the tongue pose subspace,  $\mu = \mu(\mathbf{p})$  corresponds to the mean of the original pose parameters, i.e., the rows of  $\mathbf{P}$  in Equation (3.7). This approach is also applied to the shape space represented by a PCA model: here,  $\mu$  is set to the mean of the parameters representing the original shapes.

Using a Gaussian distribution  $\mathcal{N}(\mu, \mathbf{I})$  as statistical prior allows each parameter to be restricted during optimization to an interval of size  $2h$  that is centered at the mean. For example, this means that the tongue pose parameter component  $p_i$  of  $\mathbf{p}$  is then only allowed to lie in the interval  $[\mu(p_i) - h\sigma(p_i), \mu(p_i) + h\sigma(p_i)]$  where  $\mu(p_i)$  is the mean and  $\sigma(p_i)$  the standard deviation of the parameter  $p_i$  in the original training data. As a convention, the value  $h$  is called the prior box width in the following.

By limiting the admissible values for the parameters, unrealistic shapes may be avoided during registration.

### 3.6. Conclusion

This chapter has provided some basic background on statistical shape analysis. In particular, it has described how the meshes obtained from the basic approach of the previous chapter can be analyzed to build generative shape models. Here, two models were presented: a linear one suitable for representing shape differences occurring due to one cause. The multilinear approach allows to separate different causes of shape variability into different sets of parameters. Both variants permit to truncate the obtained model, such that only the most important DoF of shape are kept. Furthermore, it was shown how the obtained models can be used to register new data. In this context, a Gaussian distribution is used as statistical prior to avoid unrealistic shapes.

The chapter also described the requirements for the mesh collection to be analyzed, which essentially uncovers another flaw of the basic mesh extraction approach presented previously: currently, it omits any sort of proper alignment of the obtained meshes for statistical analysis. Additionally, the multispeaker datasets suffer from the issue of missing data: for some speakers, specific phones are missing. The multilinear approach, however, requires shapes for all considered phones to be present for all speakers.

## 4. Deriving statistical tongue models from MRI datasets

### 4.1. Introduction

#### 4.1.1. Motivation

The previous chapter presented two core ideas of deriving a tongue model: the linear variant that is suited for modeling one specific type of variation, e.g., either an anatomical or a pose related one. The other approach, the multilinear variant, attempts at representing both types of shape variation in one single model, a multilinear model. However, the initial approach for extracting tongue meshes from magnetic resonance imaging (MRI) scans suffered from several drawbacks that drastically limited the amount of meshes that could be obtained. Moreover, a proper alignment of the obtained mesh collection was missing, which prevented a statistical analysis. Additionally, it is still unclear if the applied subjective evaluation of the obtained tongue meshes may be considered valid.

#### 4.1.2. Contribution

The contribution of this chapter is two-fold: on the one hand, it improves the basic semi-supervised approach for estimating tongue meshes from MRI data presented in Chapter 2 to such a degree that it is able to reliably register more data than the basic approach. Furthermore, results of experiments conducted on the three presented MRI datasets suggest that the improved approach may be considered speaker and tongue pose independent. Again, the approach only requires a user to provide some annotations and optionally tune a few settings. Once more, anatomical expertise may be considered optional. On the other hand, the extended approach can be used to derive a tongue model from a provided dataset in a semi-supervised way. The properties of the tongue model depend on the type of the used database: if the database consists only of one speaker, a linear model will be produced. In the case of a multispeaker dataset, the resulting model will be multilinear providing access to two sets of parameters: one manipulating the anatomy of the tongue and another one that changes the tongue pose. These properties represent an improvement over previous methods in literature (Badin, Bailly, Revéret, et al., 2002; Badin and Serrurier, 2006; Badin, Elisei, et al., 2008; Engwall, 2003; Hoole, Zierdt, and Geng, 2003; Fang et al., 2016) that relied on manually extracting tongue meshes from MRI data in order to derive tongue models.

The chapter is based on, and extends, the following papers:

#### 4. Deriving statistical tongue models from MRI datasets

Hewer, Alexander, Ingmar Steiner, Timo Bolkart, Stefanie Wuhrer, and Korin Richmond (Aug. 2015). “A statistical shape space model of the palate surface trained on 3D MRI scans of the vocal tract”. In: *International Congress of Phonetic Sciences*. Glasgow, Scotland, pp. 0724.1–0724.5. URL: <https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0724.pdf>.

Hewer, Alexander, Stefanie Wuhrer, Ingmar Steiner, and Korin Richmond (Sept. 2018). “A multilinear tongue model derived from speech related MRI data of the human vocal tract”. In: *Computer Speech & Language* 51, pp. 68–92. DOI: 10.1016/j.csl.2018.02.001.

##### 4.1.3. Overview

The chapter is roughly separated into two parts. The first part addresses the issues of the basic approach that were identified in chapters 2 and 3. Additionally, experimental results are presented that validate the applied subjective evaluation of the result meshes. Afterwards, the extended framework is again applied to the Baker, Ultrax, and USC datasets in order to assess its performance. The second part derives tongue models of the registered data. Here, three tongue models are constructed, one from each dataset. Each one is carefully evaluated. Furthermore, the tongue models derived from the Ultrax and USC datasets are directly compared in order to find the best model. The Baker tongue model is neglected because it only provides a parameter set for manipulating the tongue pose. Finally, the findings are summarized in the conclusion of the chapter.

## 4.2. Palate reconstruction

In Figure 2.22, it was observed that missing surface information due to palatal contacts prevented the original approach from properly registering the tongue shape in the corresponding contact area. This section serves the purpose of addressing this issue by reconstructing this missing information.

To this end, the following observation is helpful: the hard palate represents a natural boundary for possible tongue motions. This is due to the fact that this part of the palate is attached to the top of the mouth and can only undergo a rigid body motion involving translations and rotations, which implies that it is possible to press the tongue against it without deforming the hard palate. This motivates the idea of reconstructing the surface of the hard palate in cases where a palatal contact is occurring and using this restored surface as a replacement for the missing tongue information. For the sake of simplicity, the approach is described for data of a single speaker in the following.

In a first step, a scan of the corresponding speaker is selected where the palate is clearly visible and its shape is estimated by using, e.g., the basic approach for estimating articulator shapes from MRI data. However, adding this surface information directly into the point cloud of another scan of the same speaker might cause the palate information to be located at the wrong location, which can be seen in Figure 4.1. This issue occurs because the corresponding speaker might have moved between the individual recordings.

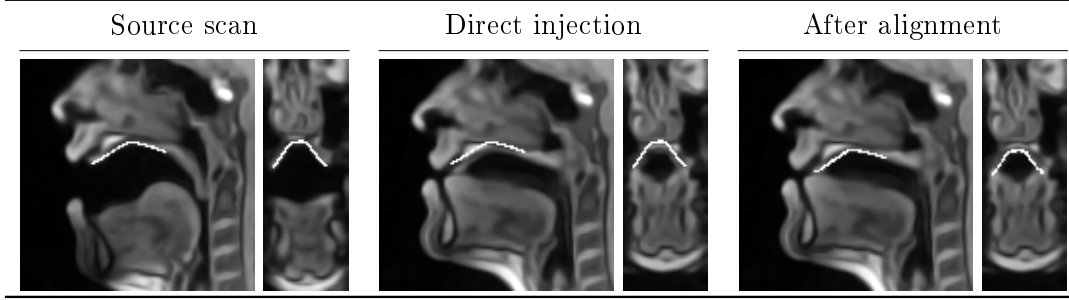


Figure 4.1.: Results for injecting the palate shape estimated from a source scan into a target scan. In the shown case, direct injection causes the information to be located at the wrong position. Aligning the palate by using the estimated head motion between both scans leads to a better result.

Thus, this head motion has to be estimated in order to align the palate information properly. For this task, an optic flow method may be used. Roughly speaking, such methods compute the movement of objects that occurred between two images  $f$  and  $g$ . An overview of such methods may be inspected in Szeliski (2010), Aubert and Kornprobst (2006), or Weickert et al. (2006). One example of such a technique is the approach by Lucas and Kanade (Lucas and Kanade, 1981; S. Baker and Matthews, 2004). Basically, this approach obtains the desired motion  $A$  by minimizing an energy:

$$E_{\text{basic LK}}(A) := \sum_{\mathbf{x} \in \Omega_T} \left( [g](A(\mathbf{x})) - [f](\mathbf{x}) \right)^2. \quad (4.1)$$

Here,  $f$  denotes the source image and  $g$  the target image where the information should be reconstructed. The domain  $\Omega_T \subset \Omega$  is a selected region of interest in  $f$ . Roughly speaking, the approach warps the target image to the source image and uses the sum of squared differences between the corresponding gray values as heuristic to determine the optimal  $A$ . As the hard palate can only undergo a rigid body motion, a simple rigid transformation is used:

$$A = M_{\text{translate}}(\mathbf{t})M_{\text{translate}}(\mathbf{o})M_{\text{rotate}}(\mathbf{x}, \alpha)M_{\text{rotate}}(\mathbf{y}, \beta)M_{\text{rotate}}(\mathbf{z}, \gamma)M_{\text{translate}}(-\mathbf{o}). \quad (4.2)$$

Here,  $\mathbf{o}$  is the center of  $\Omega_T$ . Thus, the transformation only consists of translations and rotations.

In order to use this approach, a suitable image region  $\Omega_T$  has to be chosen. In this case, it is necessary to select a region where it is known that the objects inside can only undergo a rigid motion. Moreover, the region should not suffer from the aperture problem, i.e., clear structures in the region are required. The region containing parts below and above the palate might be a suboptimal choice: as the tongue is sometimes touching the palate, objects in this region undergo more than a rigid body motion. A region fulfilling the desired properties is shown in Figure 4.2. This region only has to be

#### 4. Deriving statistical tongue models from MRI datasets

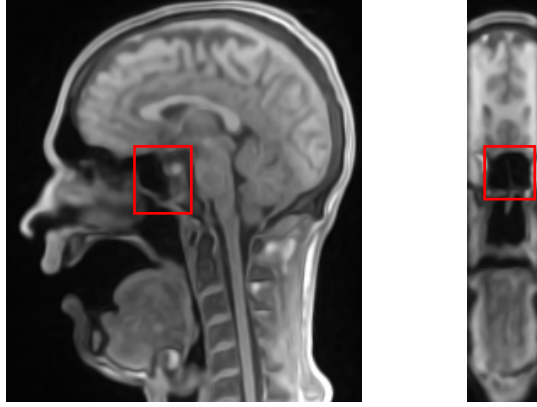


Figure 4.2.: Sagittal (left) and coronal slice (right) illustrating an example region for estimating the head motion by means of the approach of Lucas and Kanade.

defined in the source scan from which the palate shape is extracted. In order to facilitate the process of selecting the region, a user may distribute some landmarks on the scan such that the bounding box of these landmarks roughly corresponds to the desired region.

The default approach in Equation (4.1) uses the sum of squared differences to evaluate the similarity between the warped image and the source image. As it is possible that source and target differ with respect to brightness, it is worthwhile to think about using another similarity measure that makes the approach robust against such differences. Such a measure is given, for example, by the zero normalized cross correlation (ZNCC). The updated energy is thus given by:

$$E_{\text{LK}}(A) := \frac{1}{n} \sum_{\mathbf{x} \in \Omega_T} \frac{1}{\sigma_f \sigma_g} \left( [g](A(\mathbf{x})) - \tilde{g}_{A(\Omega_T)} \right) \left( [f](\mathbf{x}) - \tilde{f}_{\Omega_T} \right), \quad (4.3)$$

where  $A(\Omega_T) := \{ A(\mathbf{x}) \mid \mathbf{x} \in \Omega_T \}$ . In this formula,  $n$  refers to the number of voxels in the region  $\Omega_T$ . The value  $\sigma_f$  refers to the standard deviation of gray values of  $f$  in  $\Omega_T$ . For the image  $g$ ,  $\sigma_g$  denotes the standard deviation in  $\Omega_{A(\Omega_T)}$ . The values  $\tilde{f}_{\Omega_T}$  and  $\tilde{g}_{A(\Omega_T)}$  denote the mean gray values of  $f$  and  $g$  in the corresponding regions. As now the ZNCC is used as heuristic, the energy has to be maximized in order to achieve a high correlation and obtain the desired transformation.

Maximizing this energy provides access to the desired transformation that is required for mapping the hard palate shape to a target scan. Figure 4.1 shows how the acquired motion moves the hard palate to the correct location in the target scan. Inspecting another result of this strategy in Figure 4.3 more closely reveals an interesting observation: this kind of palate reconstruction can be used to circumvent pitfalls in manual annotation because the boundaries of the tongue can be hard to distinguish from the palate in some cases.

The vertex positions of the aligned palate mesh can afterwards be injected into the corresponding point cloud of the target scan. In addition to the vertex positions, the corresponding normal information is also added to the point cloud. This information



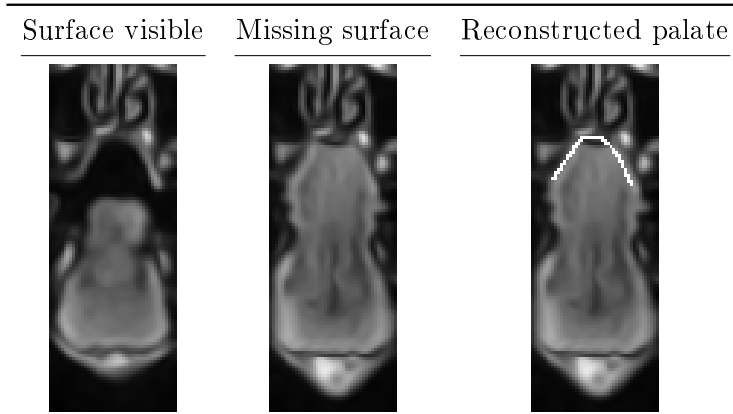


Figure 4.3.: Proposed palate reconstruction strategy may help to avoid annotation pitfalls. While the palate surface is clearly visible on the left scan, it is hard to detect on the center scan. The palate reconstruction helps to fill in the missing information.

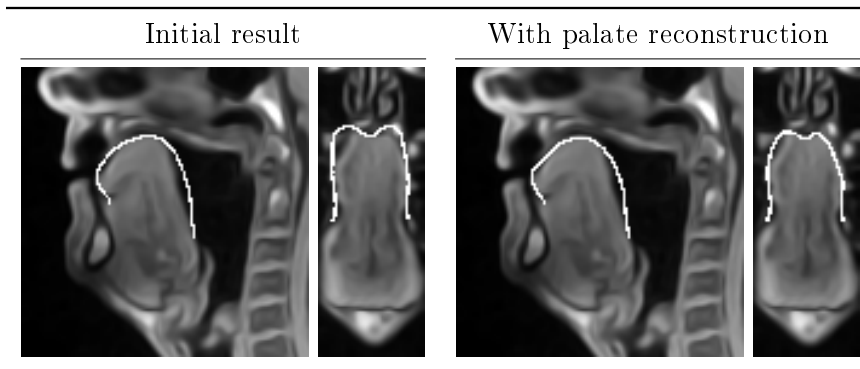


Figure 4.4.: Restoring the missing palate boundary information improves the tongue matching result to some degree.

may be derived from the mesh topology by using the approach by Max (1999). Using this modified point cloud in the template matching approach for the tongue improves the results to some degree as Figure 4.4 shows.

The soft palate, however, remains an open issue for the time being: this part of the palate can undergo highly non-rigid motions and therefore requires a more sophisticated reconstruction process.

### 4.3. Removing unrelated point cloud data

Results of the experiments in Figure 2.21 showed that the basic approach might get stuck at unrelated tissue in the volumetric point cloud. One possible explanation was that different parts of the vocal tract volume might resemble a tongue-like structure if

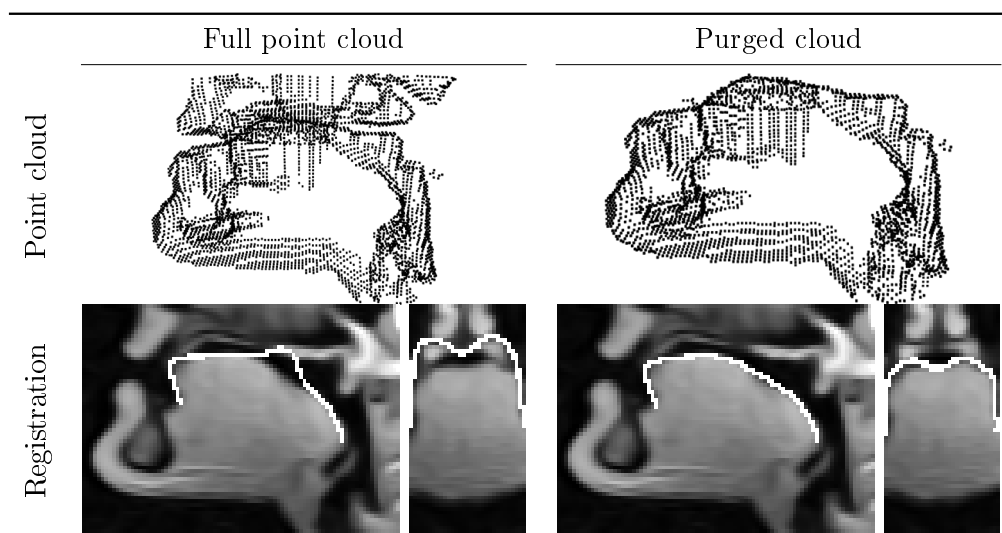


Figure 4.5.: Removing points above the hard palate improves the registration result.

merged together. Thus, a way is needed to remove such structures in order to prevent such situations. In this case, information is needed about which points can be considered as unimportant.

To this end, the previously aligned palate mesh can be used. In principle, all points lying above the hard palate are unneeded for estimating the shape of the tongue and therefore can be removed without loss of important information. Please note that the injected palate information is unaffected by this operation. Figure 4.5 shows a result of this removal strategy. It can be seen that removing unneeded data above the palate area improves the registration result.

#### 4.4. Pose normalization

Determining the variations related to speech production is the main goal of this work. However, the basic approach produced shape approximations that also included the rigid body motion of the speaker's head. In this section, a normalization is performed in order to get rid of information that is unrelated to articulation.

The Procrustes alignment technique (Dryden and Mardia, 1998) is a suitable method for removing any translational and rotational differences among meshes in a given collection. However, in the case of a tongue shape collection, removing all translational and rotational shape differences between the individual meshes actually destroys important information that is related to speech production: the tongue is connected to the lower jaw and thus can undergo a translational and rotational motion during articulation, which is unrelated to head motions. Thus, a way is needed to preserve this type of transformation while getting rid of the differences stemming from the head position.

In the case of a single speaker, the following approach could help: a single head pose

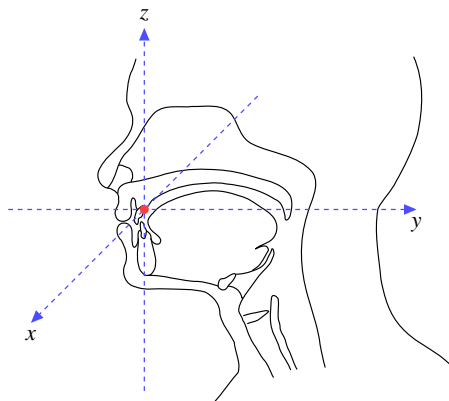


Figure 4.6.: Coordinate system used for the pose normalization. The image shows the mid-sagittal plane. The origin of the system is located near the front teeth, highlighted by a red dot; the horizontal and vertical axes represent the  $y$  and  $z$  dimensions, respectively, while the  $x$  axis is perpendicular to the image plane. Observe that compared to Figure 2.1, the roles of the axes changed.

of the speaker has to be found to which all tongue meshes can be mapped. In the section dealing with the palate reconstruction, rigid body transformations were estimated that mapped the palate shape from a source scan to all other scans of the MRI dataset. These rigid body transformations can be used to perform the pose normalization: they are inverted and applied to the corresponding tongue meshes, which brings them all into the head pose of the scan where the palate was estimated. And thus, all head motion is removed between the meshes. Additionally, the following step can be taken: the palate mesh of the corresponding scan is used to shift the origin of the coordinate system. Here, the center point in the front of the palate mesh is used as the new origin of the coordinate system. This point is roughly located at the area where the hard palate ends and the tooth region starts. This step serves the purpose of assigning a semantic value to the origin, which can be used to align data to the model. In particular, the coordinate system depicted in Figure 4.6 is used. The orientation of the axes was selected to correspond to the default orientation of the *Blender* tool (Blender Online Community, 2018) that is used throughout this work for creating renderings of meshes.

For meshes obtained from different speakers, however, the previous approach for normalizing the pose is insufficient: the original approach is only working for a single speaker where one scan is used as the reference pose. In the case of multiple speakers, however, one reference pose would then exist for each individual speaker and these poses would differ from each other with respect to position and orientation. Basically, this means that the original approach can remove head pose differences within a speaker’s data, but fails to eliminate them across speakers. As a remedy, the palate shapes of all speakers could be aligned to one single pose. This task can actually be fulfilled by applying the Procrustes alignment strategy to the palate mesh collection. This operation maps all palates into a single pose where all global translational and rotational differences between the shapes

#### 4. Deriving statistical tongue models from MRI datasets

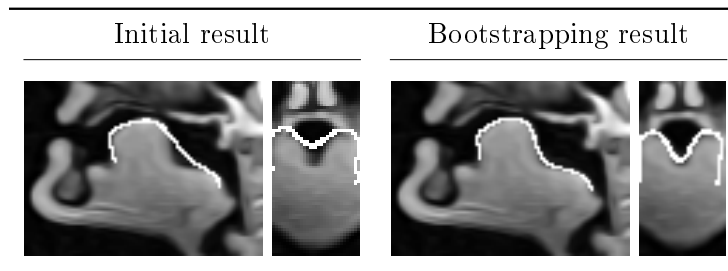


Figure 4.7.: Effect of bootstrapping.

have been minimized. Again, all palates are shifted such that the origin of the coordinate system is always located at the given point of the palate mesh. Now, all tongue shapes can be mapped into this shared pose as follows: first, estimate the motion between the palate mesh that was aligned to a scan and its Procrustes variant by using rigid alignment. Afterwards, apply this estimated motion to the corresponding tongue mesh of the scan to move the tongue to the Procrustes variant of the palate mesh. This process leads to a mesh collection of tongue shapes where transformation differences originating from the head motion within speakers and the head pose across speakers were eliminated. Additionally, translational and rotational shape differences related to speech production are preserved.

### 4.5. Bootstrapping strategy

As discovered during the experiments of the initial approach in Figure 2.21, a single template might be insufficient for registering a whole dataset of MRI recordings of the tongue. The underlying issue may be summarized as follows: the template mesh represents only one specific instance of all possible tongue shapes. In this regard, tongue forms might occur during a registration process that differ too much from this instance. Thus, the template matching could fail to estimate the correct surface.

A remedy would be to use for each scan a new template that resembles the shape that is present in the scan. This sparks the idea of first building a statistical generative model of the desired articulator and then using this model to perform a pre-registration of the corresponding point clouds. The pre-registration may be performed by using the following model fitting energy:

$$E_{\text{fit}}(\mathbf{s}, \mathbf{p}) = E_{\text{data}}(\mathbf{s}, \mathbf{p}) + E_{\text{landmark}}(\mathbf{s}, \mathbf{p}). \quad (4.4)$$

The modeling ideas correspond to the ones of the template matching in Chapter 2. This time, the mesh in the individual energy terms is obtained by reconstructing it from the current model parameters  $\mathbf{s}$ ,  $\mathbf{p}$ , and the used tongue model. As again the data term is not differentiable due to the nearest neighbor evaluation, a series of energies is optimized. It is important that the point cloud data first has to be aligned to the coordinate system of the used model because the model is unable to produce translational and rotational

motions that are related to head movements. The results of these pre-registrations are then used as the templates for the corresponding scans. However, this idea leads to a “chicken and egg” situation: on the one hand, the registrations of the corresponding articulator have to be known for building a model. On the other hand, a model is needed for obtaining the templates for performing the registrations themselves.

This observation leads to a modification of the original idea: in a first step, register all data with the initial template and perform the pose normalization. Then, build a model from the obtained registrations that are possibly partially wrong. Afterwards, use this model for performing the pre-registrations but limit the set of admissible shapes to ones that are near the mean. Restricting the shapes to this area serves the purpose of suppressing registration errors that were incorporated into the model. After that, use the pre-registration result as new template for the corresponding scan. Finally, learn a model from the newly obtained registrations. This process is known as bootstrapping and can easily be repeated iteratively until the MRI scans are registered in a satisfying way. Satisfying is in this case again a subjective measure: the user of the framework inspects the results manually and decides if the shapes are close enough to the tongue contours in the scan. An algorithmic description of this strategy is shown in Algorithm 1.

As a convention, the type of the used model in the bootstrapping strategy is assumed to be the same as the desired result model. It is, however, also possible to just use a principal component analysis (PCA) model for this purpose.

The results in Figure 4.7 show that this modification indeed solves the issue of wrongly registered shapes. The approach was only described for the tongue here, but may also be applied for registering other articulators, like the hard palate.

## 4.6. Reconstruction of missing shapes

In the case of a multispeaker dataset, the multilinear modeling approach requires the tensor  $\mathcal{X}$  to contain for each speaker the same phones. However, this requirement is violated by both the USC and the Ultrax datasets. In the Ultrax case, the phones [a, ɔ, ʌ, ə, s, j] are missing for the data of 01MRIM, the speaker of the Baker dataset. For the USC data, the phone [ə] is missing for the pilot speaker. Furthermore, [ɑ:] is missing for F5.

In these cases, the missing pose shape of a speaker is reconstructed by averaging available data: first, compute the average shape of all meshes that are present for the speaker. Afterwards, compute the mean shape of all meshes that are available for this specific pose from the other speakers in the same dataset. Finally, average both meshes. In the literature, there are more sophisticated methods to restore missing information, such as HALRTC (Ji Liu et al., 2013) or the approach by Bolkart and Wuhler (2016). In this case, however, this averaging approach was sufficient.

#### 4. Deriving statistical tongue models from MRI datasets

---

**Algorithm 1:** Bootstrapping strategy for registering articulator shapes

---

**Input:** template mesh *template*, collection of point clouds *pointClouds*, iteration amount *bootstrapIterations*

**Result:** collection of registered meshes *registeredMeshes*

```
// initialize collection of meshes
registeredMeshes ← {};
// obtain meshes from all point clouds
foreach point cloud  $P \in \text{pointClouds}$  do
    // register point cloud with original template
    mesh ← match_template(template,  $P$ );
    registeredMeshes ← registeredMeshes  $\cup$  {mesh};
end
// start the main bootstrap loop
for  $i \leftarrow 1$  to bootstrapIterations do
    // derive shape model
    model ← build_model(registeredMeshes);
    // clear results
    registeredMeshes ← {};
    foreach point cloud  $P \in \text{pointClouds}$  do
        // derive template specific for the current point cloud
        adaptedTemplate ← fit_model(model,  $P$ );
        // register point cloud with adapted template
        mesh ← match_template(adaptedTemplate,  $P$ );
        registeredMeshes ← registeredMeshes  $\cup$  {mesh};
    end
end
return registeredMeshes;
```

---

## 4.7. Validation of the subjective evaluation

As mentioned earlier, a ground truth is missing for the used MRI datasets. Thus, only a subjective evaluation of the results is possible. In particular, this evaluation serves the purpose of being an heuristic to determine if one set of settings leads to better results than another set. However, it is important to investigate if the used subjective heuristic is also shared by experts in the field. In a study (Hewer, Wuhner, et al., 2018), such an investigation was carried out. To this end, a web-based preference test was designed in order to elicit the opinion of speech experts. The goal of this experiment was to investigate whether the experts agreed with the applied subjective evaluation. In this context, it is important to note that this study used an earlier version of the proposed framework and only investigated the Ultrax dataset.

The following data was prepared for the experiment: for each of the 137 scans that were used of the Ultrax dataset, three versions of the same sagittal slice were prepared. One version showed the unannotated slice. The second version showed the slice with the projected tongue mesh contour after the initial template matching. The last version visualized the tongue mesh contour after the final bootstrapping. Afterwards, the scan set was randomly partitioned 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 to 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.

A plot summarizing the findings of the experiment can be seen in Figure 4.8. The evaluation revealed that in 83.85 % of the cases, the bootstrap result was preferred by the participants. Afterwards, it was investigated how these preferences were distributed among the different scans that were shown. For 19 scans, in 50 % or more cases, the initial template matching was preferred over the bootstrapping result. By inspecting the displayed slices of the individual scans, it can be seen that the initial and bootstrapping versions are 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 Figure 4.9.

From the relatively high acceptance rate of the obtained bootstrap results among the consulted experts two conclusions may be drawn: first, the performance of a previous version of the approach was already acceptable with regard to the quality of obtained meshes. Second, the used subjective assessment of the obtained meshes was largely shared by the experts. These conclusions can be seen as a justification for deriving a tongue model from the obtained meshes.

However, the experiment also showed that there was still some room for improvement of the given approach, which is conditioned on speaker-specific anatomy.

#### 4. Deriving statistical tongue models from MRI datasets

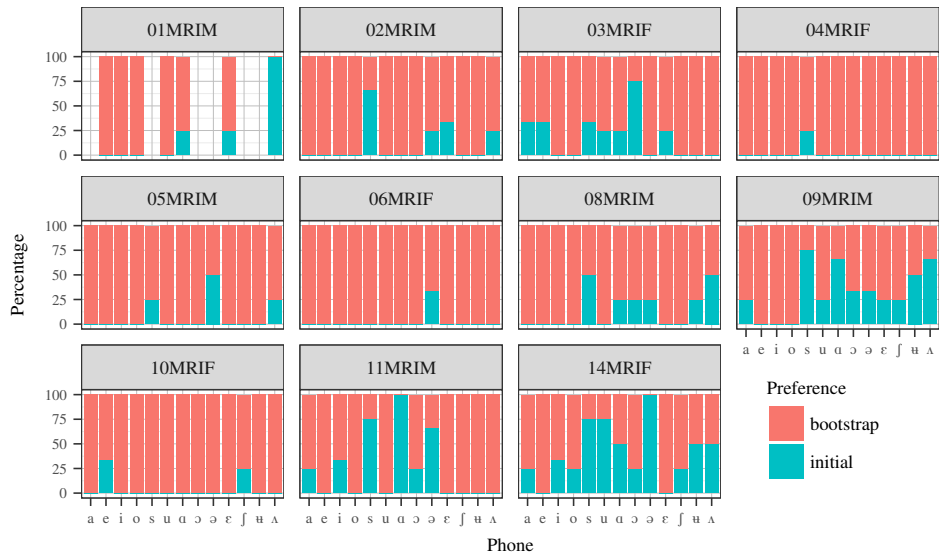


Figure 4.8.: Results of the preference test for each considered scan. Note that not all phonemes are available in the data for speaker 01MRIM. The scans were grouped by speaker to improve the visualization.

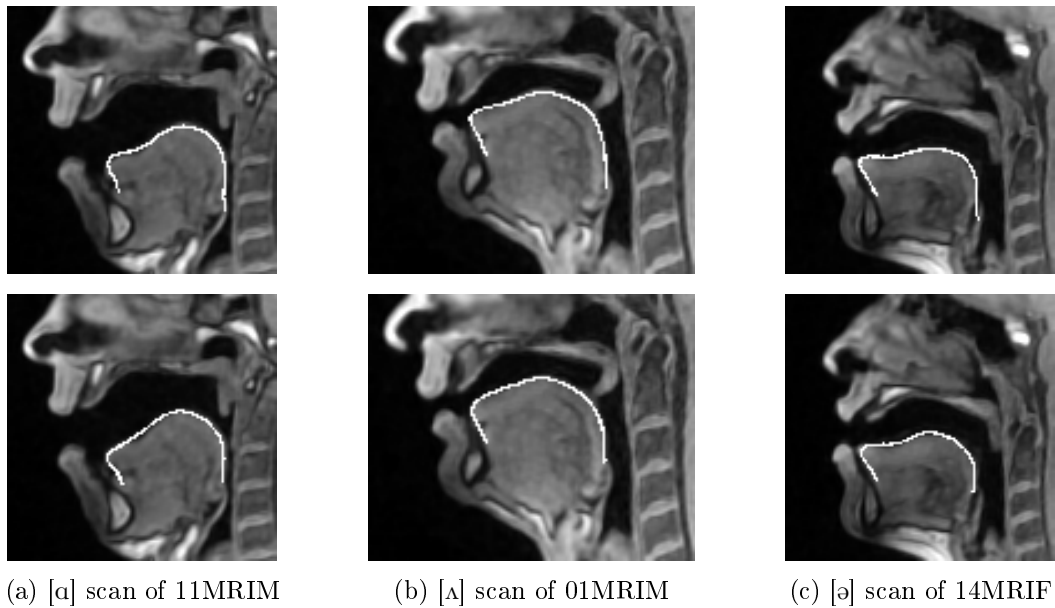


Figure 4.9.: Examples for scans where the participants preferred the initial template matching result (top row) over the bootstrapping one (bottom row).



## 4.8. Registration experiments

The previous sections addressed several issues of the initial mesh extraction approach and extended it accordingly. Moreover, the applied subjective evaluation of the meshes was validated by consulting experienced speech experts. In this section, the final approach is evaluated in order to see if the modifications lead to the desired improvements. Again, it is applied to all datasets. Due to the fact that the approach is still unable to handle contacts between the tongue and the soft palate, the following data is excluded from the experiments: all phones where a large contact area between tongue and soft palate occurs, like [ŋ, k, q]. The entire data of speaker 12MRIF in the Ultrax dataset is also ignored during the experiments because this speaker showed a high activity of the soft palate during articulation.

### 4.8.1. Settings

This section provides information about the used settings for the experiments that can be inspected in Table 4.1. Compared to the initial approach, additional settings are present. These additional settings consist of the bootstrapping settings and the settings for the palate shape estimation. The palate shape estimation is required for performing the palate reconstruction. In order to improve the acquired palate meshes, a bootstrap operation is also applied in this case. Again, the settings were manually selected. In the case of the Baker and Ultrax datasets, the segmentation is again performed by using the unweighted variant of Otsu’s method. For the USC data, again basic thresholding is applied. In the case of the bootstrapping, only very small prior boxes are used: for the tongue bootstrapping, the prior box width  $h = 0.75$  is used. The palate bootstrapping applies the width  $h = 1$ . This selection serves the purpose of avoiding overfitting during this step. The weight for the landmark term in the case of the template matching for the palate is chosen to be large and forced to remain large in order to ensure that the extremities of the resulting mesh are correctly aligned. This is important in order to recover as much of the surface of the hard palate as possible.

### 4.8.2. Observed improvements

First of all, it is of interest if the issues that were discovered in Chapter 2 were resolved by the modifications. To this end, Figure 4.10 shows a comparison between the results of the initial approach and the extended one. It becomes clear that the applied modifications solved the identified issues of the initial approach.

### 4.8.3. Results

Inspecting further results for the three datasets in Figures 4.11 to 4.13 reveals that the approach can now handle many more vocal tract configurations than before. In particular, phones where contacts between tongue and palate occur can now reliably be extracted from the datasets. Like in Chapter 2, it can be seen that speakers might have different articulation strategies for the same phone. This is, for example, the case for the

#### 4. Deriving statistical tongue models from MRI datasets

<b>Surface-enhancing diffusion</b>		
$\sigma$	standard deviation for presmoothing the image	1
$\rho$	standard deviation for combining structural information	1
$\lambda$	contrast parameter of Perona-Malik diffusivity	0.10
$t$	evolution time	2.40
<b>Point cloud extraction</b>		
$\sigma$	standard deviation for presmoothing the image	1
$\rho$	standard deviation for combining structural information	1
<b>Template matching</b>		
$\beta_{\max}$	initial weight for smoothness term	10
$\beta_{\min}$	weight for smoothness term on last iteration	6
$\gamma_{\max}$	initial weight for landmark term	0.10
$\gamma_{\min}$	weight for landmark term on last iteration	0
$n_{\text{match}}$	amount of optimization steps	100
$r_{\text{search}}$	search radius for nearest neighbor heuristic	5 mm
$\theta_{\text{search}}$	angle threshold for nearest neighbor heuristic	60 °
$n_{\text{postsmooth}}$	post smoothing iterations	1
<b>Bootstrapping</b>		
$n_{\text{bootstrap}}$	bootstrap iteration amount	2
$n_{\text{fit}}$	amount of optimization steps	10
$\sigma_i$	width of prior box for each component	0.75
$r_{\text{search}}$	search radius for nearest neighbor heuristic	5 mm
$\theta_{\text{search}}$	angle threshold for nearest neighbor heuristic	60 °
$n_{\text{postsmooth}}$	post smoothing iterations	1
<b>Template matching settings that differ for the palate</b>		
$\gamma_{\max}$	initial weight for landmark term	10
$\gamma_{\min}$	weight for landmark term on last iteration	10
$n_{\text{match}}$	amount of optimization steps	40
$r_{\text{search}}$	search radius for nearest neighbor heuristic	4 mm
<b>Bootstrapping settings that differ for the palate</b>		
$n_{\text{bootstrap}}$	bootstrap iteration amount	1
$\sigma_i$	width of prior box for each component	1
$r_{\text{search}}$	search radius for nearest neighbor heuristic	4 mm

Table 4.1.: Used settings for the experiments. Parameter name, description, and value are provided.

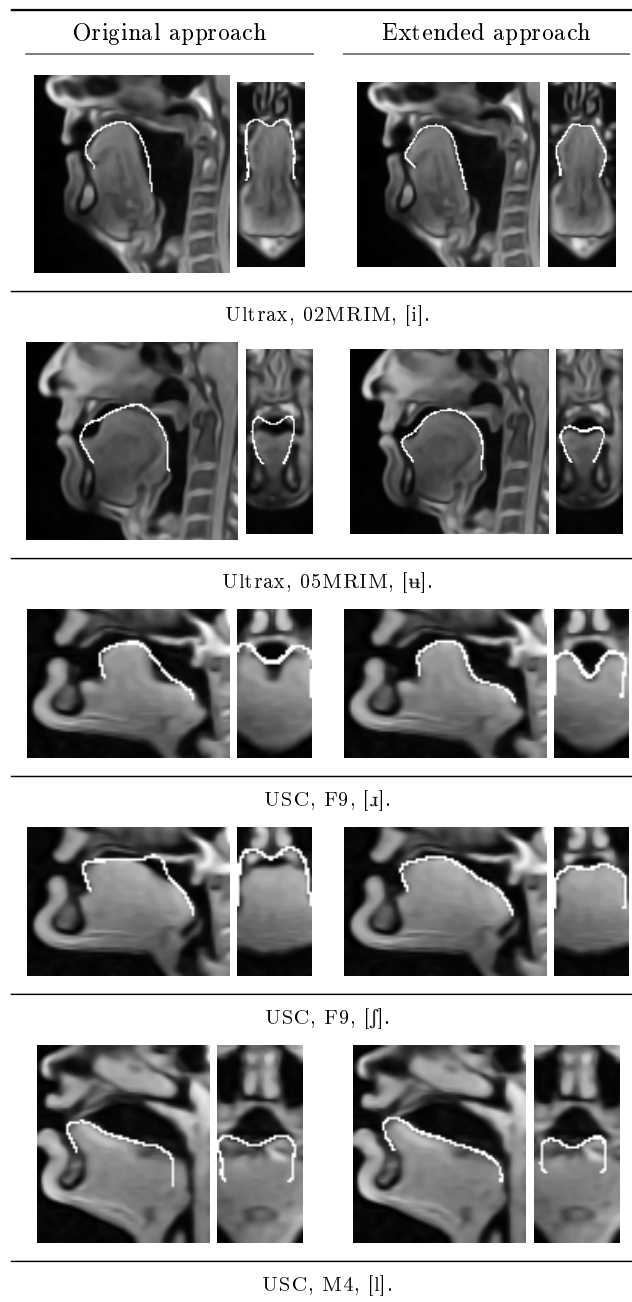


Figure 4.10.: Comparison between original approach and extended version. The registration quality significantly improves in the case of the extended approach. Dataset, speaker name, and produced phone are provided for reference.

4. Deriving statistical tongue models from MRI datasets

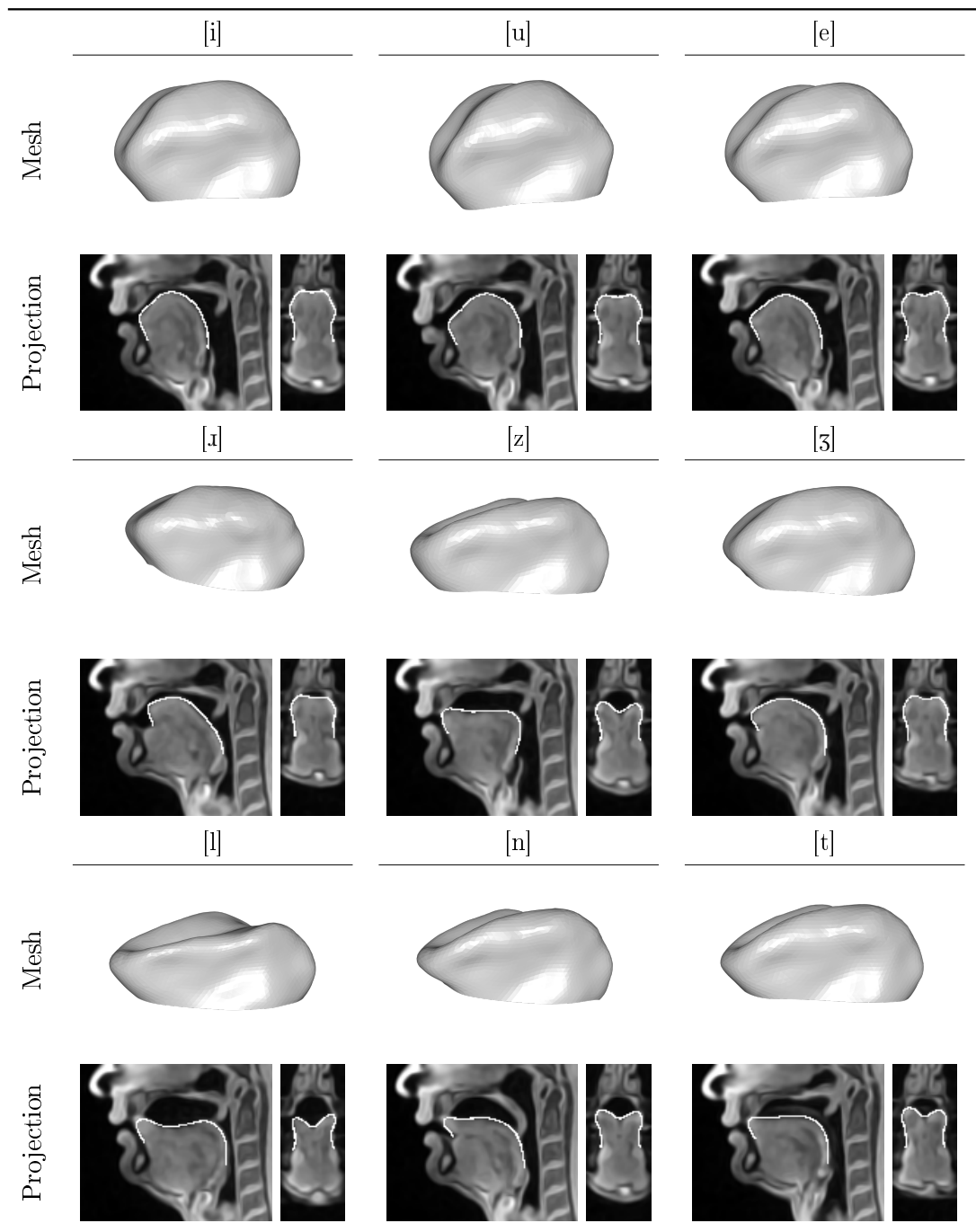


Figure 4.11.: Example results of the extended approach for selected phones in the Baker dataset.

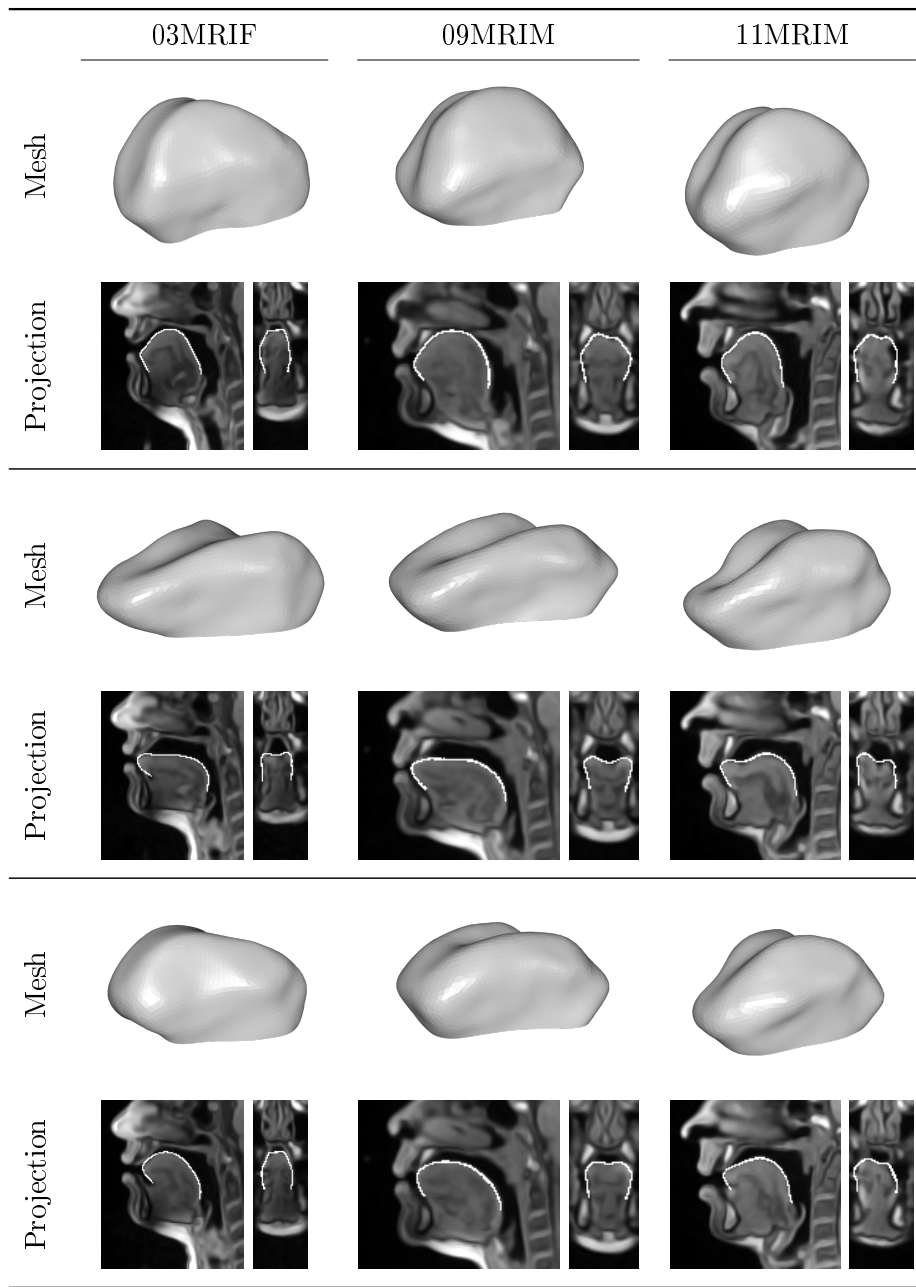


Figure 4.12.: Example results for the phones [i] (top row), [s] (center row), and [ʃ] (bottom row) in the Ultrax dataset. Registrations are shown for the speakers 03MRIF, 09MRIM, and 11MRIM.

4. Deriving statistical tongue models from MRI datasets

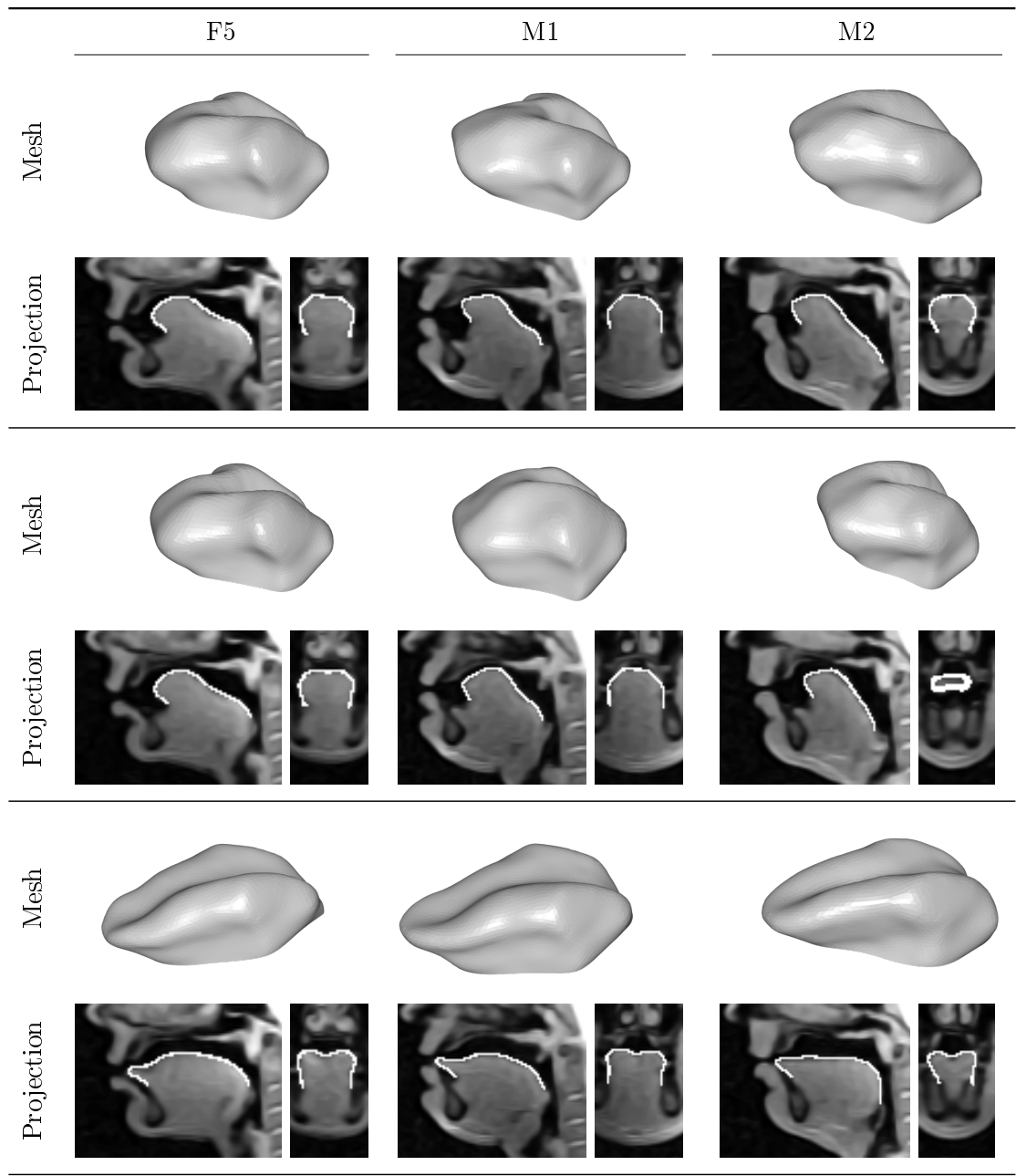


Figure 4.13.: Example results for the phones [ɪ] (top row), [ʒ] (center row), and [θ] (bottom row) in the USC dataset. Registrations are shown for the speakers F5, M1, and M2.

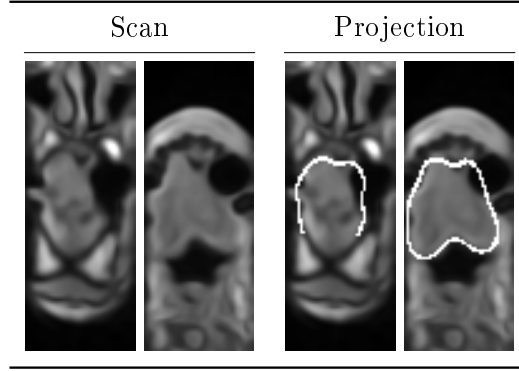


Figure 4.14.: Missing information in the scan of speaker 11MRIM for phone [i]. The estimated tongue shape (shown as projection) may be seen as reconstruction of information in this case. For visualization purposes, a coronal (left) and a transverse slice (right) are provided.

Dataset	Considered scans	Notes
Ultrax	137	Speaker 12MRIM ignored, phones [a, ɔ, ʌ, ə, s, ʃ] missing for 01MRIM
Baker	22	Phones [ŋ, k, q] ignored
USC	466	Phone [ŋ] ignored, phone [ə] missing for pilot speaker, phone [ɑ:] missing for F5

Table 4.2.: Summary of considered scans. Notes provide information about missing or ignored data in the datasets.

phone [ɹ] in Figure 4.13: the tongue tips of speakers M1 and M2 are very close to the palate during the production of this phone. The tip of F5 has some distance to the hard palate. In the same context, the speakers M1 and F5 flex the tongue back differently compared to M2.

Differences in articulation strategy can also be observed for the phone [s] in the Ultrax dataset: Figure 4.12 shows how speaker 11MRIM curls the tongue blade a bit. The other speakers, 03MRIF and 09MRIM, omit such a curling.

In the shown results, the result for phone [i] for speaker 11MRIM merits further discussion. By inspecting the coronal view of the registration, it becomes clear that the projected mesh surface has some distance from the visible tongue contour on the right side. First of all, this registration can still be regarded as acceptable. Furthermore, inspecting a transverse visualization in Figure 4.14 of this scan reveals that in this area data is actually missing. A possible explanation could be that a tooth filling was interfering with the MRI scanning procedure. In this case, the registered tongue surface in this area can be seen as a reconstruction of missing information.

In total, the extended approach successfully registered all the considered scans with Table 4.2 summarizing the results. This means that the proposed approach succeeded in 625 cases where always the same settings were used. Only the landmarks were distributed

#### 4. Deriving statistical tongue models from MRI datasets

manually for each scan. Furthermore, a segmentation technique was selected for each of the datasets. Under these circumstances, it can be claimed that the approach is speaker and tongue pose independent.

### 4.9. Model creation and evaluation

The new approach provides access to a sufficient amount of tongue meshes and an appropriate alignment of the data, such that a statistical analysis can take place. In the case of the Baker dataset, a simple PCA model is built as the data only contains information about a single speaker. Thus, only the degrees of freedom (DoF) related to actual tongue motion can be estimated from this data. For the Ultrax and the USC datasets, a multilinear model can be constructed.

During the bootstrapping strategies, unaltered versions of these models are used. The final model may be evaluated in order to decide if the parameter spaces can be truncated in order to remove unimportant information. In the multilinear case, these are two spaces: the speaker or anatomy space  $\mathbb{S}$  and the pose space  $\mathbb{P}$ . The PCA model only has one parameter space. The desired truncation is important as some parameters of the original model may represent noise in the training data. It is common to evaluate such statistical models by analyzing their compactness, generalization, and specificity (Styner et al., 2003) in order to make this decision. The parameter spaces of a good model are assumed to be compact, general, and specific. For performing these evaluations, variants of the approach by Bolkart (2016, Section 4.2.2) are used.

Moreover, it is important to compare the models with each other. This comparison serves the purpose of deciding which model is better suited for applications.

#### 4.9.1. Compactness

Compactness investigates how much the individual components of one parameter space contribute to the description of the used training data. In Figure 4.15, results for this evaluation can be inspected for all three derived tongue models.

In the Baker case, it can be seen that 4 components of the parameter space describe 93 % of the data. The original parameter space dimension is given by 22. The observation that only 4 parameters are needed to explain the majority of the data can lead to two conclusions: either the training data contains redundant information or the tongue only has a few DoF with respect to articulation.

For the Ultrax dataset, information about the compactness of the speaker and the pose parameter space is available. In the speaker case, 6 components describe 92 % of the data. In the pose case, 4 components are required to represent 90 % of variability. Again, the observation can be made that relatively few parameters in the pose space are needed to describe the majority of the data. Once more, this could be an indicator for redundancies in the training data. Interestingly, the speaker space requires more components than the pose space in order to represent the data.

Finally, the USC model also provides access to the compactness of the speaker and pose parameter space. This time, 6 parameters are needed for the pose parameter space



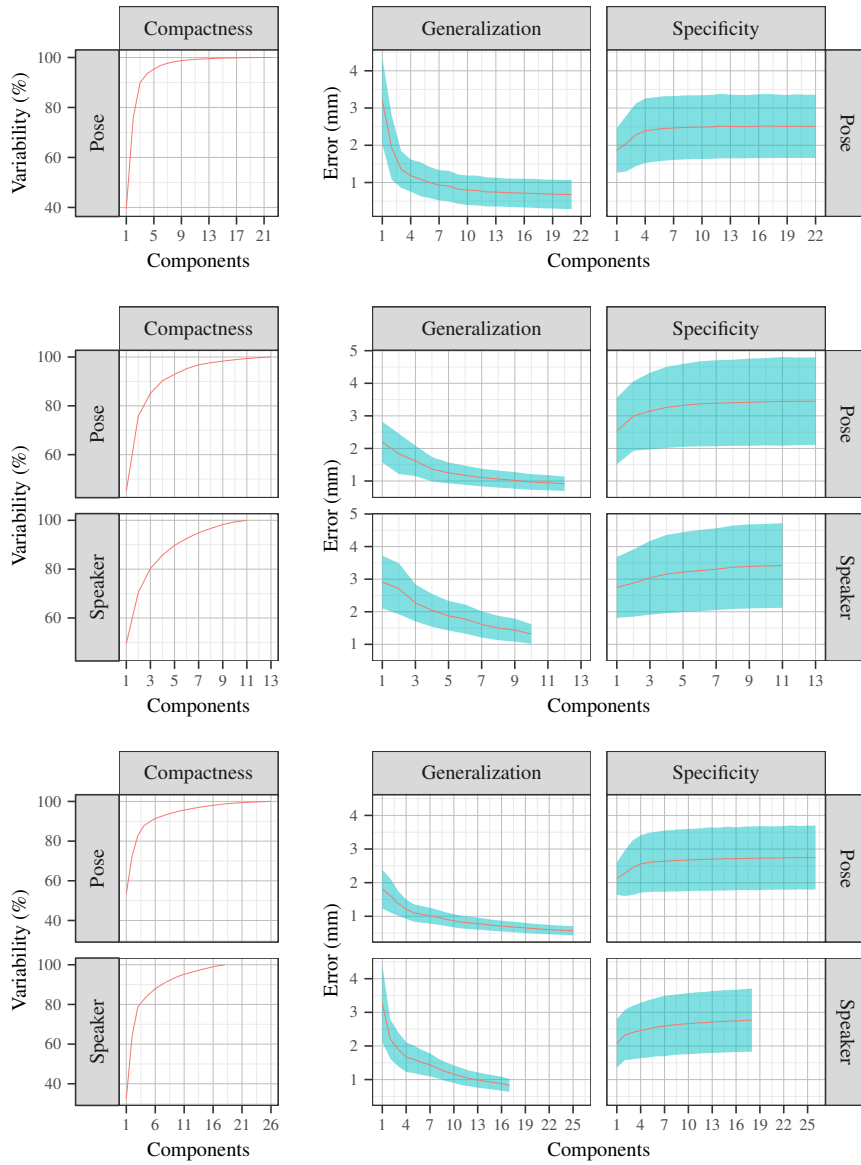


Figure 4.15.: Compactness (left), generalization (center), and specificity (right) of the evaluated models. For the generalization and specificity, the mean (line) and the standard deviation (ribbon) are shown. The plots provide the results for the Baker (top row), Utrax (center row), and USC dataset (bottom row).

#### 4. Deriving statistical tongue models from MRI datasets

to represent 91% of the training data. In the speaker case, 8 components are required to cover 91% of the data. Once again, more parameters are needed for the speaker space than for the pose space to approximately represent the same amount of data.

On the whole, the conclusion may be drawn that the pose space always requires a small amount of parameters to describe the majority of present data, which indicates that only a few DoF of the tongue are needed for speech production. Furthermore, the speaker space needs more parameters to describe the same amount of data. This could be an indicator that more DoF exist for the anatomy of the tongue than for the actual speech production.

##### 4.9.2. Generalization

Generalization measures how well the model can register data that was not part of the training. For the PCA case, the generalization is computed in a leave-one-out fashion: the mesh for each phone is once excluded from the mesh collection and a model is derived from the remaining meshes. The mesh that was excluded is then registered with the obtained model. Afterwards, the error is recorded by computing the average Euclidean vertex distance between registered mesh and target mesh. This experiment is repeated multiple times where each time a different size of the parameter space is used.

In order to evaluate the generalization for the multilinear models, both parameter spaces are examined separately: to evaluate the speaker generalization, the following steps are performed: for each speaker, a tongue model is derived from the meshes of all other speakers. Then, the tongue meshes of the excluded speaker are registered with this model. Afterwards, the average Euclidean distance between the registered meshes and the original ones are measured. Again, the experiments are repeated multiple times where each time a different size of the speaker parameter space is used. The size of the pose subspace is fixed to approximately represent 90% of the data during these experiments to prevent overfitting caused by this subspace.

In the analysis of the pose generalization, the same approach is used: for each phone, a tongue model is derived from the meshes of all other phones. Then, the meshes of the excluded phone are registered with the obtained model. The size of the speaker space is fixed to represent 90% of the data.

The results of these experiments for all tongue models are depicted Figure 4.15. During this evaluation, the width  $h = 2$  was used for the prior box in the model fitting optimization. It can be seen that increasing the size of the parameter spaces leads to better fitting results. The generalization experiments show that only a few components of the parameter spaces are needed to reliably register unseen data, which implies that the model can adapt to new tongue anatomies or poses. In particular, small parameter space sizes are enough to reach an average error that is slightly above the average measurement resolution of the original MRI scan data. For the Baker and Ultrax datasets, this measurement resolution is given by 1.2 mm. In the USC case, the resolution is 1.5625 mm. A high number of parameters leads to errors below the average resolution, which can be considered as overfitting. On the whole, the pose subspace shows better generalization properties than the speaker subspace: fewer pose parameters than speaker parameters

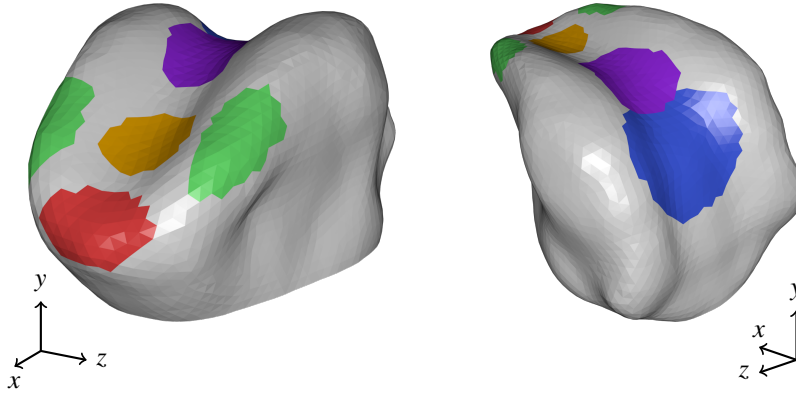


Figure 4.16.: Speech related regions of the tongue surface: Tongue tip (red), tongue blade (brown), tongue back (violet), tongue dorsum (blue), and the lateral regions (green).

are required to reach the same level of precision. A similar observation could already be made in the case of the compactness evaluation. Again, an explanation could be that the training data contains redundant information. In the Ultrax case, for example, the phone pairs  $[\Delta, \mathfrak{u}]$ ,  $[e, i]$ , and  $[e, \varepsilon]$  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.

### 4.9.3. 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 towards the tongue. In this work, it is especially of interest to investigate how large these differences are for the regions of the tongue mesh that are speech related. Figure 4.16 shows an overview of those regions.

In the PCA case for the Baker dataset, the experiment was conducted as follows. First, all tongue meshes of the training data are assumed to be valid shapes. These meshes are used to derive tongue models where the size of the parameter space is again varied, as in the evaluation of the generalization. For each size, random tongue shapes are generated from the resulting models by assuming a multivariate Gaussian distribution. These tongue meshes are compared with the ones in the mesh collection. Here, the average Euclidean distance between the generated mesh and the closest one in the collection is recorded. During this comparison and distance evaluation, a region consisting of all speech related parts is considered.

For evaluating multilinear models, the approach is changed somewhat. The specificity of the speaker and pose parameter space are evaluated separately. For measuring the specificity of the speaker subspace, the pose space is again fixed to the same size as in the generalization case. Thus, only the size of the speaker space is varied during the experiment. In the case of the pose evaluation, the speaker space is fixed and only the

#### 4. Deriving statistical tongue models from MRI datasets

pose parameter space size is changed. The size for the speaker space is chosen to be the same as in the evaluation of the generalization.

The results are shown in Figure 4.15. In total, 1 000 000 samples were always generated. On the whole, it can be seen 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. By comparing the different models, it becomes clear that the Ultrax model shows the worst specificity: in general, its mean error is higher than the ones of the Baker and the USC model.

##### 4.9.4. Fixed phone specificity

Finally, it is of interest in this work to find out how much the tongue shapes belonging to specific phones differ from the corresponding ones generated by the model. In particular, such an evaluation provides insight in how specific the model is towards the given phones. This evaluation can only be performed for multilinear models as in the corresponding training data multiple instances of the same phone are available.

The following experiment was performed for each phone: the parameters in the pose space were frozen to the ones belonging to the given phone. Then, for each size of the speaker subspace, samples are generated and the average Euclidean distance to the closest mesh is computed. Here, only meshes belonging to that phone are used to compare the generated meshes. This time, in the distance evaluation and comparison, parts of the tongue that are considered critical for this specific phone (Jackson and Singampalli, 2009) are used. In general, this region consists of the tongue blade, tongue back, and the tongue dorsum. However, for the phones [s, ʃ, θ, ð, z, ʒ, l, n] a different region is chosen that consists of the tongue tip and the tongue blade. The phones [h, m, v, f] without specific critical parts on the tongue are excluded from this evaluation.

The results of these experiments are shown in Figure 4.17. Again, 1 000 000 samples were generated. In these experiments, the phone [ʌ] 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. One explanation could be 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. Overall, the fixed phone specificity results are better for the USC model than for the Ultrax model: the observed errors are smaller and more consistent among the examined phones. The errors in the Ultrax case tend to be larger and show different standard deviations depending on the phone.

##### 4.9.5. Comparison between Ultrax and USC model

Currently, two models are of interest for applications, namely the multilinear versions. This is due to the fact that these models allow speaker adaptation because the anatomy of the tongue can be adjusted by tuning the speaker parameters. In the case of the Baker model, the anatomy is fixed. Thus, it is worthwhile to investigate which multilinear

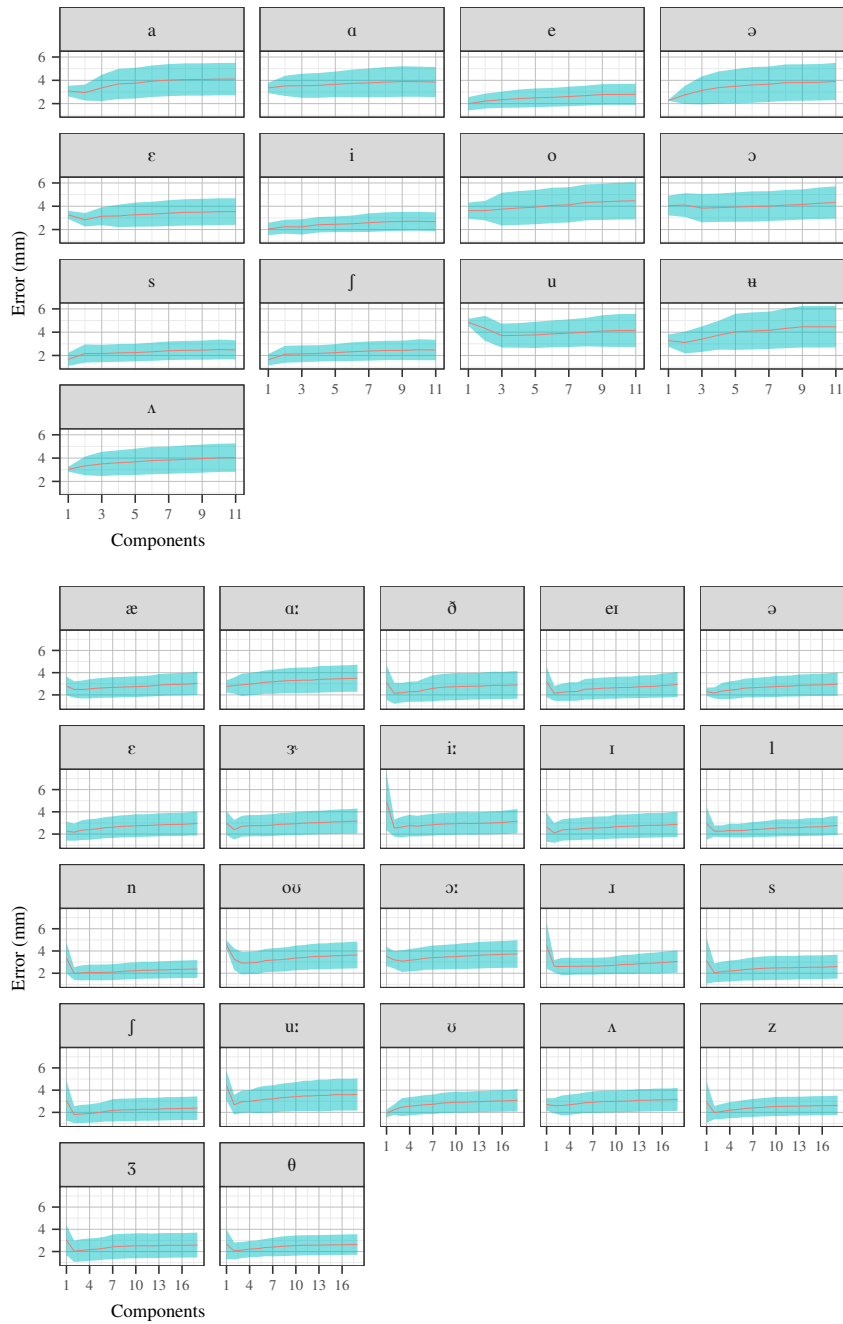


Figure 4.17.: Specificity results for the fixed phone experiments of the Ultrax (top) and USC dataset (bottom). Plots show mean (line) and standard deviation (ribbon).

4. Deriving statistical tongue models from MRI datasets

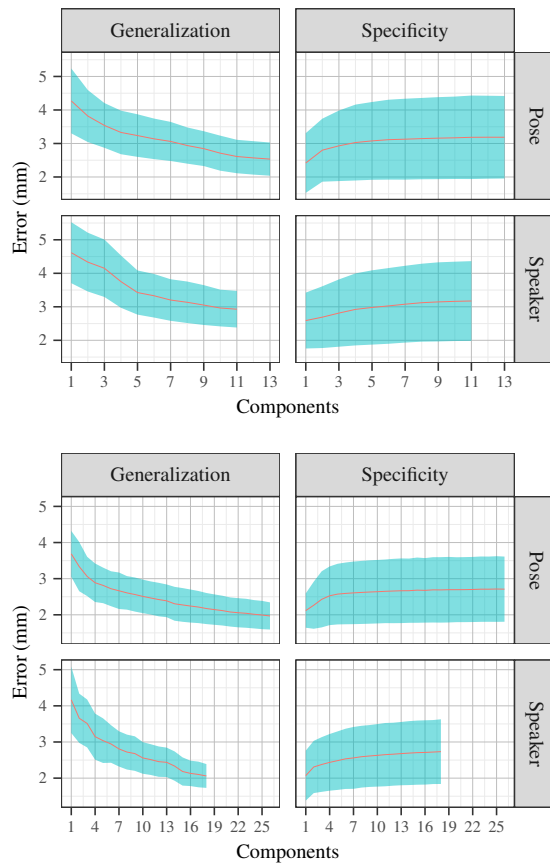


Figure 4.18.: Comparison between Ultrax (top) and USC model (bottom). For the generalization and specificity, the mean (line) and the standard deviation (ribbon) of the experiments are shown.

model is better.

So far, the models were only evaluated by using the meshes in the corresponding training data collection. In the current evaluation, the experiments are altered as follows: for computing the generalization, always the whole training mesh collection is used for deriving the models without excluding any data. The derived models are then used to register all meshes from a different dataset. This means that the Ultrax model is used to register data of the USC mesh collection and vice versa. The specificity evaluation is also modified. Instead of using only the meshes of the training data, also the meshes from a different dataset are used as examples of valid tongue shapes. For both experiments, the same settings as in the original experiments are used.

Results of these new evaluations are shown in Figure 4.18. It can be seen that the errors that were measured during the specificity evaluation are lower for the USC model than those for the Ultrax model. This observation implies that the USC model is more specific towards the tongue than the Ultrax variant. Moreover, the USC model is better than the Ultrax one at registering a different dataset, which can be seen in the generalization evaluation: fewer parameters are required for the USC model to achieve the same level of accuracy as the Ultrax model. In summary, it can be said that the USC model is superior to the Ultrax model. An explanation for this behavior could be that more data was available in the USC case for building the model. Additionally, the phone selection was different for the USC data: a balanced collection of consonants and vowels was recorded. This property might have caused the tongue model to learn about more DoF of the tongue during speech production.

#### 4.9.6. Final model

The above experiment revealed that the USC tongue model is the most versatile tongue model of the investigated ones: it allows the anatomy and tongue pose to be changed, and additionally can register unknown data. Based on the results of the compactness, specificity, and generalization evaluation, the sizes of the parameter spaces are chosen as follows: the size  $m_s = 8$  is used for the speaker parameter space  $\mathbb{S}$  and  $m_p = 6$  for the pose parameter space  $\mathbb{P}$ . These settings serve the purpose of having a good compromise between compactness, specificity, and generalization.

## 4.10. Conclusion

This chapter extended the initial approach for deriving tongue meshes from speech-related MRI recordings of the vocal tract. The modifications enabled the new approach to register many more different tongue shapes. In this context, indicators were found that the presented framework is speaker and tongue pose independent. Furthermore, the used subjective evaluation heuristic for determining the quality of an obtained registration was validated by consulting experts of the field. This was an important step because it justified using the obtained meshes for deriving tongue models.

The resulting mesh collections were used to derive three tongue models, each one representing the nature of the used dataset: a linear model derived from a single speaker

#### 4. *Deriving statistical tongue models from MRI datasets*

dataset and two multilinear models obtained from multispeaker datasets. The acquired models were evaluated in order to examine their compactness, generalization, and specificity properties. During these evaluations, it was observed that only a few tongue pose parameters were needed to explain the majority of the corresponding data. Here, also a fixed phone specificity analysis was conducted to evaluate how specific the models are towards certain phones. Additionally, the Ultrax model was compared to the USC one, which revealed that the USC model was the superior tongue model.

However, some open issues remain: an objective evaluation of the estimated tongue shapes is still missing. Furthermore, the current version of the approach is unable to adequately handle tongue configurations that have a large contact area with the soft palate. Moreover, it is unclear how good the current tongue model actually separates the anatomical shape variations from the tongue pose related ones.



# 5. Registering sparse motion capture data

## 5.1. Introduction

### 5.1.1. Motivation

The previous chapters investigated the tongue shape using static information. In the area of speech science, it is also of interest to understand the dynamics of speech production. To this end, it is common to acquire motion capture data of the human tongue for analyzing articulation. In this context, electromagnetic articulography (EMA) can nowadays be seen as the state-of-the-art modality for obtaining this kind of data. EMA is a modality that is able to track the positions of selected flesh points (Schönle et al., 1987; Perkell, Cohen, et al., 1992; Hoole and Zierdt, 2010). Roughly speaking, the tracking is performed as follows: the used EMA device or articulograph uses transmitter coils to generate an electromagnetic field. Then, sensor or receiving coils of the corresponding articulograph are attached to flesh points of interest. As soon as these coils enter the electromagnetic field generated by the transmitter coils, a current is induced in the respective sensor. By analyzing this current, the position of the sensor coil can be estimated. This modality allows to track such coils with a high temporal resolution by using electromagnetic fields, which makes it well-suited for recording motion capture data of the tongue. In contrast to the X-ray microbeam (XRMB) modality that involves ionizing radiation, EMA can be considered as harmless to the human body (Hoole and Nguyen, 1997). Currently, two major systems are available for recording such kind of data: The Wave system by Northern Digital Inc.<sup>1</sup> and the AG501 system by Carstens Medizintechnik GmbH<sup>2</sup>. Predecessors of the AG501 are the AG100, AG200, and AG500 systems. The accuracy of such systems has been widely researched. In particular, Savariaux et al. (2017) found that the AG501 and the Wave systems provide a suitable accuracy for investigating the tongue motion during speech production. Table 5.1 provides information on the systems relevant for this work. Here, it becomes apparent that depending on the use case, one system is better than another. If a study requires a device that is small, highly portable, and high accuracy is unneeded, the NDI Wave will be a good choice. However, if the study depends on high accuracy and the portability and size of the device are unimportant, the AG501 will be better suited for such a purpose.

During recordings with current articulographs, the recorded subject is mostly visible and the EMA acquisition is noiseless. This means that the EMA modality offers the

---

<sup>1</sup><https://www.ndigital.com/msci/products/wave-speech-research>

<sup>2</sup>[http://www.articulograph.de/?page\\_id=711](http://www.articulograph.de/?page_id=711)

## 5. Registering sparse motion capture data

	Wave	AG500	AG501
Number of trackable points	16	12	24
Sampling rate	100 Hz to 400 Hz	200 Hz	250 Hz to 1250 Hz
Positional accuracy (RMS)	1.5 mm	not available	0.3 mm
Coil type	single use	multi-use, requires calibration	similar to AG500
Format	portable, 7 kg	stationary, 130 kg	stationary, 64 kg

Table 5.1.: Comparison between different articulographs. The RMS values are provided by the manufacturers where no information is available on the AG500.

Study	Research topic
Ackermann et al. (1993)	speech freezing in Parkinson’s disease
Hoole and Nguyen (1997)	coarticulation effects
Jackson and Singampalli (2009)	speech production
Akdemir and Çiloglu (2008)	speech segmentation
Geng and Mooshammer (2009)	speaker normalization
Steiner and Ouni (2011)	differences between upright and supine posture
Yunusova, Baljko, et al. (2012)	palate shape estimation
Tomaschek et al. (2013)	vowel length and vowel quality
Steiner, Knopp, et al. (2014)	effect of posture and noise on speech production
Wieling et al. (2016)	dialectal differences
Hermes et al. (2018)	age-related effects on speech motor control

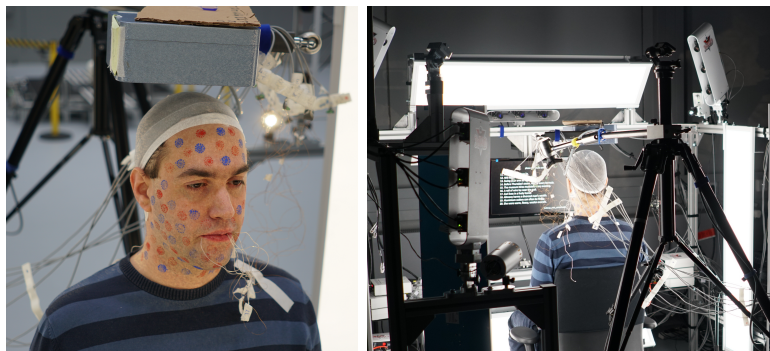
Table 5.2.: Selection of literature that applies EMA for analysis.

advantage that it can be combined with other modalities that are recorded at the same time, like, e.g., video, audio, or ultrasound. This is, for example, different from magnetic resonance imaging (MRI) recordings where special microphones and noise cancellation techniques are required to simultaneously record the audio. In Figure 5.1, pictures of such combined setups can be inspected.

In literature, EMA recordings have been carried out for researching a variety of different topics as it can be seen in Table 5.2. In addition to performed studies, researchers also worked on creating and distributing databases of such articulatory data. Some of these databases are depicted in Table 5.3. The availability of such public databases eliminates the need for researchers to record their own data and provides the advantage of making their obtained results comparable to other studies.

Despite the well established use of EMA, however, this modality has one distinctive disadvantage if compared, for example, to MRI: whereas EMA provides a high temporal resolution, its spatial coverage is quite low: sometimes, only 5 or 3 points on the tongue are tracked. This has several reasons: on the one hand, a specific distance between the individual coils is required because otherwise the electromagnetic fields generated by the coils themselves might interfere with each other (Steiner, Richmond, and Ouni, 2013). On the other hand, having too many coils glued to the tongue might make it too difficult to articulate properly for the recorded subject.

This sparseness of the data makes it hard to interpret in a visually meaningful way



(a) Acquisition of face capture and EMA data. Setup uses the NDI Wave device located above the head of the subject. Front and back view of subject are provided.



(b) Recording of ultrasound and EMA data. The subject holds the ultrasound probe while sitting in the Carstens AG501 articulo-graph.

Figure 5.1.: Examples for multimodal recording setups involving EMA. In both cases, the device including the transmitter coils is above the head of the subject.

Database	Subjects	Articulo-graph	Sampling rate
MOCHA TIMIT (Wrench, 2000)	1 male, 1 female	AG100	500 Hz
USC-TIMIT (Narayanan, Toutios, et al., 2014)	2 female, 2 male	Wave	100 Hz
TORGO (Rudzicz et al., 2012)	3 female, 4 male	AG500	200 Hz
mngu0 (Richmond, Hoole, et al., 2011)	1 male	AG500	200 Hz

Table 5.3.: Selection of databases that provide EMA recordings. The individual entries provide information about recorded subjects, the used articulo-graph, and the used sample rate for the EMA data. Notes on the MOCHA TIMIT dataset: this dataset actually contains more speakers. However, in literature, only two are used. The quality of the other recordings of this dataset is unknown.

## 5. Registering sparse motion capture data

because the data only consists of a few points moving in space without any visible context. Such context could be added as follows: the recorded trajectories of the flesh points could be used to animate a simplified representation of the tongue or even a full three-dimensional (3D) model. These representations may be helpful to researchers for studying the dynamics of the vocal tract. Moreover, such animations would be helpful in the area of computer-aided pronunciation training (CAPT). This is backed by observations in literature that visual information can help during language acquisition (Mills, 1987) and is also important for speech perception (Sumbly and Pollack, 1954; Benoit and Le Goff, 1998). Additionally, it has been shown that humans possess a general articulatory awareness, i.e., they know about the current shape and position of their tongue (Montgomery, 1981). In this regard, Engwall and Bälter (2007) found that animations of the vocal tract are desirable by students learning a language. Here, Engwall (2008) showed that such animations can indeed help language learners to correct their pronunciation.

### 5.1.2. Related work

Thus, it is understandable that creating such visualizations is an active field of research: the VisArtico tool (Ouni et al., 2012) is intended for providing easy visual access to recorded EMA data. It uses simplified representations of the lips, tongue, and jaw, and can optionally reconstruct the palate shape from given EMA data.

Badin, Elisei, et al. (2008) used a talking head to visualize EMA data. They used computed tomography (CT), MRI, and multiview video recordings to derive statistical surface models for the individual parts of the talking head. This representation provides the advantages of being fully 3D and also giving information about other parts of the vocal tract. Furthermore, the usage of statistical models may lead to realistic motions. However, this system is speaker specific and the EMA data of the same speaker was used for creating the animation.

In the work by Engwall (2001) and Engwall (2003), a linear component analysis (LCA) tongue model derived from MRI data of a single speaker was animated using EMA data of the same speaker. This model was used in a talking head application intended for CAPT (Bälter et al., 2005). Again, such an approach provides the advantage of producing 3D visualizations of the tongue that may be realistic because a statistical model is used. A limitation is the fact that the tongue model is speaker-specific and may fail to register the data of another speaker in a proper way.

Steiner and Ouni (2012) and Steiner, Richmond, and Ouni (2013) proposed a skeleton based approach (Magenat-Thalmann et al., 1988) to animate a 3D tongue mesh by using EMA data. Their approach offers the advantage of being modular, i.e., the tongue mesh can easily be integrated into various applications because of the usage of open source software. Furthermore, the tongue mesh can be customized by the user, which makes the approach adaptable to new speakers. However, the authors state that their approach is not intended to provide an accurate model for tongue shapes and motions.

Katz et al. (2014) followed a similar approach: they offer a real-time capable system that uses EMA data to animate a generic head and tongue model by applying standard

computer graphics methods. A generic model, however, may be incapable to represent the specific articulation strategy of a speaker.

### 5.1.3. Contribution

The work in this chapter also addresses the issue of creating a meaningful visualization of the tongue’s dynamics during speech production by using EMA data. In particular, it uses the multilinear model derived from the USC dataset for this purpose. The chapter is based on the following papers:

- Hewer, Alexander, Ingmar Steiner, and Korin Richmond (Mar. 2019). “Analysis of coarticulation using EMA data with a statistical shape space model of the tongue”. In: *Conference on Electronic Speech Signal Processing*. Dresden, Germany, pp. 296–303. URL: [http://www.essv.de/pdf/2019\\_296\\_303.pdf](http://www.essv.de/pdf/2019_296_303.pdf).
- Hewer, Alexander, Stefanie Wuhler, Ingmar Steiner, and Korin Richmond (Sept. 2018). “A multilinear tongue model derived from speech related MRI data of the human vocal tract”. In: *Computer Speech & Language* 51, pp. 68–92. DOI: 10.1016/j.csl.2018.02.001.
- James, Kristy, Alexander Hewer, Ingmar Steiner, and Stefanie Wuhler (Sept. 2016). “A real-time framework for visual feedback of articulatory data using statistical shape models”. In: *Interspeech*. San Francisco, CA, USA, pp. 1569–1570. URL: [http://www.isca-speech.org/archive/Interspeech\\_2016/abstracts/2019.html](http://www.isca-speech.org/archive/Interspeech_2016/abstracts/2019.html).
- Steiner, Ingmar, Sébastien Le Maguer, and Alexander Hewer (Dec. 2017). “Synthesis of tongue motion and acoustics from text using a multimodal articulatory database”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12, pp. 2351–2361. DOI: 10.1109/TASLP.2017.2756818.

The main contribution of this work is a 3D visualization approach for the tongue. However, in contrast to earlier work, it allows for speaker adaptation by estimating the anatomy and the articulation strategy of the observed speaker. Due to the usage of a statistical model, it also aims at producing natural and realistic shapes and motions of the tongue. Earlier work (James et al., 2016) has shown that this visualization technique is real-time capable and that it can easily be integrated into a talking head framework for providing feedback to a language learner. Furthermore, registering the EMA data with a tongue model can be seen as a dimensionality reduction strategy: the coordinates of all tongue coils in a single time frame are essentially assigned a single shape space coordinate represented by the parameters of the tongue model. This kind of simplification of the data could be helpful for data-driven analysis studies.

### 5.1.4. Overview

This chapter first describes how the EMA data is represented in this work. Afterwards, preprocessing steps are discussed that are needed for preparing EMA data for registration. Then, the registration process is described. Experiments are conducted to analyze the

## 5. Registering sparse motion capture data



Figure 5.2.: EMA sensor coils are glued to points of interest on the tongue.

performance of the approach by using data from 4 different speakers that was acquired from three different EMA systems. Finally, a conclusion summarizes the chapter.

### 5.2. Data Representation

During an EMA recording, multiple points of interest are tracked by gluing coils at these points as it can be seen in Figure 5.2. This means that in a continuous setting, an EMA recording can be seen as a finite set of continuous functions  $E := \{\mathbf{e} \mid \mathbf{e} \in \mathcal{C}(I)\}$  defined on the interval  $I \subset \mathbb{R}$  where  $I$  represents the time span for which the articulatory data was collected. One function  $\mathbf{e} : \mathbb{R} \rightarrow \mathbb{R}^3$  of this set denotes the trajectory of a single EMA coil over time.<sup>3</sup> Thus,  $\mathbf{e}(t)$  is the position of the respective coil at time  $t \in I$ . In practice, however, EMA devices sample these trajectories at specific time steps and thus only give access to discretized versions of these functions. Such a discretized trajectory can be modeled as

$$[\mathbf{e}]_t \approx \mathbf{e}(th_t), \quad (5.1)$$

where  $h_t$  represents the distance between two samples. As a convention, the set  $[E]_t$  is defined to contain the positions of all EMA coils attached to the tongue at time  $t$ .

### 5.3. Preprocessing

This section summarizes the preprocessing steps that are performed in this work to prepare EMA data for registration.

#### 5.3.1. Denoising

The positional data of the coils acquired from the EMA measurements is often degraded due to high frequency measurement noise. Therefore, it is important to remove this noise from the data as otherwise the registrations by the model might also suffer from

---

<sup>3</sup>It is important to note that this is a simplification of an EMA recording. Such recordings may also provide additional information like for example the orientation of the coils.

this degradation. In literature, a low-pass filter is often used to improve the quality of the data (Kroos, 2012; Yunusova, Baljko, et al., 2012; Tomaschek et al., 2013). However, care has to be taken in this regard because this procedure could also remove information related to the articulation. Therefore, it is important to inspect results manually after the denoising step in order to make sure that important motions are still preserved. Examples of low-pass filters that may be applied to the trajectories of the EMA coils are a Butterworth filter or a Gaussian filter. This work uses a Gaussian filter.

### 5.3.2. Reconstruction of missing data

In some cases, positional data might also be missing at specific time steps because the measured root mean square error (RMSE) during acquisition was too high. This work applies linear interpolation to reconstruct this missing information.

### 5.3.3. Removing head motion

It is important to note that during the recording, it is possible that the recorded subject is moving his head continuously. This motion is then also incorporated into the obtained position measurements of the tongue, which causes a problem: in general, only the motion of the articulator is of interest in research. Moreover, the derived tongue models are unable to reproduce this kind of motions as these rigid body transformations were explicitly removed from the mesh collection. Of course, this motion can be prevented by taking the necessary precautions, for example, by fixating the head of the participant, but this would make the recording very uncomfortable. As a remedy, reference coils are used that are located at specific locations on the subject such that only the rigid head motion is recorded. Example locations are, e.g., behind the ears, at the forehead, or at the upper jaw. For the following pose normalization approach, these coils must be positioned in such a way that they describe a plane, i.e., at least three coils are required. In order to remove the head motion, these reference coils are first used to build a local coordinate system at each time frame. Afterwards, the positions of all coils from the same time step are projected into this local coordinate system. This mapping removes most translational and rotational differences between time frames that originated from head motions. However, due to potential errors that affect the construction of the local coordinate system, some minor motion artifacts might remain. These errors might be the result of coils moving on the skin or measurement noise that was still present after applying the denoising filter. In this work, these errors are assumed to be negligible.

### 5.3.4. Mapping the data into a canonical coordinate system

The previous correction step removes the rigid body motion originating from head motion, which means that the orientation of the projected data depends now on the constructed local coordinate system and thus on the chosen reference coils. This leads to issues because, on the one hand, the positions of these coils differ from speaker to speaker, which makes the resulting data difficult to compare across different speakers. On the other hand, the orientation of the resulting coordinate system might be different from

## 5. Registering sparse motion capture data

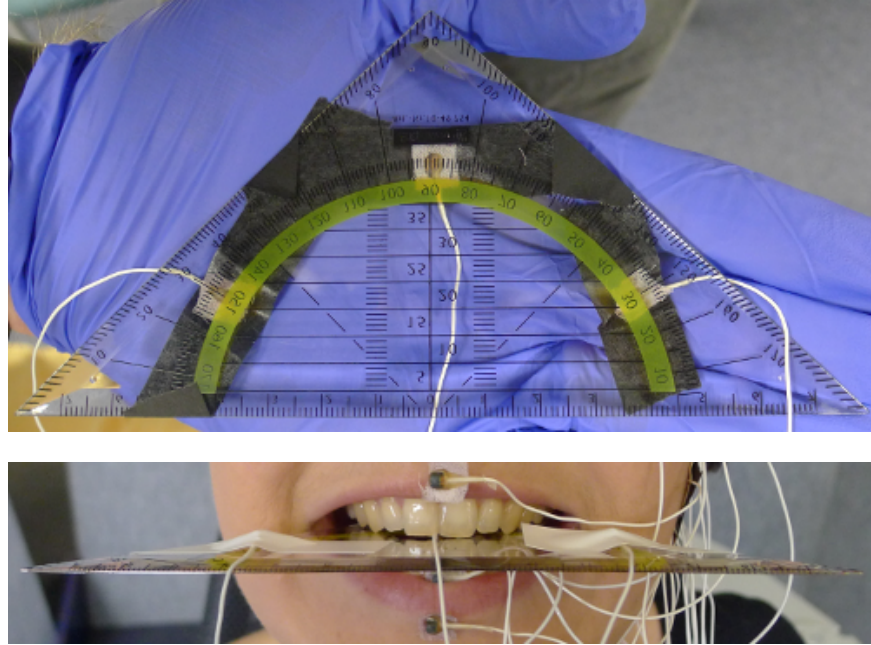


Figure 5.3.: Example setup for recording the bite plane. EMA coils are placed on a triangle ruler (top). Subject puts ruler into mouth and bites on it (bottom). Images taken from Musche (2014).

the one of the tongue model space, which prevents this kind of data from being registered properly. In literature, a so-called bite plane mapping is performed in order to bring data recorded from different speakers into a common coordinate system (Tomaschek et al., 2013). To this end, first a bite plane EMA recording is performed: here, three or more coils are attached to a plane-like surface, like, e.g., a common triangle ruler or an object created specifically for this purpose. The corresponding speaker of the recording session is then asked to put this object into his mouth and bite on it. A picture showing such a recording setup can be seen in Figure 5.3. The recorded positions of the coils of this recording are then denoised, head corrected, and then used to build a new local coordinate system: the bite plane coordinate system. It is important to note here that this bite plane coordinate system is only computed at one specific time step. This is different from the head correction step where such a coordinate system had to be constructed at each time step of the recording. This is due to the fact that the bite plane is assumed to be constant over time whereas the head motion is continuously changing. In a final step, all head corrected EMA data is then mapped into this coordinate system, which makes it comparable across different speakers.

In terms of the tongue model, this mapping has the following effect: as the  $xy$ -plane of this new coordinate system corresponds to the bite plane of the speaker, it is now also similar to the orientation of the tongue model space where the  $xy$ -plane plays a similar role. Basically, this means that the orientation of the  $z$ -axis is now similar to the coordinate system of the tongue model space.



### 5.3.5. Rotating the data

Whereas the EMA data is now aligned to the  $z$ -axis of the tongue model coordinate system, the remaining axes still need to be adapted. The  $y$ -axis describes the direction from the front of the tongue to the tongue back in the model space, as Figure 4.6 shows. Thus, the EMA data might have to be rotated in order to fulfill this property. In literature (Tomaschek et al., 2013), such an alignment is often applied to make data recorded from different speakers comparable to each other by eliminating pose related inter-speaker differences. Now the question arises how this specific direction can be estimated from the available EMA data, such that the data can be rotated. Basically, the positions of the coils attached to the mid-sagittal region of the tongue could be helpful in this case. Roughly speaking, the following steps are then performed to rotate the data accordingly. The positions for all time steps of these coils are projected into the  $xy$ -plane of the bite plane coordinate system by setting the  $z$ -coordinate to 0. Intuitively, the most dominant direction described by the resulting point cloud should correspond to the desired direction because during articulation, this is the expected behavior of the mid-sagittal region of the tongue. In order to estimate the most dominant direction, a principal component analysis (PCA) is applied to the result. The obtained direction is now used to rotate the EMA data such that this direction corresponds to the  $y$ -axis.

### 5.3.6. Mapping to origin of model

Now the orientation of the EMA data corresponds to the one of the coordinate system of the tongue model. However, the origin still has to be adjusted. As stated earlier, the origin of the tongue model is roughly located at the mid-sagittal point where the hard palate ends and the upper teeth start. This origin is illustrated in Figure 4.6. In order to obtain this point from the recorded EMA data, multiple approaches may be used.

In the easiest case, this specific point was recorded as part of the EMA acquisition process. However, if this recording is missing the respective point has to be estimated. Often, datasets also provide a so-called palate trace for each recorded speaker. Such a palate trace is obtained by creating a recording where a single EMA coil is moving along the surface of the palate. The accumulated positions of this coil then roughly describe the surface of the palate. This point cloud can then be used to manually select the point where the origin of the tongue model is located.

If a recording of a palate trace is missing, it may be estimated from the EMA data of all recordings. This is due to the fact that the hard palate represents a natural boundary for all possible articulations, an observation that was already helpful in Chapter 4 to reconstruct the palate shape in MRI scans with a palatal contact. Here, it motivates the following approach to obtain a palate trace for the corresponding speaker, which can be seen as an extension of the strategy in Ouni et al. (2012) to 3D. First, a point cloud containing all observed positions of the tongue coils is created. Now, this point cloud has to be processed in such a way that only the wanted surface information remains. This is achieved by estimating points on the upper part of the cloud's convex hull. This specific approach for computing the palate surface information can fail: the recorded tongue

## 5. Registering sparse motion capture data

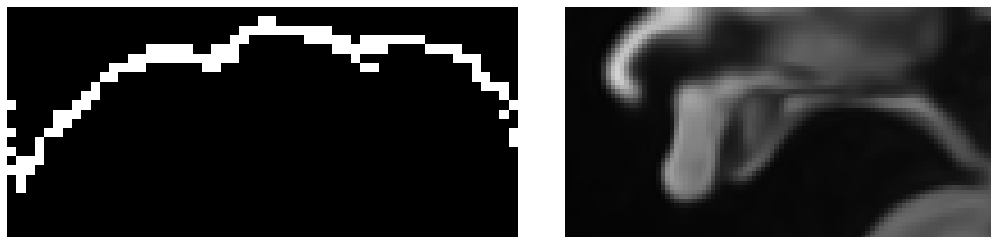


Figure 5.4.: Sagittal view of an example visualization of an estimated palate trace (left). For reference, a sagittal slice of an MRI scan of the Baker dataset (right) is shown to illustrate the shape of a palate.

flesh points might have stopped below the hard palate, e.g., because the performed articulations made it unnecessary to touch the hard palate. Thus, it is crucial to choose the appropriate data.

However, both, the palate trace and the estimated version of it are still point clouds. This makes it difficult to select the origin point. As a remedy, an image representation of such point clouds can be derived. Basically, an empty image is created with dimensions that match the bounding box of the point cloud. Before, an appropriate spacing between the voxels is chosen. Now, the point cloud is projected into this empty image. This means that voxels that contain at least one point are colored white and all other points are shown in black. A result can be seen in Figure 5.4. This image representation facilitates the process of selecting the point. The entire process of obtaining this point can be assumed to be error-prone because the used point cloud lacks semantic information. For example, this means that it is unclear which points belong to the palate and which points belong to the teeth.

Finally, the obtained reference point is used to shift the data such that the selected point corresponds to the origin of the coordinate system.

### 5.3.7. Simplification of the data

This work currently only uses EMA coils that are roughly located in the mid-sagittal area of the tongue. In order to enforce this property and simplify the data, the processed EMA data is projected into the mid-sagittal plane. Basically, this means that the  $x$ -coordinate is set to 0. This coordinate plane corresponds now to the plane where the origin of the tongue model is located. Of course, this mapping may lead to information loss by potentially removing lateral movements. However, such motions are infrequent during speech and thus the information loss may be regarded as negligible.

### 5.3.8. Comparison to related work

Other approaches also preprocessed the articulatory data before creating an animation from it. Steiner, Richmond, and Ouni (2013) used a manual cross-modal registration approach that involved data from EMA, MRI, and dental casts of the corresponding speaker for performing the alignment. However, the authors state that this approach

might be error-prone. Furthermore, the requirement for additional data, like the MRI data and the dental casts of the speaker, limits its applicability to EMA databases where this information is missing. In Badin, Elisei, et al. (2008), the authors mention that an appropriate scaling and alignment of the data was performed to match the coordinate system of the used models, but omit details. Engwall (2003) represented the EMA positions as deviations from a given reference configuration and required a means of computing the height of the lower jaw in order to properly adapt the used tongue model to the recorded data. Katz et al. (2014) mention only that the EMA data is head-corrected, but details are missing on how the data was mapped to a speaker.

## 5.4. Registering EMA data

After the preprocessing steps, an approach is needed for registering the transformed EMA data with the provided USC tongue model  $f(\mathbf{s}, \mathbf{p})$ . The strategy for fitting a tongue model to a point cloud was presented in Equation (4.4) and is given by:

$$E_{\text{fit}}(\mathbf{s}, \mathbf{p}) = E_{\text{data}}(\mathbf{s}, \mathbf{p}) + E_{\text{landmark}}(\mathbf{s}, \mathbf{p}). \quad (5.2)$$

Thus,  $[E]_t$  that corresponds to the tongue coil positions at time frame  $t$  can be used as a point cloud that should be fitted by the model. Having a look at the roles of the different energy terms, however, reveals some properties that make this approach suboptimal for registering the EMA data: first of all, this specific technique would require the model to be fitted to every point cloud  $[E]_t$  individually without using information from previous time frames. As this property only guarantees a data nearness in the individual frame, it may produce shapes that differ from the previous time step too much such that temporal inconsistencies may be observed. Moreover, this method tries to estimate a nearest neighbor for each of the vertices of the generated tongue model mesh in order to setup the correspondences that are used for optimizing  $\mathbf{s}$  and  $\mathbf{p}$ . However, each EMA data frame only contains a small number of points, which makes it difficult to estimate appropriate nearest neighbors for every vertex. Furthermore, the approach might use different nearest neighbors for a vertex at each time frame as it processes them individually, which again would increase the likelihood of temporal inconsistencies. Finally, a small number of points might be insufficient for describing the whole tongue surface. This data sparseness could lead to situations where the tongue model produces tongue shapes that are close to the data, but are highly unrealistic.

In order to address these issues, the energy has to be changed. The new version should take the temporal nature of the EMA data into account and also use a fixed correspondence between EMA coils and vertices of the tongue model mesh. An energy fulfilling these wanted properties is given by:

$$E_{\text{track}}([\mathbf{s}]_t, [\mathbf{p}]_t) = \alpha E_{\text{track data}}([\mathbf{s}]_t, [\mathbf{p}]_t) + \beta E_{\text{bias}}([\mathbf{s}]_t, [\mathbf{p}]_t) + \gamma E_{\text{coherence}}([\mathbf{s}]_t, [\mathbf{p}]_t), \quad (5.3)$$

where the quantities  $[\mathbf{s}]_t$  and  $[\mathbf{p}]_t$  depict the tongue model parameters at the given time step  $t$ .

## 5. Registering sparse motion capture data

The adapted data term  $E_{\text{track data}}(\cdot)$  uses a fixed correspondence between tongue model mesh vertices and EMA coils for computing the data nearness. Essentially, this modeling implies that at each time frame, the parameters of the tongue model are optimized such that the vertices of the generated mesh are near the positions of the associated EMA coils at the corresponding time. Therefore, this term may be seen as a modification of the landmark term that uses dynamic trajectories instead of static points. Using always the same correspondences contributes to ensuring a temporal consistency. In the following, the weight  $\alpha$  of the data term is again assumed to be set to 1. The new term  $E_{\text{bias}}(\cdot)$  penalizes deviations of the tongue model parameters from their respective means:

$$E_{\text{bias}}([\mathbf{s}]_t, [\mathbf{p}]_t) := \|\mathbf{s}_t - \mu(\mathbf{s})\|^2 + \|\mathbf{p}_t - \mu(\mathbf{p})\|^2. \quad (5.4)$$

In terms of registration, this model idea tries to address the data sparseness of the EMA modality and provides the approach with additional information about the shape of the tongue, namely that its parameters should be close to the mean, which implies that the associated shape is more realistic than shapes generated from parameter values that are further away. However, care should be taken: setting the weight  $\beta \geq 0$  of this term to a value that is too high keeps the strategy from properly registering the present data. Finally, the added term  $E_{\text{coherence}}(\cdot)$  weighted by  $\gamma \geq 0$ , the temporal smoothness or coherence term, produces energy if the current values for  $[\mathbf{s}]_t$  and  $[\mathbf{p}]_t$  differ from the ones of the previous time step:

$$E_{\text{coherence}}([\mathbf{s}]_t, [\mathbf{p}]_t) := \|\mathbf{s}_t - \mathbf{s}_{t-1}\|^2 + \|\mathbf{p}_t - \mathbf{p}_{t-1}\|^2. \quad (5.5)$$

This term serves the purpose of enforcing a temporal consistency of the shape over time. In the current modeling, von Neumann boundary conditions are used, i.e., this smoothness term produces zero energy for the first time frame.

### 5.5. Finding correspondences between EMA coils and model vertices

The energy in Equation (5.3) requires correspondences between the used EMA coils and the tongue model vertices to be known before a registration can be performed. These correspondences can be set manually. However, such a selection could be tedious and additionally lead to suboptimal results. As a remedy, a semi-supervised approach may be used to find a good correspondence.

In order to find a good correspondence, it is important to first formulate a heuristic how to compare two correspondences. Given mappings between coils and vertices, register one time frame of the EMA data with the tongue model and compute distances between the position of the EMA coil and the corresponding vertex of the registered tongue shape. A mapping is called better than another one if the mean of computed distances is smaller than the one of the other mapping. However, such a purely objective evaluation would permit correspondences where a EMA coil is associated to a vertex that is highly unlikely to be the position of an EMA coil. This motivates the idea of complementing the heuristic

by a subjective inspection. For example, the found correspondences can be compared to the information provided in the associated documentation of the data. Such information might be descriptions, simplified figures, or photographs of the coil layout.

In the following, two variants of semi-supervised methods are discussed.

### 5.5.1. Randomized approach

In earlier studies (Hewer, Wuhler, et al., 2018; Steiner, Sébastien Le Maguer, et al., 2017; Sébastien Le Maguer et al., 2017), the estimation between coils and vertices was performed by using a semi-supervised randomized approach. In this strategy, a single time frame of the EMA recordings is selected manually and the following steps are performed.

First, a random tongue shape is sampled from the tongue model and an initial nearest neighbor on the mid-sagittal area of the tongue mesh is determined for each coil. Then, these correspondences are iteratively refined by fitting the model and updating the nearest neighbors, which can be seen as an iterative closest point fitting strategy. These two steps are repeated multiple times and the best correspondences are kept according to the heuristic that was described earlier. However, care should be taken during this optimization in order to avoid overfitting. This was avoided in earlier studies by selecting a relatively small prior box width for the tongue model parameters during the correspondence estimation.

Clearly, some drawbacks of this method can be identified: the iterative closest point fitting strategy depends on the initialization. If the initialization is suboptimal, the approach may fail to find a proper mapping. This drawback can be circumvented, e.g., by repeating the approach multiple times, which then increases the running time. Another disadvantage of the strategy is the fact that it allows each coil to be associated to each tongue model vertex in the mid-sagittal area, which may lead to associations that have a small mean distance, but wrong vertex-coil mappings. While the visual inspection protects somewhat against such correspondences, this drawback might also increase the number of times the approach has to be performed before an acceptable mapping is found. Moreover, the results of this approach might be difficult to reproduce because it uses randomization as part of the strategy.

### 5.5.2. Combinatorial approach

Another way to find mappings may be a brute-force approach that probes all possible combinations of correspondences. Here, the model is fitted to the coils by using every possible correspondence between EMA coil and tongue model vertex. However, this means in the case of the used tongue models that for three coils, e.g.,  $3100^3$  combinations have to be probed. On the one hand, this is infeasible because the resulting space of all combinations is simply too large to be adequately processed. On the other hand, this space also contains implausible combinations: for example, a combination that maps all coils to a single vertex or a combination that uses vertices where EMA coils are never attached. This is a similar issue as in the randomized approach described previously.

## 5. Registering sparse motion capture data

The question arises how such problematic combinations may be avoided. In this regard, the following observation is helpful: literature that recorded EMA data and published the acquired dataset provides information about where the coils were attached. This information can be used to drastically reduce the search space for the brute-force approach: each coil is now assigned to a specific subset of mesh vertices where the respective coil could have been located. Then only combinations have to be probed where the mappings take this assignment into account. Assuming that these subsets consist on average of 30 vertices, this restriction reduces the space to approximately  $30^3$  combinations. As each one of the combinations can be evaluated independently, parallelization may be used to speed up the computation. This approach represents a clear advantage over the randomized variant: it avoids producing mappings that involve vertices that are unrelated to the region used for the coil placement. Furthermore, it is deterministic, which means that results of this strategy are reproducible. However, disadvantages may also be identified: this technique requires additional input, namely the regions where the coils were located. It may also be slower because all possible combinations are probed.

Again, care should be taken during this optimization in order to avoid overfitting. Here, the same countermeasures are taken as in the randomized approach: the prior box width for the tongue model parameters is chosen to be relatively small.

Due to the identified advantages of this approach, it is used instead of the randomized variant in this work.

## 5.6. Estimating the speaker anatomy

After the correspondences have been estimated, it is possible to register the EMA data with a tongue model. This registration process, however, optimizes both the speaker space parameters and the pose space parameters, which may lead to animations that might appear unrealistic. In particular, the energy in Equation (5.3) currently allows the anatomy to change over time. While this strategy uses these additional degrees of freedom to better register the data, it can lead to tongue animations where the tongue is for example continuously changing its anatomical properties, like, for example, its size, which appears unnatural. This can, e.g., be inspected in the supplementary material of Hewer, Wuhler, et al. (2018) where an animation of the tongue is shown that was obtained from optimizing both sets of parameters.

Such a behavior may be avoided by simply freezing the speaker parameters to certain values during the registration and only optimizing the pose parameters. Basically, this means that the tongue model is adapted to the speaker anatomy and essentially represents a speaker-specific PCA model. Of course, a reasonable choice of parameters is needed that models the anatomy of the speaker that should be registered. If only EMA data is available of the corresponding speaker, the anatomy may be estimated from this data: first, all recordings of this speaker are registered by minimizing the original energy in Equation (5.3). Afterwards, the obtained speaker parameters are averaged. These mean speaker parameters are assumed to represent the anatomical features of the speaker. Here, it is important to note that the very sparse motion capture data of the tongue

Gaussian filter		
$\sigma$	standard deviation of Gaussian kernel	10
Tracking		
$\beta$	weight for mean bias term	5
$\gamma$	weight for temporal coherence term	5
$h$	width of prior box	5

Table 5.4.: Used settings for the EMA registration experiments. Parameter name, description, and value are provided.

might be insufficient for estimating the true anatomy of the respective speaker.

## 5.7. Experiments

In this section, the tongue model derived from the USC dataset is used to register the EMA data of three different sources. Performing the registration with the other tongue models is omitted here because during the evaluation of the models in Chapter 4, the USC model proved to be the most versatile one. These experiments serve the purpose of investigating if the tongue model is able to reliably register dynamic and sparse speech production data. Furthermore, it is of interest to explore if the trajectories of the obtained shape space coordinates may be used for data-driven analysis of EMA data.

Basically, the following registration steps are always performed for each dataset: all EMA recordings are registered with the USC tongue model twice. First, all tongue model parameters are optimized to fit the data. Afterwards, the speaker anatomy is estimated, and all data is fitted again where the anatomy is frozen and only the pose parameters are optimized. The preprocessing steps depend on the considered dataset and are discussed in the corresponding section.

However, care should be taken when working with EMA data: there are indicators in literature (Weismer and Bunton, 1999; Meenakshi et al., 2014; Hoole and Nguyen, 1997) that having coils or other markers glued to the tongue may affect the articulation strategy of the speaker, which means that the recorded motion capture data is different from the natural articulation of the speaker. As a consequence, the registered tongue shapes might also represent this unnatural articulation strategy.

### 5.7.1. Settings

Table 5.4 shows the settings that were used for the experiments. Like the settings in chapters 2 and 4, these were manually selected to reach acceptable results. Here, it can be seen that a rather high value is used for the width of the prior box, which basically gives the tongue model a lot of freedom during the registration. Such a freedom might be needed because the model might encounter unknown tongue shape configurations in the data. The vertex subsets that are used for the correspondence optimization are visualized in Figure 5.5.

## 5. Registering sparse motion capture data

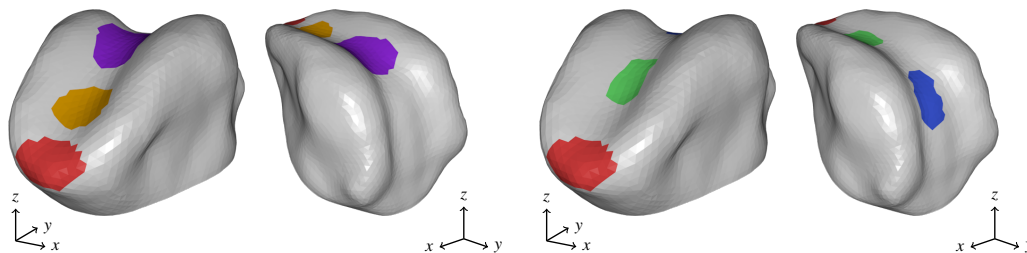


Figure 5.5.: Visualization of vertex subsets that are used in the combinatorial correspondence optimization approach: tongue tip (red), tongue blade (brown), tongue body (green), tongue back (violet), and tongue sagittal dorsum (blue). Due to overlapping vertex sets, multiple renderings of the tongue mesh are shown.

### 5.7.2. The EMA subset of *mngu0*

The *mngu0* dataset (Richmond, Hoole, et al., 2011; Steiner, Richmond, Marshall, et al., 2012) is a multimodal corpus of articulatory data: it contains audio, video, EMA recordings, and time-aligned phonetic transcriptions of a single British speaker. Furthermore, volumetric MRI recordings of the speaker’s phonetic inventory are available, as well as 3D scans of dental plaster casts taken from his lower and upper jaw. In this experiment, the EMA subset of the corpus is of interest that contains recordings of over 2000 utterances. The utterances were selected from English newspaper text with the goal of maximizing the coverage of context sensitive diphones. A diphone is one possible representation of a discrete unit of speech. It consists of two contiguous half-phones ranging from the center of one phone to the center of the next. All data was acquired with the AG500 at a sampling rate of 200 Hz. The recordings were conducted over two days where a different layout for the EMA coils was used on each day. In this experiment, the recordings of the first day are used that consist of 1354 utterances, which results in 67 minutes of speech production material. On this day, three coils were attached to the mid-sagittal region of the tongue as depicted in Figure 5.7.

This dataset is publicly available for research purposes. For the experiment, the following distribution packages were downloaded from the *mngu0* website, <http://mngu0.org>:

1. Day1 basic EMA data, head corrected and unnormalized (v1.1.0)
2. Day1 transcriptions, Festival utterances and ESPS label files (v1.1.1)

This experiment uses the unnormalized data as it still contains the silent intervals before and after an utterance in order to also investigate how the tongue model behaves during non-speech activities. By using the label files, it is possible to investigate which phones are present and how often they occur in the dataset. A histogram providing this information is shown in Figure 5.6. It becomes apparent that this dataset contains phones that are unknown to the USC tongue model because instances of them were missing in the training data. Examples of these phones are given by [ŋ, g, k]. The label files also fulfill another purpose: as they describe when a specific phone was produced in an utterance, they add phonetic context to registrations later on for visualization purposes; during the registration itself, there are ignored.



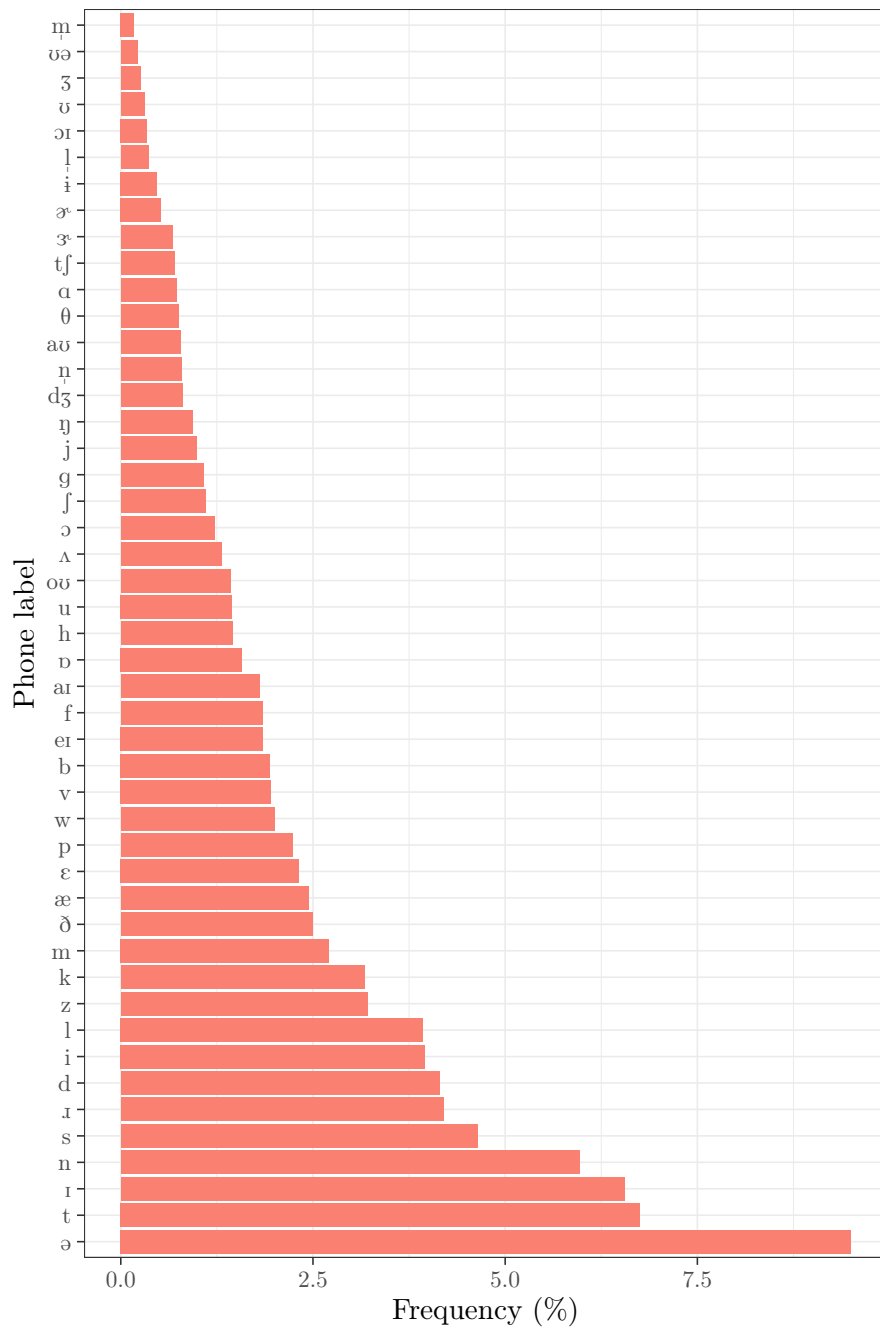


Figure 5.6.: Histogram of phone occurrence in the mngu0 dataset. The phone [ʒ] is omitted because it only occurs once.

## 5. Registering sparse motion capture data

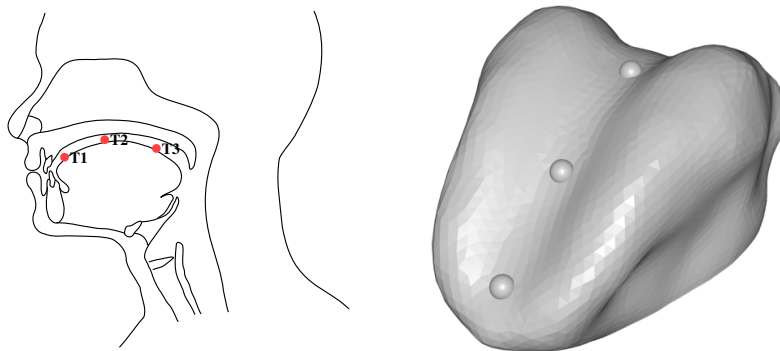


Figure 5.7.: Illustration of tongue coil layout for the mngu0 dataset (left, adapted from Richmond, Hoole, et al., 2011) and rendering of combinatorial correspondence optimization result (right). Spheres on tongue mesh highlight the vertices that were found.

It is important to note that this dataset provides a specific feature: the EMA coils stayed in the same location during one recording day, which distinguishes it from datasets like MOCHA TIMIT (Wrench, 2000) where coils became detached and reattached during the recordings. In the case of reattaching a coil, the preprocessing step of estimating the correspondence between coil and tongue model vertex has to be repeated because it is very unlikely that the coil was attached to the exact same location as before.

### Preprocessing

As the downloaded data itself is already processed to some degree, i.e., head correction is assumed to be already performed, all preprocessing steps up to and including the bite plane mapping were omitted. As this dataset is missing a recording for the needed reference point, it was estimated by reconstructing the palate trace using the EMA recordings. For estimating the correspondence between tongue coils and tongue model vertices, the vertex subsets *tongue tip*, *tongue body*, and *tongue sagittal dorsum* were used. During the estimation, the width  $h = 0.5$  was used for the prior box. The combinatorial approach produced the result depicted in Figure 5.7.

### Results

On a global scale, the following information of the results were computed: Figure 5.8 shows the cumulative error of the two registration runs where the error was measured by computing the average Euclidean distance between coils and corresponding vertices at each time step. The distribution of the tongue model parameters is shown in Figure 5.9. Note that the normalized versions of the model parameters are shown, i.e.,  $v$  represents the value  $\mu(x) + v \sigma(x)$  where  $\mu(x)$  and  $\sigma(x)$  are the mean value and standard deviation of the corresponding parameter entry  $x$  in the training data of the tongue model, respectively. This scaling helps to identify issues of the registration approach: very high or low values indicate that the produced shape is implausible to the tongue model, which might

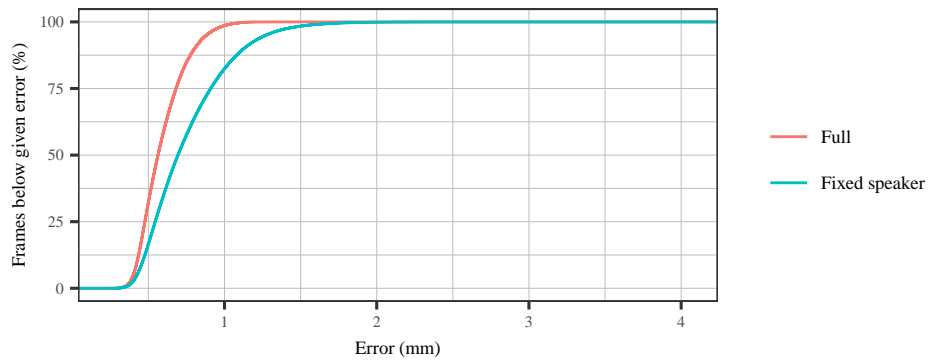


Figure 5.8.: Cumulative error for the two registrations of the mngu0 data.

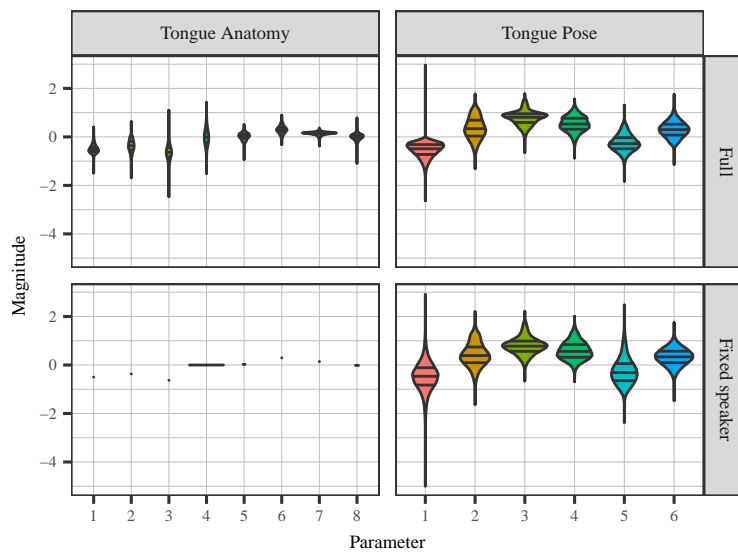


Figure 5.9.: Tongue model parameter distribution for the two registrations of the mngu0 data. The violin plots show the density, with the mean and interquartile range marked by horizontal lines.

## 5. Registering sparse motion capture data

	Experiment 1				Experiment 2			
	Mean	Std. dev.	Min	Max	Mean	Std. dev.	Min	Max
$s_1$	-0.50	0.17	-1.49	0.41	-0.50	0	-0.50	-0.50
$s_2$	-0.37	0.25	-1.69	0.63	-0.37	0	-0.37	-0.37
$s_3$	-0.63	0.33	-2.47	1.10	-0.63	0	-0.63	-0.63
$s_4$	-0.00	0.33	-1.51	1.42	-0.00	0	-0.00	-0.00
$s_5$	0.03	0.14	-0.93	0.50	0.03	0	0.03	0.03
$s_6$	0.29	0.16	-0.32	0.89	0.29	0	0.29	0.29
$s_7$	0.14	0.07	-0.37	0.37	0.14	0	0.14	0.14
$s_8$	-0.02	0.16	-1.10	0.79	-0.02	0	-0.02	-0.02
$p_1$	-0.54	0.30	-2.63	2.95	-0.47	0.60	-5	2.90
$p_2$	0.36	0.46	-1.31	1.78	0.43	0.49	-1.64	2.21
$p_3$	0.78	0.30	-0.65	1.78	0.79	0.36	-0.66	2.20
$p_4$	0.54	0.32	-0.87	1.56	0.59	0.37	-0.68	2.00
$p_5$	-0.25	0.35	-1.83	1.31	-0.26	0.54	-2.37	2.47
$p_6$	0.30	0.33	-1.13	1.75	0.33	0.35	-1.47	1.74
Error (mm)	0.59	0.15	0.23	1.41	0.75	0.27	0.24	4.05

Table 5.5.: Statistics of tongue parameters and error for the two registrations of the mngu0 data.

indicate that it encountered an unknown shape. Finally, Table 5.5 provides statistics about the measured error and the tongue model parameters.

### General observations

By inspecting the cumulative error, it can be seen that the registration approach produced acceptable results: if all tongue model parameters are optimized, then nearly all errors are below 1 mm. In the fixed speaker anatomy scenario, the errors increase, but are mostly below 1.5 mm. This loss of accuracy can be linked to the smaller number of degrees of freedom (DoF) of the speaker-specific PCA model, which only offers 6 DoF instead of 14 like the original model. It is interesting to observe that the variance of the pose parameters increased in the fixed speaker case. This was to be expected: the fitting approach now has fewer DoF to work with. Thus, the other parameters have to be used more to perform the fitting. From the data perspective, the full model registration is superior to the fixed speaker one. However, from a perception perspective, the fixed speaker type of registration may be preferable over the full optimization variant because it offers anatomy consistency over time, which may be seen in Figure 5.10.

### Silence analysis

In the parameter distribution and the statistics, one phenomenon requires an investigation: the distribution of the  $p_1$  tongue pose parameter assumes the lowest value that was allowed by the optimization in the fixed speaker registration: it is located at  $\mu(p_1) - 5\sigma(p_1)$ , which clearly indicates that the model might have encountered a shape

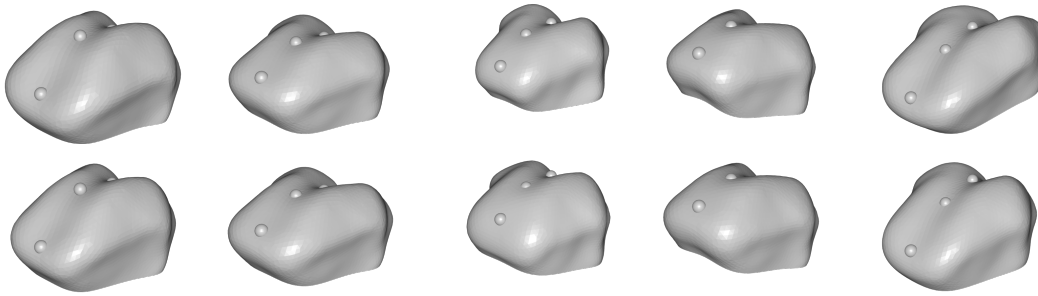


Figure 5.10.: Visual comparison between results obtained by full (top) and fixed speaker optimization (bottom). Spheres indicate EMA coil positions. The full optimization is continuously adapting the anatomy over time: the tongue shrinks or grows. The anatomical features remain consistent in the fixed speaker case. Registrations were obtained from utterance *s1\_0002*.

that is unknown to it. Moreover, the maximum for the measured errors increases from 1.41 mm in the full model registration to 4.05 mm, which could be related to the behavior of the  $p_1$  parameter. As the label information for the utterances is available, it is possible to determine the context for the region where the extreme value for  $p_1$  occurs. It is located in the utterance *s1\_0711*:

“There is no clear idea on where all this fresh water is going to come from.”

Figure 5.11 shows the trajectories of the EMA coils and the corresponding trajectories of the tongue model vertices for the fixed speaker registration. Furthermore, the associated Euclidean distance between vertex and coil position is shown over time. In addition to that, Figure 5.13 shows the evolution of the associated tongue pose parameters for this specific utterance.

The figures show that the model in general fails to stay perfectly in the mid-sagittal plane, which may be explained as follows: the corresponding tongue meshes might lack vertices that lie exactly in this specific plane. It becomes clear that the tongue model seems to struggle with the silence interval after the speech segment. This claim is also supported by the reconstructed tongue mesh and the coil positions at time stamp 4.26 s that can be seen in Figure 5.15. In particular, the tongue model fails to properly register the tongue blade (T2) and back (T3) coils. One explanation might be that the speaker used this silence interval to relax, which led to a non-speech motion of the tongue. As a consequence, this observation could be an indicator that non-speech tongue shapes exist that describe other DoF that are unused during speech production. Of course, the question arises how the full registration approach handled this situation. Figures 5.12 and 5.14 show the corresponding results. First of all, it can be seen that this strategy performs better than the fixed speaker one because, on the one hand, the occurring errors are smaller and, on the other hand, the parameters stay within a reasonable interval. However, the parameter evolution show that this approach is continuously adapting the anatomy of the tongue shape, which may lead to temporal inconsistencies. Moreover, the resulting tongue mesh at the corresponding time stamp shown in Figure 5.15 reveals

## 5. Registering sparse motion capture data

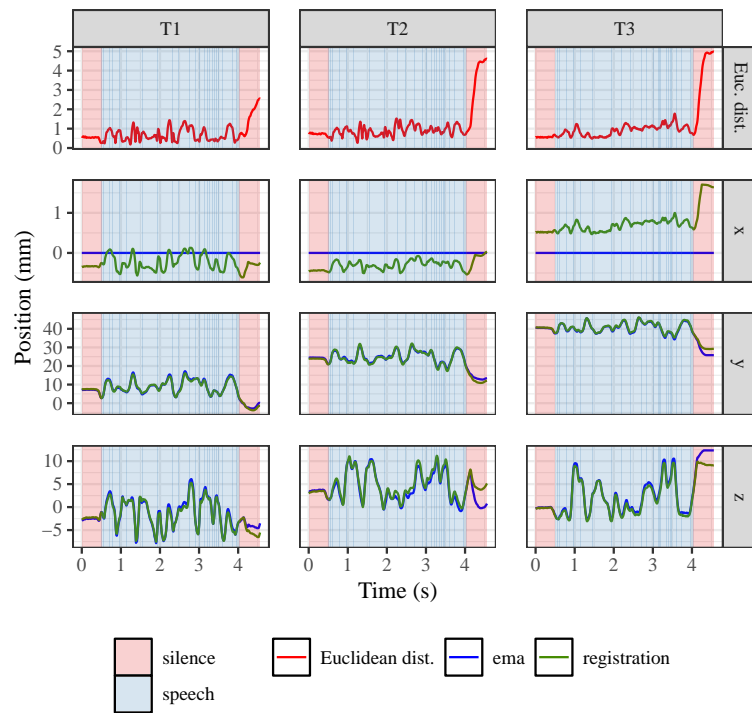


Figure 5.11.: Trajectories of EMA coils and corresponding vertices of registered tongue model for utterance *s1\_0711*. The top row shows the Euclidean distance between coil and corresponding vertex over time. The registration was obtained by optimizing only the tongue pose parameters and fixing the anatomy.

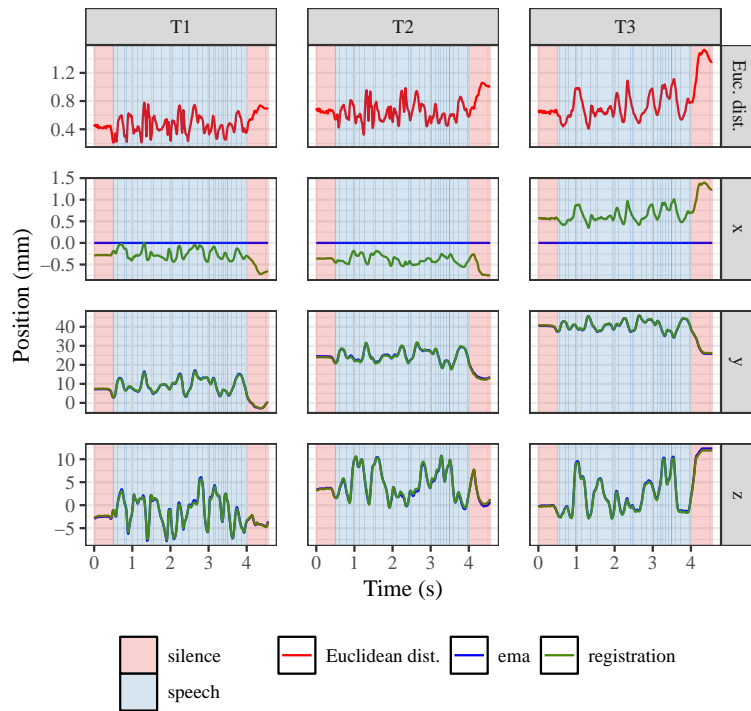


Figure 5.12.: Trajectories of EMA coils and corresponding vertices of registered tongue model for utterance *s1\_0711*. The top row shows the Euclidean distance between coil and corresponding vertex over time. The registration was obtained by optimizing all parameters.

5. Registering sparse motion capture data

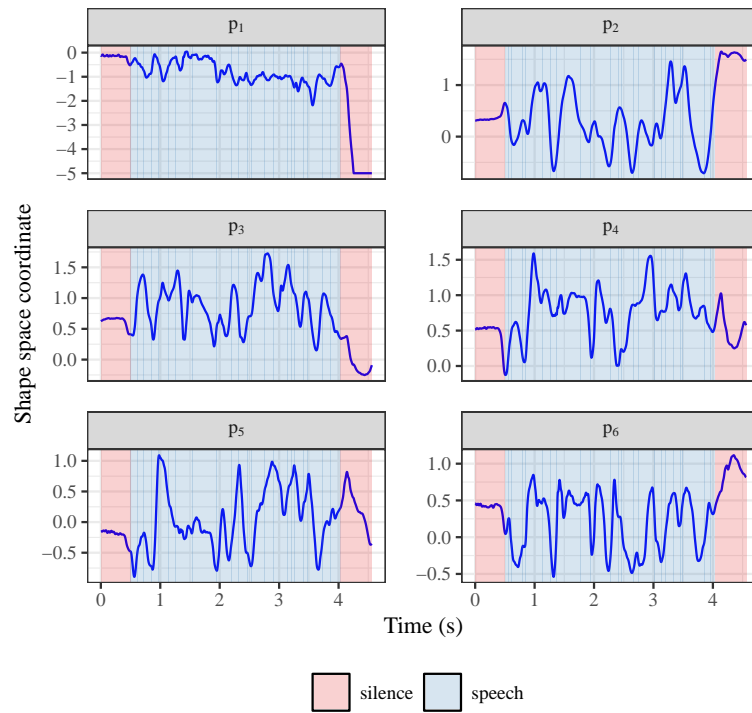


Figure 5.13.: Trajectories of tongue pose parameters over time for utterance  $s1\_0711$  in the fixed speaker case.



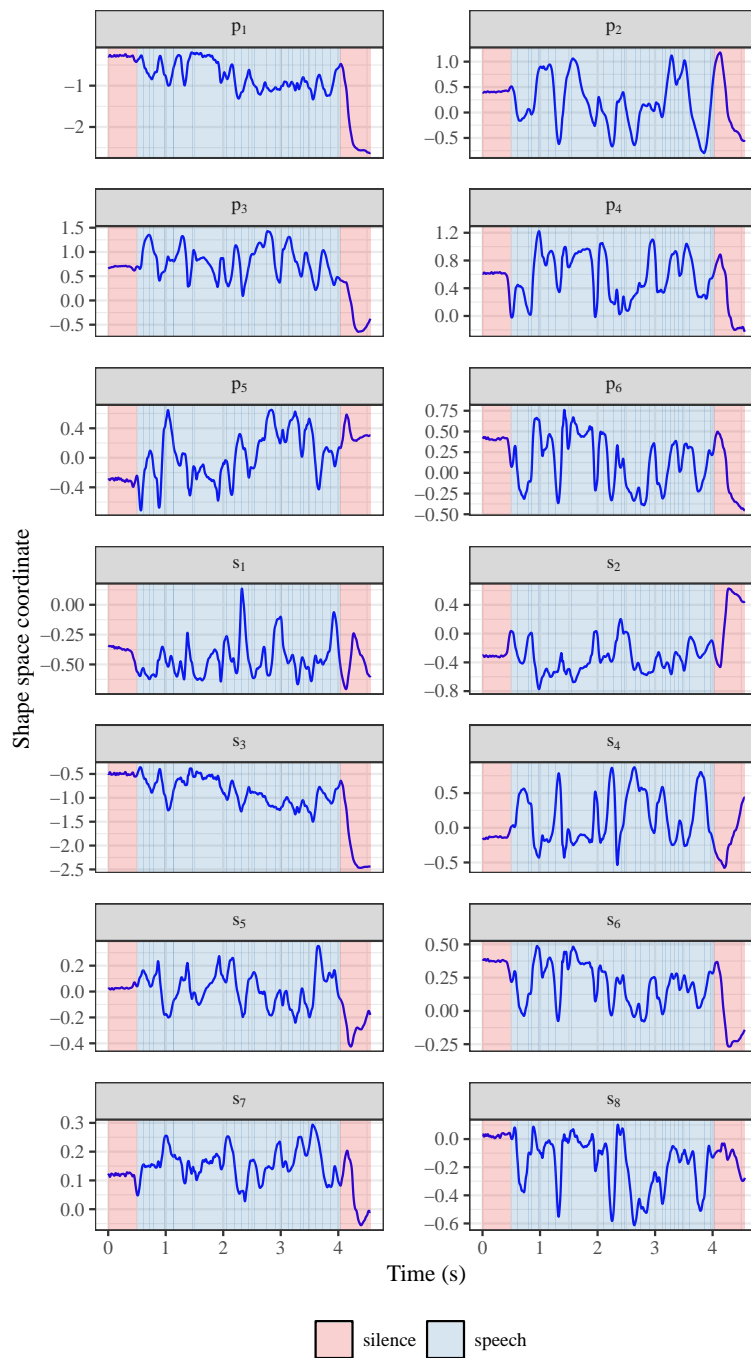


Figure 5.14.: Trajectories of all tongue model parameters over time for utterance  $s1\_0711$  in the full optimization case.

## 5. Registering sparse motion capture data

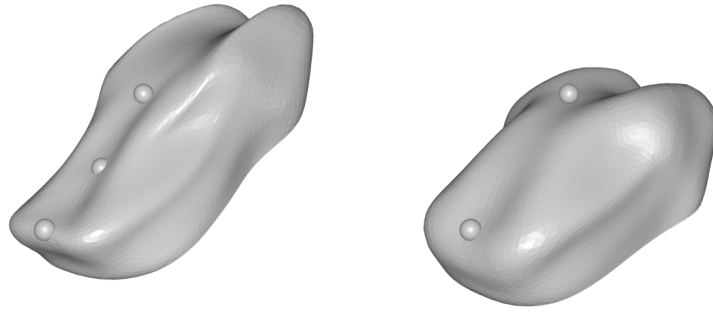


Figure 5.15.: Reconstructed tongue meshes from the obtained model parameters at time stamp 4.26 s for utterance *s1\_0711*. Results for the full (left) and fixed speaker (right) registration are shown. Spheres representing the EMA coil positions are added for reference.

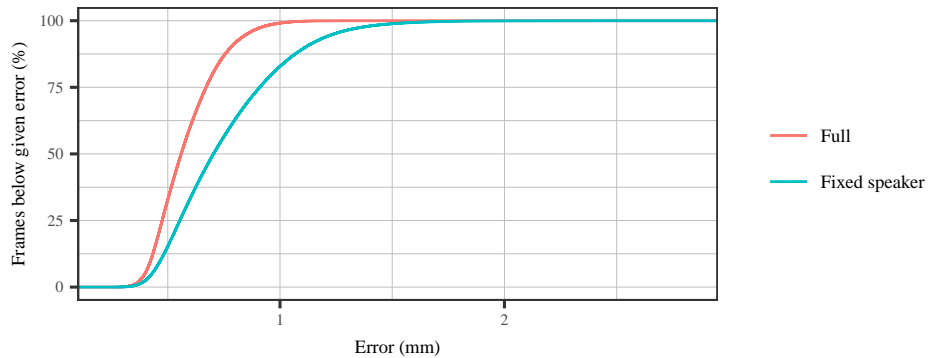


Figure 5.16.: Cumulative error for the two registrations for the *mngu0* data without silence intervals.

a rather unnatural appearance: it is very long, rather slim, and its back is raised very high. Here, it is important to note that this judgment is based on conjecture: the true shape of the tongue at this time frame is unknown because only three positions measured by EMA are available.

Now it is of interest to see if the error and the parameter behavior improve for the fixed speaker registration by removing such silence intervals from the evaluation. To this end, the silence intervals before and after an utterance are removed, silences occurring during an utterance are still preserved. The results of this operation are shown in Figure 5.16, Figure 5.17, and Table 5.6. The errors improved and the cumulative error plot reveals that the ones stemming from these silence intervals had only a minimal effect on the original evaluation.

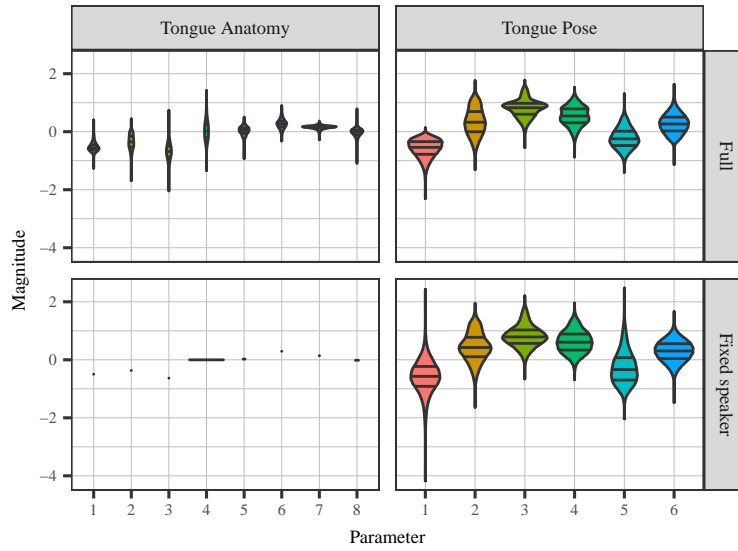


Figure 5.17.: Tongue model parameter distribution for the two registrations for the mngu0 data without silence intervals. The violin plots show the density, with the mean and interquartile range marked by horizontal lines.

	Experiment 1				Experiment 2			
	Mean	Std. dev.	Min	Max	Mean	Std. dev.	Min	Max
$s_1$	-0.53	0.16	-1.26	0.40	-0.50	0	-0.50	-0.50
$s_2$	-0.35	0.25	-1.69	0.44	-0.37	0	-0.37	-0.37
$s_3$	-0.69	0.31	-2.04	0.74	-0.63	0	-0.63	-0.63
$s_4$	0.04	0.33	-1.34	1.42	-0.00	0	-0.00	-0.00
$s_5$	0.04	0.14	-0.93	0.50	0.03	0	0.03	0.03
$s_6$	0.27	0.15	-0.32	0.89	0.29	0	0.29	0.29
$s_7$	0.15	0.06	-0.28	0.37	0.14	0	0.14	0.14
$s_8$	-0.04	0.16	-1.10	0.79	-0.02	0	-0.02	-0.02
$p_1$	-0.58	0.31	-2.31	0.13	-0.58	0.56	-4.17	2.43
$p_2$	0.34	0.47	-1.31	1.76	0.44	0.50	-1.64	1.94
$p_3$	0.80	0.31	-0.55	1.78	0.81	0.37	-0.66	2.20
$p_4$	0.54	0.33	-0.87	1.52	0.62	0.38	-0.68	1.95
$p_5$	-0.22	0.36	-1.40	1.31	-0.27	0.57	-2.03	2.47
$p_6$	0.27	0.33	-1.13	1.63	0.30	0.36	-1.47	1.66
Error (mm)	0.58	0.14	0.23	1.35	0.75	0.26	0.24	2.82

Table 5.6.: Statistics of tongue parameters and error for the two registrations of the mngu0 data without silence intervals.

## 5. Registering sparse motion capture data

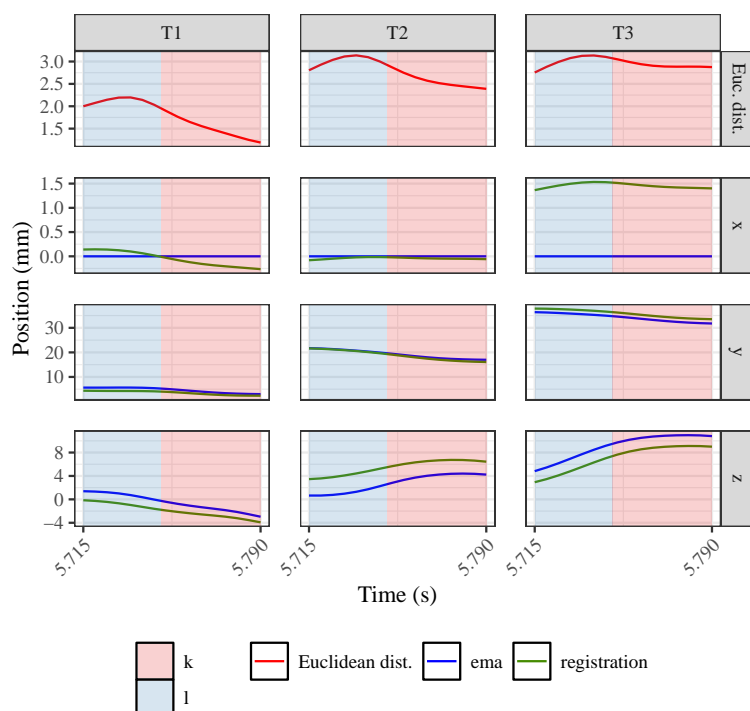


Figure 5.18.: Trajectories of EMA coils and corresponding vertices of registered tongue model for diphone segment [l\_k] in utterance *s1\_1233*. The top row shows the Euclidean distance between coil and corresponding vertex over time. The registration was obtained by optimizing only the tongue pose parameters and fixing the anatomy.

### Diphone analysis

While the maximum for the error reduced to 2.82 mm and the range of tongue pose parameters became smaller, the minimum  $-4.17$  of  $p_1$  is still very low, which requires further investigation. It turns out that this specific value occurs in the utterance *s1\_1233*:

“The spokesman explained that most of the easy oil had already been extracted and that now oil companies were having to look at ways of removing the more inaccessible mineral deposits.”

Figures 5.18 and 5.19 reveal that this behavior of the tongue model happens during a speech interval, namely the diphone [l\_k] in “oil companies”. The pose component  $p_1$  assumes a very low value and also the error becomes large. Like before, the blade and back coil are problematic. This can also be seen in Figure 5.20 where the generated mesh and the corresponding coils at this time frame are shown: the blade coil is below the mesh surface. This could be related to the speaker’s using a vocal tract configuration for the phone [ɬ] to produce [l] in this context. In the case of [ɬ], the configuration of [l] is altered by raising the tongue body. For the diphone [l\_k], this serves the purpose

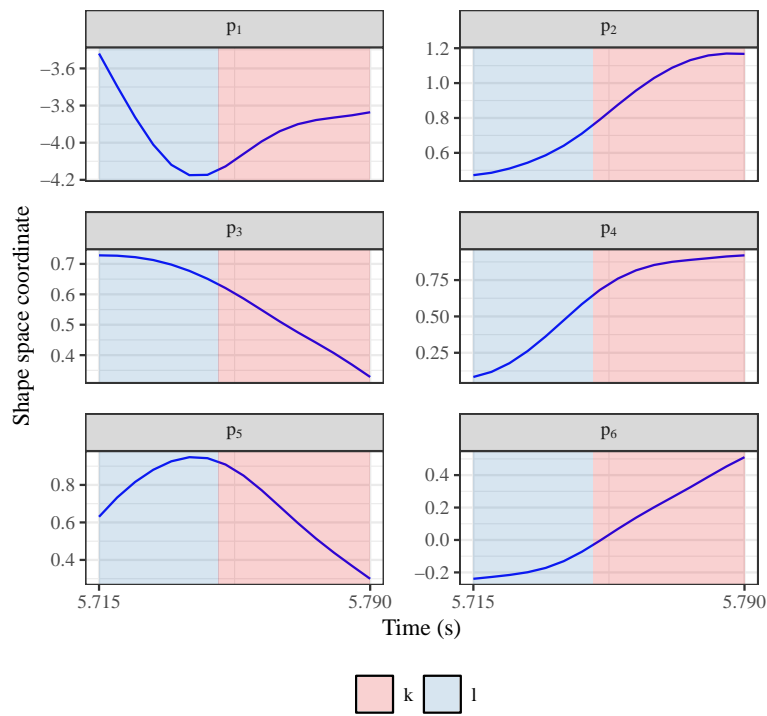


Figure 5.19.: Trajectories of tongue model parameters over time for diphone segment [l\_k] in utterance *s1\_1233*. The registration was obtained by optimizing only the tongue pose parameters and fixing the anatomy.

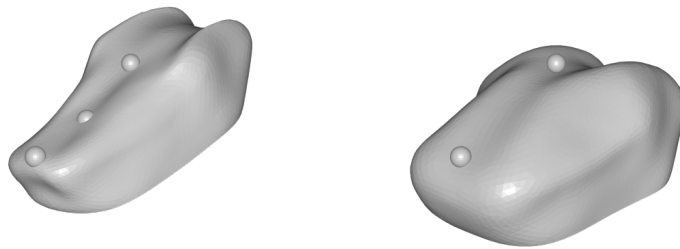


Figure 5.20.: Reconstructed tongue meshes from the obtained model parameters at time stamp 5.745s for utterance *s1\_1233*. Results for the full (left) and fixed speaker (right) registration are shown. Spheres representing the EMA coil positions are added for reference.

## 5. Registering sparse motion capture data

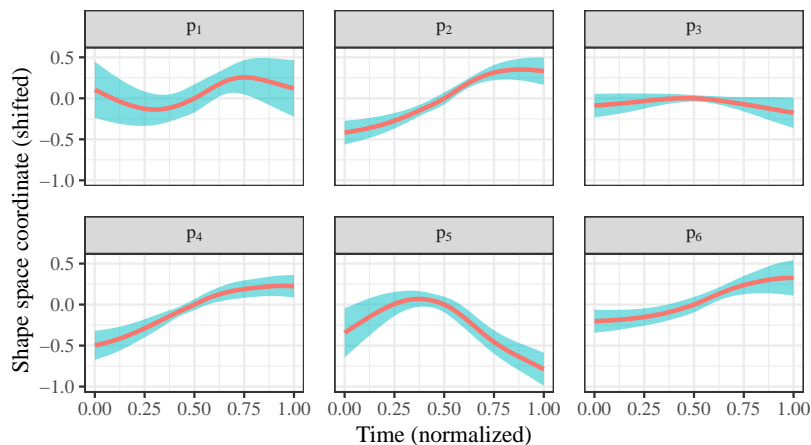


Figure 5.21.: Analysis of the tongue pose trajectories for the diphone [l\_k]. Plots show the mean trajectory (line) and the standard deviation (ribbon).

of bringing the tongue body into position for producing the dorsal consonant [k] that requires a contact between tongue back and palate. On the whole, this creates a tongue shape in this time frame, which might be unknown to the tongue model. The generated mesh in Figure 5.20 for the full registration appears again unnatural.

An analysis of the label data shows that 43 instances of the diphone [l\_k] are present in the considered dataset. In order to investigate how the pose parameters behave in general for this diphone, the results for the fixed speaker registration are processed as follows. First, the individual instances are time normalized, such that the diphone starts at time 0, the second phone of the diphone starts at 0.5, and the diphone ends at time 1. Then, the instances are interpolated with natural splines and the results are sampled at equidistant points in order to compare the individual instances of the diphone. As only the shape of the trajectories is of interest during this evaluation, any translational differences have to be removed. To this end, the mean trajectory of each parameter is computed. Afterwards, the individual parameter trajectories are shifted along the  $y$ -axis in order to minimize the distance to the corresponding mean trajectory. Results of this analysis are shown in Figure 5.21. These results provide valuable information: the behavior of the tongue pose parameters seems to be consistent near time 0.5 for different instances of the diphone, which is reflected by small standard deviations. At the boundaries of the diphone, the standard deviations increase. In this context, the standard deviation for  $p_1$  is large compared to the other parameters. The small sample size might be the reason for this result.

This observation sparks the idea of performing this kind of analysis also on other diphones. To this end, the three most frequent diphones in the dataset are investigated:

1. [ð\_ə] (693 instances)
2. [t\_ə] (582 instances)

## 3. [ə\_n] (523 instances)

Here, only instances are counted where at least one sample per participating phone is available in the diphone. Samples might be missing due to the acquisition rate of the EMA device.

Results are shown in Figure 5.22 where the same observation can be made like in the [l\_k] case: the shape of the trajectories is consistent near the center of the diphone, i.e., the standard deviation of the shape is small. Near the boundaries of the diphone, the standard deviation becomes larger. This could be related to the context of the diphone: at the boundary of the diphone area, the tongue shape is influenced by the adjacent phone next to the current diphone. This effect is also known as coarticulation.

On the whole, the obtained shape space coordinates may be useful for analyzing the typical trajectories of diphones during speech production. This means, the diphone data could be analyzed to train a model for synthesizing diphone trajectories in the pose parameter space. Afterwards, these trajectories could be transferred to an arbitrary speaker to create animations of the entire tongue surface by adapting the speaker parameters accordingly. Furthermore, the trajectories in the pose parameter space seem to display patterns that could be used to detect diphones or articulatory gestures.

### 5.7.3. Tübingen dataset

The Tübingen dataset was recorded as part of a pilot study at the Max Planck Institute for Intelligent Systems in Tübingen. This dataset consists of synchronized audio, EMA, and face capture recordings. The EMA data was acquired at a sampling rate of 400 Hz by using an NDI Wave articulograph. Face capturing was performed by an active stereo system (3dMD LLC, Atlanta, GA). The system uses six color cameras, six gray-scale stereo camera pairs, five speckle pattern projectors, and six white LED panels to capture geometry and texture at 60 fps. In order to avoid facial occlusions, the articulograph was positioned above the head of the recorded subject, which is shown in Figure 5.1a. Additionally, registrations for the face capture data are available that were obtained by using a sequential registration approach (T. Li et al., 2017).

In total, 3 male speakers were recorded: sp01, sp02, and sp03. The data of sp01 is missing a lot of EMA recording material due to difficulties in gluing the sensor coils to the tongue. Thus, this speaker is ignored in this experiment. The speech material consists of 80 English sentences of the TIMIT corpus (Garofolo et al., 1993). Figure 5.23 shows the EMA coil layout of this study. Here, it becomes clear that this dataset focuses on the front part of the tongue including tip and blade. This experiment only takes the mid-sagittal coils into account.

Currently, the dataset is only available for internal use.

### Preprocessing

For this dataset, all preprocessing steps are performed. The dataset provides a recording for the reference point that is required for mapping the data to the origin of the tongue

5. Registering sparse motion capture data

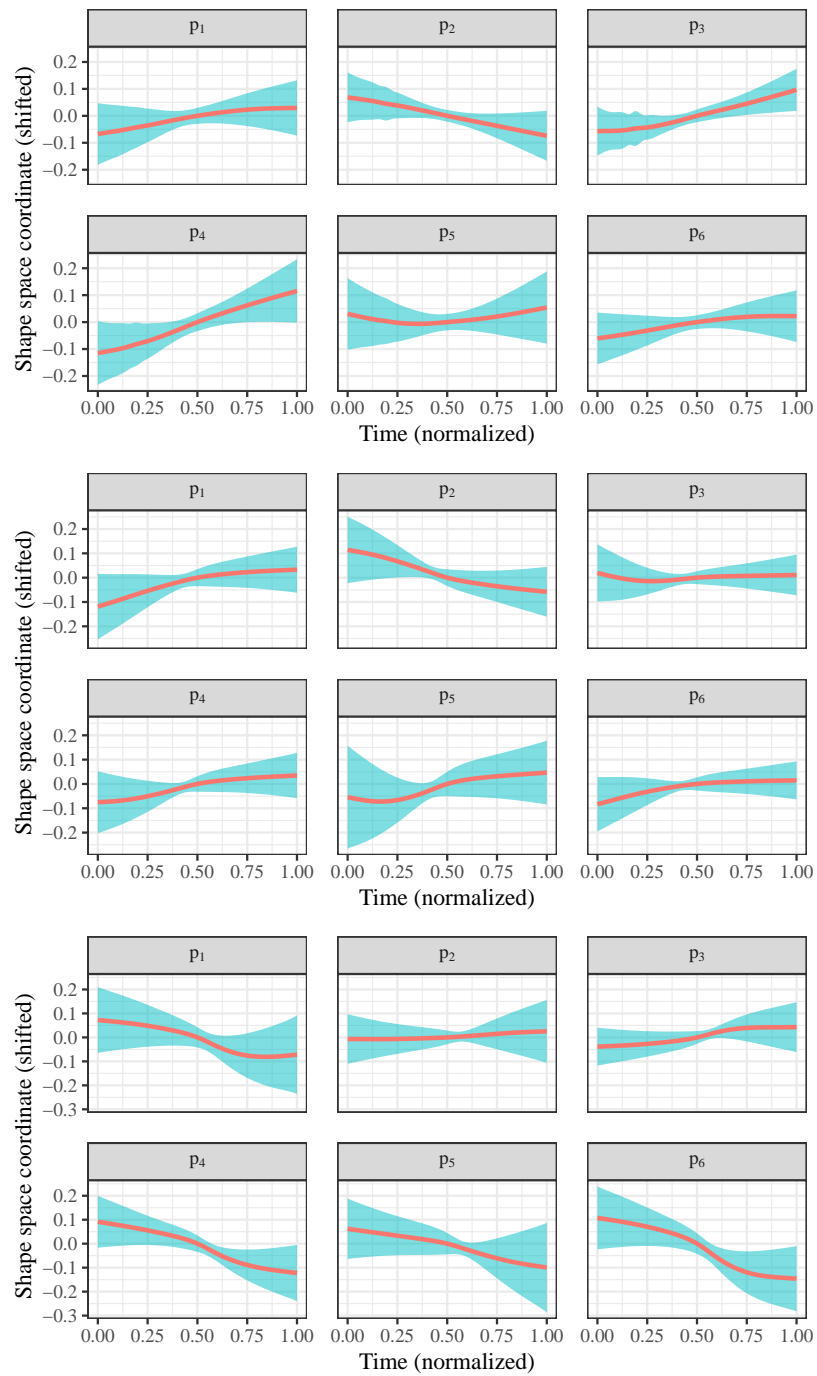


Figure 5.22.: Analysis of the tongue pose trajectories for the diphones [ð\_ə] (top), [t\_ə] (center), and [ə\_n] (bottom). Plots show the mean trajectory (line) and the standard deviation (ribbon).



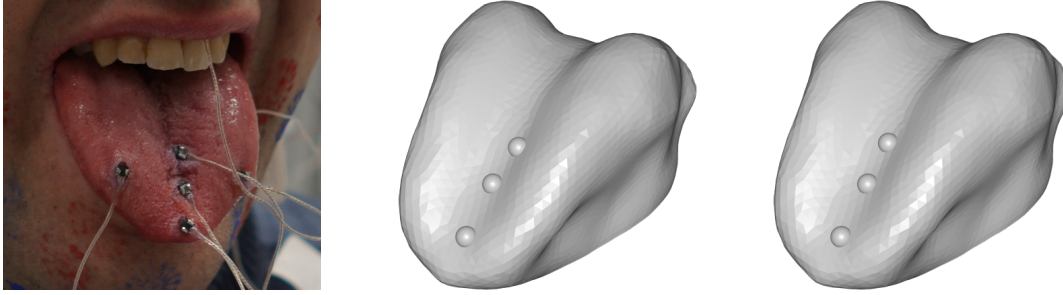


Figure 5.23.: Photograph of tongue coil layout for subject sp02 of the Tübingen dataset is shown on the left. Combinatorial correspondence optimization result for mid-sagittal coils is provided for subjects sp02 (center) and sp03 (right).

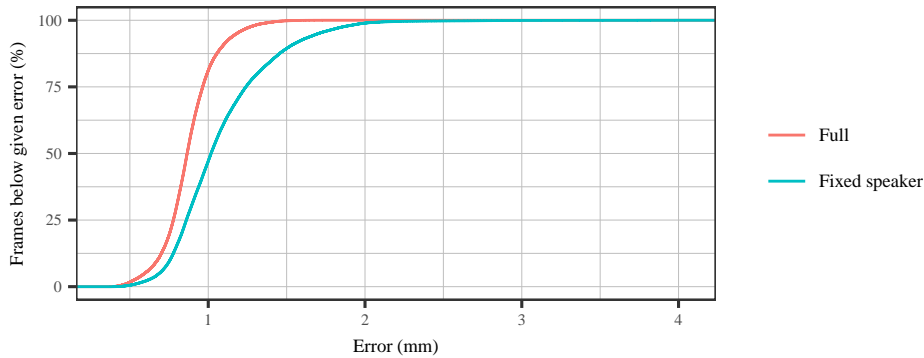


Figure 5.24.: Cumulative error for the two registrations of sp02 in the Tübingen dataset.

model. For estimating the correspondence between tongue coils and tongue model vertices, the vertex subsets *tongue tip* and *tongue blade* were used. The subset *tongue blade* was used for the two coils located there. For speaker sp02, the prior box width  $h = 1$  was used. In the case of sp03, a width of  $h = 0.5$  was applied. Correspondence results of the combinatorial approach are shown in Figure 5.23. It can be seen that the found vertex-coil correspondences for the two speakers are nearly identical, they only differ with respect to the correspondence for the back coil.

## Results

The results for sp02 can be seen in Figure 5.24, Figure 5.25, and Table 5.7. The ones for sp03 are depicted in Figure 5.26, Figure 5.27, and Table 5.8.

On the whole, the same observations can be made as in the mngu0 case: the errors and variance of the tongue pose parameters increase for the fixed speaker registration approach. For sp02, nearly all errors are below 1.5 mm in the full optimization case. For sp03, the results are better: here the errors are mostly below 1 mm. In the fixed speaker case, the errors for sp02 are mostly below 2 mm. Nearly all errors for speaker

## 5. Registering sparse motion capture data

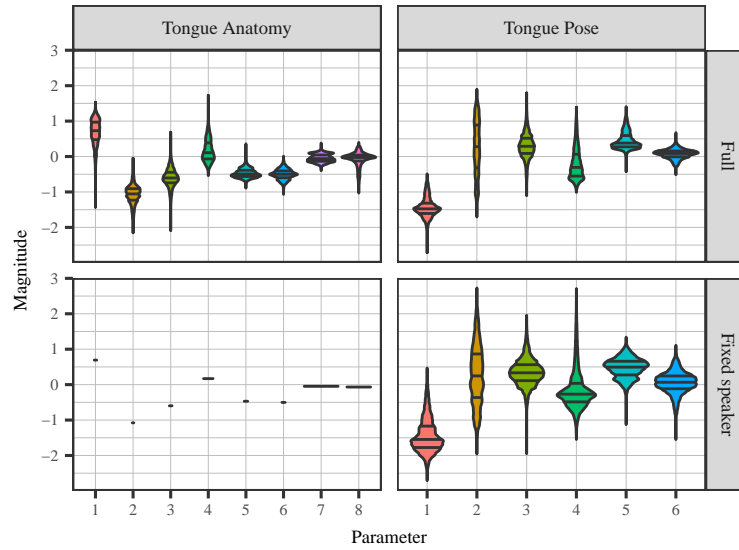


Figure 5.25.: Tongue model parameter distribution for the two registrations of sp02 in the Tübingen dataset. The violin plots show the density, with the mean and interquartile range marked by horizontal lines.

	Experiment 1				Experiment 2			
	Mean	Std. dev.	Min	Max	Mean	Std. dev.	Min	Max
$s_1$	0.69	0.36	-1.43	1.53	0.69	0	0.69	0.69
$s_2$	-1.08	0.27	-2.15	-0.05	-1.08	0	-1.08	-1.08
$s_3$	-0.60	0.26	-2.08	0.69	-0.60	0	-0.60	-0.60
$s_4$	0.17	0.31	-0.54	1.73	0.17	0	0.17	0.17
$s_5$	-0.47	0.13	-0.89	0.36	-0.47	0	-0.47	-0.47
$s_6$	-0.50	0.14	-1.07	0.01	-0.50	0	-0.50	-0.50
$s_7$	-0.05	0.11	-0.40	0.38	-0.05	0	-0.05	-0.05
$s_8$	-0.07	0.20	-1.03	0.40	-0.07	0	-0.07	-0.07
$p_1$	-1.44	0.27	-2.72	-0.49	-1.44	0.48	-2.71	0.47
$p_2$	0.25	0.76	-1.71	1.89	0.27	0.83	-1.96	2.73
$p_3$	0.30	0.34	-1.11	1.79	0.35	0.37	-1.95	1.96
$p_4$	-0.21	0.43	-1.01	1.39	-0.15	0.54	-1.55	2.70
$p_5$	0.45	0.24	-0.43	1.40	0.47	0.28	-1.13	1.34
$p_6$	0.07	0.13	-0.51	0.67	0.06	0.31	-1.55	1.11
Error (mm)	0.87	0.17	0.34	1.79	1.08	0.32	0.37	4.05

Table 5.7.: Statistics of tongue parameters and error for the two registrations of sp02 in the Tübingen dataset.

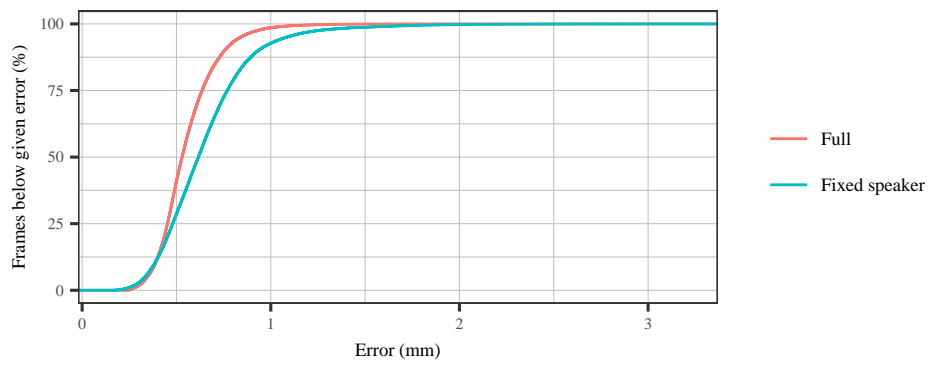


Figure 5.26.: Cumulative error for the two registrations of sp03 in the Tübingen dataset.

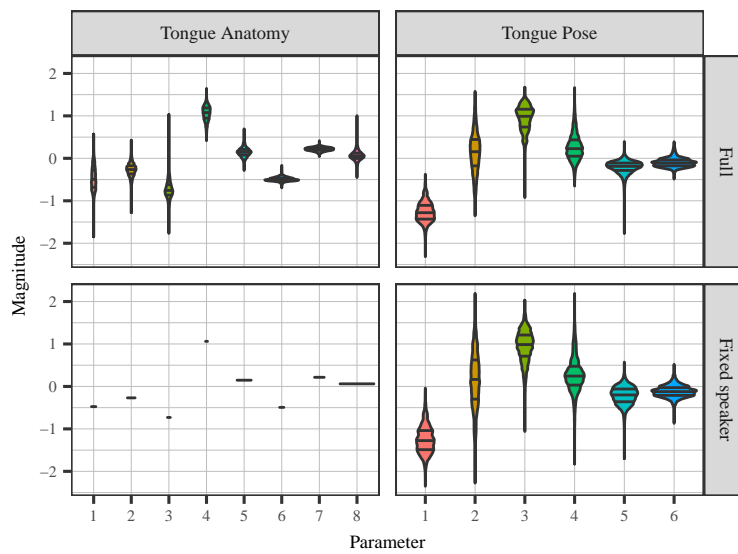


Figure 5.27.: Tongue model parameter distribution for the two registrations of sp03 in the Tübingen dataset. The violin plots show the density, with the mean and interquartile range marked by horizontal lines.

## 5. Registering sparse motion capture data

	Experiment 1				Experiment 2			
	Mean	Std. dev.	Min	Max	Mean	Std. dev.	Min	Max
$s_1$	-0.48	0.28	-1.85	0.57	-0.48	0	-0.48	-0.48
$s_2$	-0.27	0.16	-1.28	0.43	-0.27	0	-0.27	-0.27
$s_3$	-0.73	0.22	-1.76	1.03	-0.73	0	-0.73	-0.73
$s_4$	1.06	0.17	0.42	1.65	1.06	0	1.06	1.06
$s_5$	0.15	0.11	-0.29	0.69	0.15	0	0.15	0.15
$s_6$	-0.49	0.05	-0.70	-0.17	-0.49	0	-0.49	-0.49
$s_7$	0.22	0.04	0.04	0.41	0.22	0	0.22	0.22
$s_8$	0.06	0.12	-0.45	1.00	0.06	0	0.06	0.06
$p_1$	-1.26	0.23	-2.32	-0.38	-1.25	0.32	-2.35	-0.04
$p_2$	0.14	0.48	-1.35	1.58	0.16	0.68	-2.28	2.19
$p_3$	0.93	0.31	-0.92	1.67	0.94	0.38	-1.06	2.03
$p_4$	0.26	0.31	-0.65	1.67	0.28	0.41	-1.83	2.19
$p_5$	-0.21	0.15	-1.76	0.40	-0.22	0.22	-1.70	0.57
$p_6$	-0.10	0.11	-0.49	0.39	-0.12	0.14	-0.87	0.52
Error (mm)	0.56	0.16	0.16	1.97	0.65	0.26	0.13	3.22

Table 5.8.: Statistics of tongue parameters and error for the two registrations of sp03 in the Tübingen dataset.

sp03 are below 1.5 mm for this registration. In contrast to the mngu0 registration results, anomalies in the model parameter distribution are absent, i.e., the parameters stay within reasonable intervals.

### Fusing tongue and face animation

The face registrations of the dataset also provide access to the vertex locations where the EMA reference coils were located. This information can be used to map the obtained tongue meshes to the corresponding face mesh. Of course, the tongue registrations have to be downsampled first to match the sampling rate of the face scanner.

Basically, the mapping between tongue and face mesh may then be performed as follows: all transformations that were applied to the EMA data at a specific time step up to and excluding the mapping to the local coordinate system are inverted and applied to the corresponding reconstructed tongue mesh. Then a local coordinate system is built by using the positions of the vertices on the face mesh that correspond to the locations of the reference coils. A mapping to this coordinate system is inverted and applied to the transformed tongue mesh. In theory, this operation would move the tongue to the correct position in the face mesh. In practice, however, this approach produced suboptimal results. One reason might be that the correspondences between face vertex and reference coil are incorrect or the positions are degraded due to measurement noise.

Instead, the following simplified approach is used: first, the result of the original approach is computed. Second, the untransformed tongue mesh is rotated around the  $x$ -axis by  $-90^\circ$  to match the orientation of the face mesh. Then, the center of the rotated mesh is shifted to the one of the result of the original approach. Finally, the shifted mesh is translated along the  $x$ -axis to lie in the center of the mouth. This shift is set manually and is constant for all samples of a speaker.

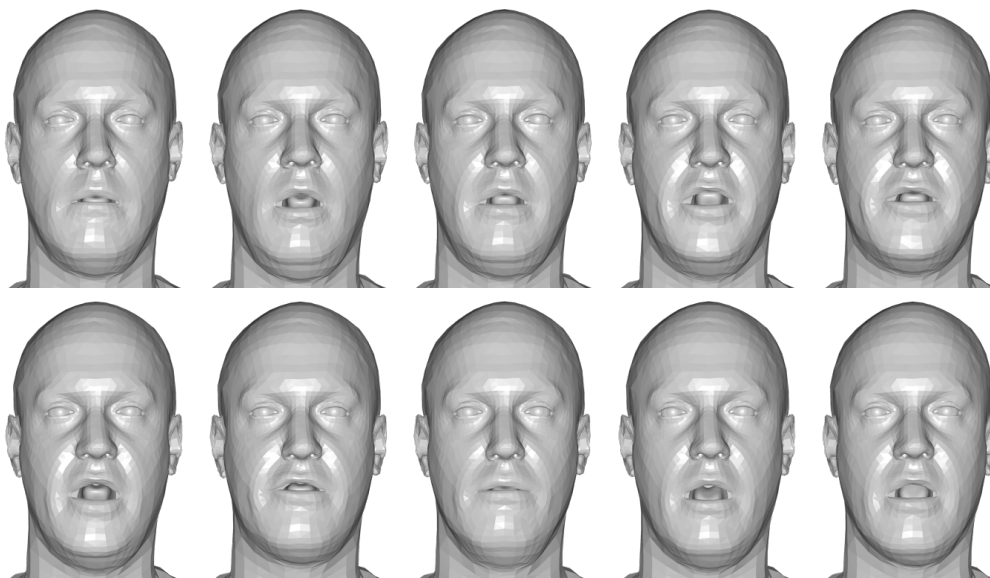


Figure 5.28.: Example frames of an animation showing subject sp02 created by fusing multimodal registration results of face and tongue. Frames belong to prompt 27 of the Tübingen dataset: “Did dad do academic bidding?”

These operations make it possible to fuse tongue and face meshes into a single animation. Results of such a fusing are shown in Figure 5.28. Despite the lack of the teeth in the visualization, the appearance of the presented frames looks acceptable. Of course, a subjective study is needed to verify this claim.

#### 5.7.4. Trier dataset

The Trier dataset is a multimodal database of articulatory data for analyzing the effect of posture and acoustical noise on articulation (Steiner, Knopp, et al., 2014). It consists of audio, video, ultrasound tongue imaging (UTI), and EMA recordings. Furthermore, for some speakers, 3D intraoral scans are available that were obtained by means of a 3shape TRIOS scanner<sup>4</sup>. In total, data of 7 speakers (3 female and 4 male) was acquired. In terms of speech material, it focuses on the German language where the inventory can be summarized as follows:

- sustained vowels and diphthongs
- consonant phonemes of German in a [aCa] context where C is a German consonant
- consonant-vowel repetitions
- German translation of the “Northwind and the Sun” passage, a standard specimen in phonetic research (International Phonetic Association, 1999)

---

<sup>4</sup><https://www.3shape.com>

## 5. Registering sparse motion capture data

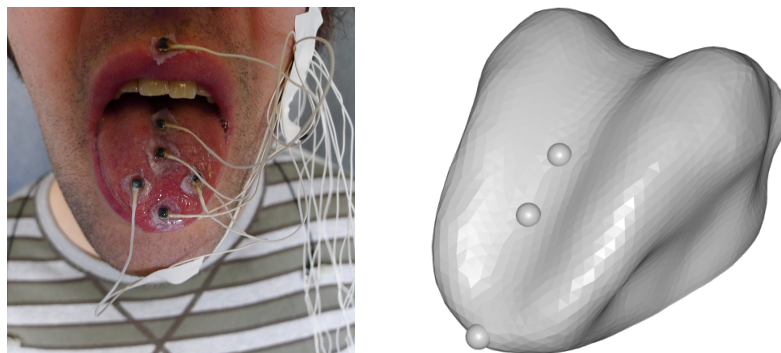


Figure 5.29.: Photograph of tongue coil layout for subject VP08 of the Trier dataset (left) and rendering of combinatorial correspondence optimization result of mid-sagittal coils for subject VP05 (right). Spheres on tongue mesh highlight the vertices that were found.

- 10 utterances designed to study German vowels

Furthermore, palate traces were acquired for the participants. The EMA was obtained by using an AG501 with a sample rate of 250 Hz. The recording uses a five-coil layout for the tongue: along the mid-sagittal plane, coils were placed at the tongue tip, the blade, and at the dorsum. In contrast to the *mngu0* dataset, it also uses two lateral coils on either side of the tongue blade. These lateral coils are ignored here. This layout can be inspected in Figure 5.29. Currently, this data is unavailable for public use.

The most interesting property of this dataset is given by the fact that it provides access to articulatory data of the German language. This implies that a tongue model derived from speakers of American English is used to register dynamic speech production data of another language. However, the phonetic inventories of the English and German languages are very similar. Thus, the pose parameter space of the tongue model should be compatible with motion capture data originating from producing German speech.

In this experiment, all data of the female speaker *VP05* is used that was acquired in upright position without acoustical noise.

### Preprocessing

For this data, all preprocessing steps are performed. Like in the *mngu0* case, the dataset is missing a recording for the reference point required for mapping the data to the origin of the tongue model. The recorded palate traces showed some inconsistencies and were therefore ignored. Instead, the palate surface was estimated like in the *mngu0* case. For estimating the correspondence between tongue coils and tongue model vertices, the vertex subsets *tongue tip*, *tongue body*, and *tongue back* were used. This time, the width  $h = 1$  was used for the prior box. The combinatorial approach produced the result depicted in Figure 5.29.

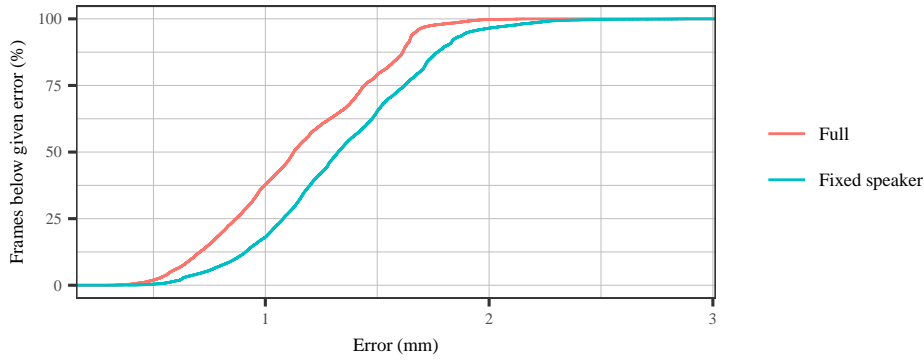


Figure 5.30.: Cumulative error for the two registrations of VP05 in the Trier dataset.

	Experiment 1				Experiment 2			
	Mean	Std. dev.	Min	Max	Mean	Std. dev.	Min	Max
$s_1$	-0.77	0.47	-2.13	0.35	-0.77	0	-0.77	-0.77
$s_2$	-0.70	0.23	-1.21	0.05	-0.70	0	-0.70	-0.70
$s_3$	-0.81	0.22	-1.69	-0.29	-0.81	0	-0.81	-0.81
$s_4$	-0.45	0.36	-1.59	0.61	-0.45	0	-0.45	-0.45
$s_5$	-0.08	0.21	-0.71	0.71	-0.08	0	-0.08	-0.08
$s_6$	0.79	0.20	-0.21	1.20	0.79	0	0.79	0.79
$s_7$	-0.05	0.21	-0.93	0.70	-0.05	0	-0.05	-0.05
$s_8$	-0.86	0.26	-1.66	-0.06	-0.86	0	-0.86	-0.86
$p_1$	-1.47	0.11	-2.16	-1.10	-1.45	0.12	-1.80	-0.89
$p_2$	0.85	0.40	-0.98	1.49	0.89	0.64	-1.48	1.73
$p_3$	0.90	0.28	-0.44	1.69	0.95	0.25	-0.14	1.68
$p_4$	0.69	0.31	-1.01	1.40	0.78	0.29	-0.22	1.86
$p_5$	-0.23	0.53	-1.97	1.14	-0.21	0.96	-2.79	2.74
$p_6$	0.88	0.46	-0.23	2.65	0.80	0.41	-0.73	1.73
Error (mm)	1.15	0.36	0.28	2.16	1.34	0.37	0.31	2.88

Table 5.9.: Statistics of tongue parameters and error for the two registrations of speaker VP05 in the Trier dataset.

## Results

The evaluation results are provided in Figure 5.30, Figure 5.31, and Table 5.9.

The results for the full optimization show acceptable errors where 62% of the errors are below 1.25 mm. Again, the errors increase if the anatomy is fixed: now, 62% of the errors are below 1.5 mm. Once more, the variance of the pose parameters increases in the fixed speaker experiment. Like for the data of the Tübingen dataset, anomalies are absent in the parameter distribution.

## 5. Registering sparse motion capture data

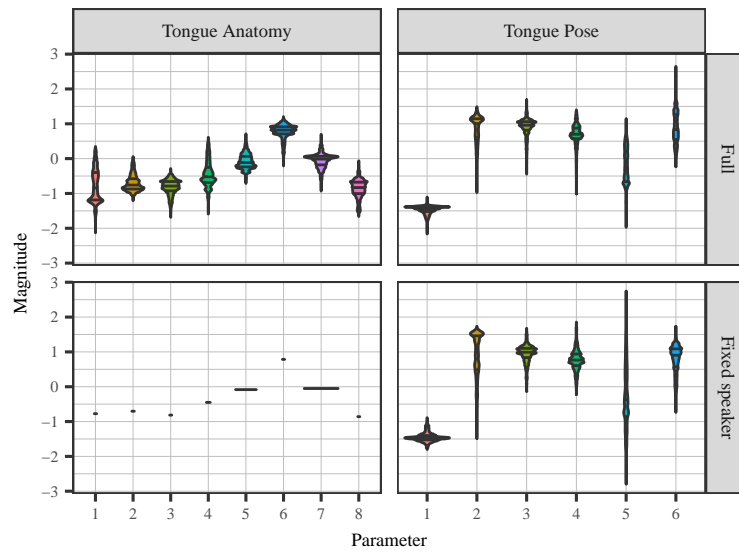


Figure 5.31.: Tongue model parameter distribution for the two registrations of VP05 in the Trier dataset. The violin plots show the density, with the mean and interquartile range marked by horizontal lines.

## 5.8. Conclusion

This chapter described an approach for visualizing sparse articulatory data by using a multilinear tongue model. In particular, preprocessing steps were described that have to be applied to EMA data before a registration can take place. Afterwards, the registration approach was described that roughly consists of two steps: one step registers the data while using all parameters of the tongue model. The second step registers the data again while freezing the speaker parameters to the estimated anatomy of the corresponding speaker.

The performance of the approach was assessed in several experiments. In total, EMA data of 4 different speakers was registered during the experiments. The performed experiments provided valuable insights: first of all, the proposed approach seems to be able to register unknown speakers and even data from a language unknown to the tongue model. To a large extent, the experiments used the same settings to register the data. Thus, the presented registration framework seems to be independent of the speaker and the used articulo-graph.

Another observation was that fixing the speaker parameters during registration provided access to a shape sequence that is temporally consistent.

Moreover, the registration provided access to a representation of the EMA in the form of trajectories in the parameter space of the model. These trajectories seem to be consistent for diphones, which might make this representation useful for further studies.

Finally, the tongue meshes obtained from the registration could be readily combined with face meshes originating from face capture data registration, which led to animations



of the tongue and face. Here, the perceptive quality of these animations still has to be evaluated.

However, it is important to note that it is difficult to evaluate from these tracking experiments how close the estimated tongue shape is to the true tongue of the speakers, 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. Such an evaluation might make use of an additional modality, such as UTI or real-time magnetic resonance imaging (rtMRI), to determine parts of the true tongue contour.



## 6. Multimodal speech synthesis

### 6.1. Introduction

#### 6.1.1. Motivation

The previous chapter presented a viable way for estimating the full three-dimensional (3D) tongue shape from articulatory motion capture data. This data has a very high temporal resolution, but offers only very sparse sample points in the spatial domain. Being able to reconstruct the tongue shape from such data offers the ability to visualize such a modality in a way that is easier to interpret and understand than the original representation. Such animations can be used, e.g., in computer-aided pronunciation training (CAPT) software to show to a language learner how to move the tongue to produce a specific sound or word. So far, the shape of the tongue was estimated from recorded electromagnetic articulography (EMA) data, which represents a limitation for such systems: only words or phones that were recorded can be shown to the user. Additionally, the anatomy of the tongue and the articulation strategy were derived from the used EMA data, which might be different from the current user of the software. Thus, an approach is needed that, on the one hand, is able to synthesize tongue motions from given text in order to produce animations for words that are missing in the underlying EMA database. On the other hand, it should be possible to adapt the anatomy to the target speaker's anatomy. Furthermore, it would be desirable to provide synchronized audio for such a synthesized animation to provide the user with a correspondence between tongue motion and produced speech. Such a system would also be helpful for other scenarios where audiovisual speech synthesis is required, like for example virtual avatars which can be equipped with realistic tongue motions.

In the area of speech technology, text-to-speech (TTS) is an active field of research. Here, approaches are investigated how to best synthesize speech from text that sounds natural. In particular, statistical parametric synthesis strategies have been proven to be suitable for addressing this task. A standard technique of this class is given by the HMM/DNN-based speech synthesis system (HTS) framework that was first presented by Zen and Toda (2005). Originally, it was based on a Gaussian mixed model/decision tree modeling paradigm. In the meantime, deep neural networks have proven to be advancing statistical TTS by replacing this paradigm with a feed forward deep neural network framework.

A key observation is now that these systems parameterize the recorded speech and derive the model from the resulting parameters. This operation is similar to the registration of the EMA data with the tongue model: the articulatory data is parameterized in the form of the tongue model parameters  $\mathbf{s}$  and  $\mathbf{p}$  during such a registration. This

## 6. Multimodal speech synthesis

observation sparks the idea of combining the obtained speech parameters with the ones from the articulatory data in an HTS framework.

### 6.1.2. Related work

In the area of speech science, there is a growing body of work on application-oriented research to combine articulatory data, and features derived from it, with speech technology applications, such as to recover articulatory movements from the acoustic signal (articulatory inversion mapping, cf. King et al. (2007) and Mitra et al. (2011) for examples), provide articulatory control for reactive TTS synthesis (e.g., Astrinaki et al. (2013) and Ling, Richmond, and Yamagishi (2013)), or predict sparse articulatory movements from a symbolic representation (e.g., Ling, Richmond, and Yamagishi (2010a) and Cai et al. (2015)).

Previous studies (e.g., Engwall, 2002; Fagel and Clemens, 2004), combined intraoral motion capture data obtained from EMA (Schönle et al., 1987) with concatenative speech synthesis to animate a geometric tongue model simultaneously with synthesized audio. Other approaches (e.g., Ben Youssef (2011)) for hidden Markov model (HMM) based TTS with intraoral animation also rely on acoustic-articulatory inversion mapping. Among more recent implementations, the statistical parametric speech synthesis paradigm introduces greater flexibility in the modeling and therefore in the combination of multiple modalities. Consequently, several studies (Ling, Richmond, Yamagishi, and Wang, 2009; Ling, Richmond, and Yamagishi, 2010a; Ling, Richmond, and Yamagishi, 2010b) have successfully used HMM based multimodal speech synthesis with EMA data. However, a study is missing that has presented an end-to-end system to directly synthesize acoustics and the motion of a full 3D model of the tongue surface from text using statistical parametric speech synthesis, particularly with a tongue model that can be easily adapted to the anatomy of different speakers.

### 6.1.3. Contribution

The main contribution presented in this chapter is an end-to-end TTS system that synthesizes full 3D tongue animations synchronized with synthesized audio. In particular, it summarizes the findings of the publications:

Le Maguer, Sébastien, Ingmar Steiner, and Alexander Hewer (Aug. 2017). “An HMM/DNN comparison for synchronized text-to-speech and tongue motion synthesis”. In: *Interspeech*. Stockholm, Sweden, pp. 239–243. DOI: 10.21437/Interspeech.2017-936.

Steiner, Ingmar, Sébastien Le Maguer, and Alexander Hewer (Dec. 2017). “Synthesis of tongue motion and acoustics from text using a multimodal articulatory database”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12, pp. 2351–2361. DOI: 10.1109/TASLP.2017.2756818.

In this context, it is important to note that the above papers used a prototype version of a tongue model derived from the Ultrax dataset and an earlier version of the previously discussed EMA registration approach. The chapter therefore extends the previous work

by reproducing the most relevant experiment with the final USC tongue model and the current registration approach.

#### 6.1.4. Overview

The chapter is organized as follows: first, the HTS framework is presented, a standard approach for synthesizing speech from text. This framework is then extended to synthesize articulatory data from text. Before turning to the experiments, the tongue model and the EMA registration strategy used in the given studies are presented and discussed. This is necessary because they are different to the ones presented earlier in this work. Afterwards, experiments are conducted and different variants of the aforementioned extension of the HTS framework are evaluated. In this section, the most relevant experiment is reproduced by using the presented EMA registration strategy of the previous chapter and the final USC model. Finally, the conclusion outlines the findings of the experiments and provides ideas for possible future work.

## 6.2. Adapting the HTS framework

### 6.2.1. Standard approach

The HTS framework proposed by Zen and Toda (2005) is a standard statistical parametric speech synthesis system. Its default architecture consists of the four parts:

1. parameterization of the audio signal
2. training of the models
3. parameter generation from the models
4. rendering of the signal from generated parameters

The parameterization of the signal can be performed using any suitable signal processing tool, as long as it is kept consistent with the signal rendering in the final step. In the standard procedure, this is generally accomplished by coupling STRAIGHT (Kawahara et al., 1999) with a mel log spectrum approximation (MLSA) filter (Fukada et al., 1992). First, STRAIGHT is used to extract the spectral envelope, the fundamental frequency ( $F_0$ ), and the aperiodicity. Generally, the  $F_0$  values are transformed into the logarithmic domain, to be more consistent with human hearing. In a final step, the MLSA filter is used to parameterize the coefficients used for the spectral envelope and the aperiodicity. The result of this operation are the mel-generalized cepstral coefficients (MGC) and the aperiodicity per band (BAP), respectively. This step is needed because the original number of coefficients is too high to be processed adequately.

The training step is either based on the standard HTS training paradigm proposed by Zen and Toda (2005), or on the default deep neural network (DNN) training described in Zen, Senior, et al. (2013). If the DNN training variant is used, the  $F_0$  trajectory is interpolated and the voiced/unvoiced property is extracted in order to respect the standard DNN training proposed by Zen, Senior, et al. (2013). The generation level consists of applying the algorithm presented by Tokuda et al. (2000).

### 6.2.2. Multimodal extension

The original version of the HTS approach is intended for performing the synthesis of a speech audio signal. However, it can easily be modified to synthesize other modalities. This is due to the fact that it uses the parameterization of the audio instead of the signal itself, which implies that also parameters stemming from other modalities might be used. For example, several studies (Ling, Richmond, Yamagishi, and Wang, 2009; Ling, Richmond, and Yamagishi, 2010b; Ling, Richmond, and Yamagishi, 2010a) used EMA data as an articulatory representation in a synthesis approach. That means that some steps of the original architecture may be modified:

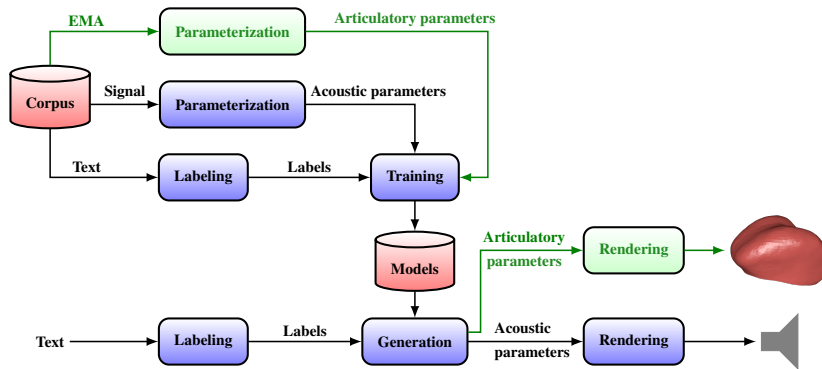
1. parameterization of the audio signal and corresponding articulatory data,
2. training of the models,
3. parameter generation from the models, and
4. rendering of the signal and articulatory data from the generated parameters.

Here, only the first and last step are adapted in order to take articulatory data into account. The new architecture would therefore be able to synthesize synchronized speech audio and articulatory data. In Figure 6.1a, the modified HTS system can be inspected.

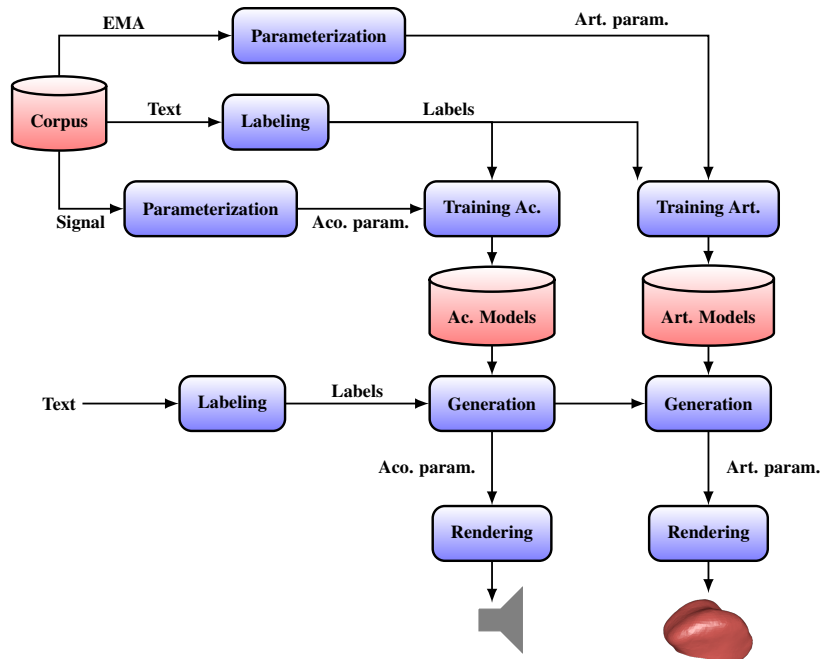
By considering that EMA data was already successfully used as a parameterization in an HMM scenario, it appears worthwhile to investigate if tongue model pose parameters  $\mathbf{p}$  are also a viable parameterization of articulatory data for this purpose. These parameters can be obtained from EMA data by performing a registration like the one presented in the previous chapter. In such a case, the system would be able to provide access to synthesized pose parameters from a given text, which afterwards can be turned into a full 3D animation of the tongue. Here, also the anatomy of the tongue can be adapted during the mesh generation process, which provides the advantage of producing a user-specific tongue shape.

### 6.2.3. Separating the articulatory model from the acoustical model

The original multimodal modeling in Figure 6.1a represents a significant drawback: it requires the system to be trained on a multimodal dataset, i.e., a dataset that contains speech audio data with synchronized articulatory data. However, the number of such articulatory databases is extremely small, which partly can be attributed to the required effort for creating such databases. On the other hand, conventional acoustic databases are widely available. In order to increase flexibility, the proposed system can thus be modified as shown in Figure 6.1b. This new variant decouples the training of the acoustical model from that of the articulatory model, which offers the following advantages: first of all, the acoustical model may now be derived from conventional acoustic databases and easily be combined with the articulatory model. This way, different acoustical models representing different voices can be incorporated into the framework without the requirement of having the articulatory data available for the respective speakers. Such a feature might be useful in a CAPT application where the user can then easily choose between different voices. Moreover, in this new variant, it is also possible to update the articulatory model separately from the acoustical one. This might occur if new articulatory data is available



(a) Architecture for multimodal based synthesis adapted from Zen and Toda, 2005; the multimodal extensions are highlighted.



(b) Adapted architecture for multimodal based synthesis where acoustical and articulatory components have been separated.

Figure 6.1.: Diagrams of the different architectures used in the experiments.

## 6. Multimodal speech synthesis

or if a new tongue model is used to register the data. In this context, it is important to ensure the synchronization between the acoustical and the articulatory trajectories. To do so, the durations produced by the acoustic generation stage are imposed onto the articulatory generation stage at the phone level. This behavior is represented in Figure 6.1b by an arrow connecting the acoustical generation step with the one of the articulatory data.

### 6.3. Experiments

#### 6.3.1. Overview

The previous sections described two ideas for a multimodal extension of the HTS framework. This section is dedicated to performing experiments that serve the following purposes: a basic experiment investigates if the used database is actually suited for building a conventional TTS system. Furthermore, the viability of the proposed multimodal modeling ideas is validated. In this context, it is of interest to explore if the tongue pose parameters are a suitable parameterization for the articulatory data. Moreover, the HMM and DNN approaches are compared.

#### 6.3.2. Setup

The conducted experiments used the default configuration for both HTS variants. In particular, the HTS 2.3 setup (HTS Working Group, 2015) was used for the HMM variant and the default HTS 2.3.1 setup<sup>1</sup> for the DNN version. This means that the DNN configuration applied 3 hidden layers containing 1024 nodes each. For both configurations, the limits of the fundamental frequency were adapted to the interval 60 Hz to 300 Hz.

#### 6.3.3. Database

The data for the experiments is again taken from the mngu0 corpus that was already explored in the previous chapter. This time, the following distribution packages were of interest:

1. Day1 basic audio data downsampled to 16 kHz (v1.1.0)
2. Day1 basic EMA data, head corrected and unnormalized (v1.1.0)
3. Day1 transcriptions, Festival utterances and ESPS label files (v1.1.1)

In contrast to the previous chapter, all EMA coils play a role in these experiments. The full EMA coil layout is shown in Figure 6.2; the coils are explained in Table 6.1.

From the provided acoustic data, signal parameters were extracted using STRAIGHT with a frame rate of 200 Hz, matching that of the EMA data. In order to follow the standard HTS methodology, the same parameters were also kept. Therefore, the signal parameters were 50 MGC, 25 BAP, and one coefficient for the  $F_0$ .

From the 1354 utterances in the data, 152 (11.2 %, around 10 min) were randomly selected and held back as a test set; the remaining 1202 utterances (around 105 min)

---

<sup>1</sup><http://hts.sp.nitech.ac.jp/?Download#f2602aa9>



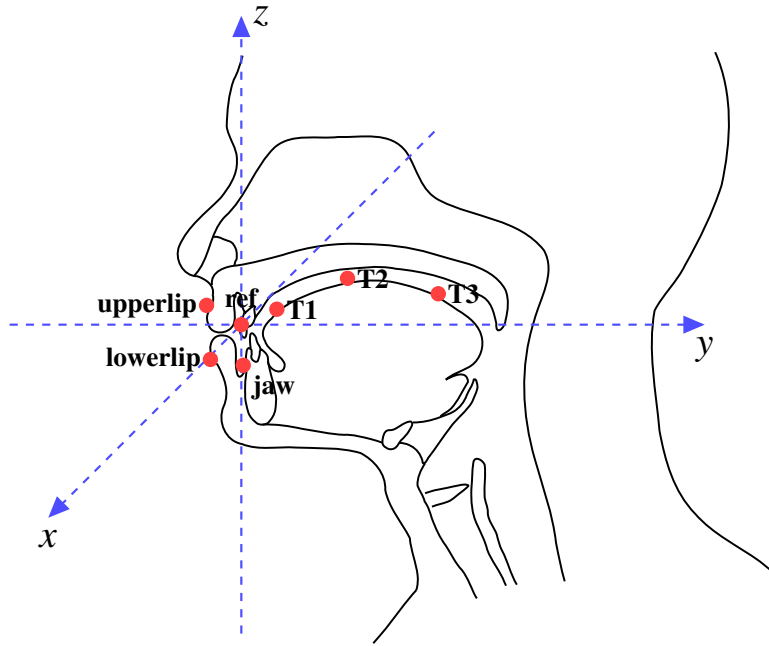


Figure 6.2.: Full EMA coil layout of the used data of the mngu0 corpus. All coils are close to the mid-sagittal plane. The ref coil on the upper incisors forms the origin of the coordinate space.

were used as the training set to build HTS synthesis voices. A comparison of phone distributions in the training and test sets shows a satisfactory match (cf. Figure 6.3).

#### 6.3.4. Used tongue model

The studies performed in Steiner, Sébastien Le Maguer, et al. (2017) and Sébastien Le Maguer et al. (2017) used a prototype version of the tongue model. In particular, it was derived from the Ultrax data using an earlier version of the proposed strategy in chapters 2 and 4. Differences to the current version include, for example:

- different approach for performing the palate reconstruction,
- omitted truncation of the model that finds a good compromise between generalization, specificity, and compactness,
- missing Laplacian smoothing of training meshes

As a truncation was omitted, the resulting multilinear model offers 12 and 13 degrees of freedom (DoF) for the anatomy and tongue pose, respectively.

## 6. Multimodal speech synthesis

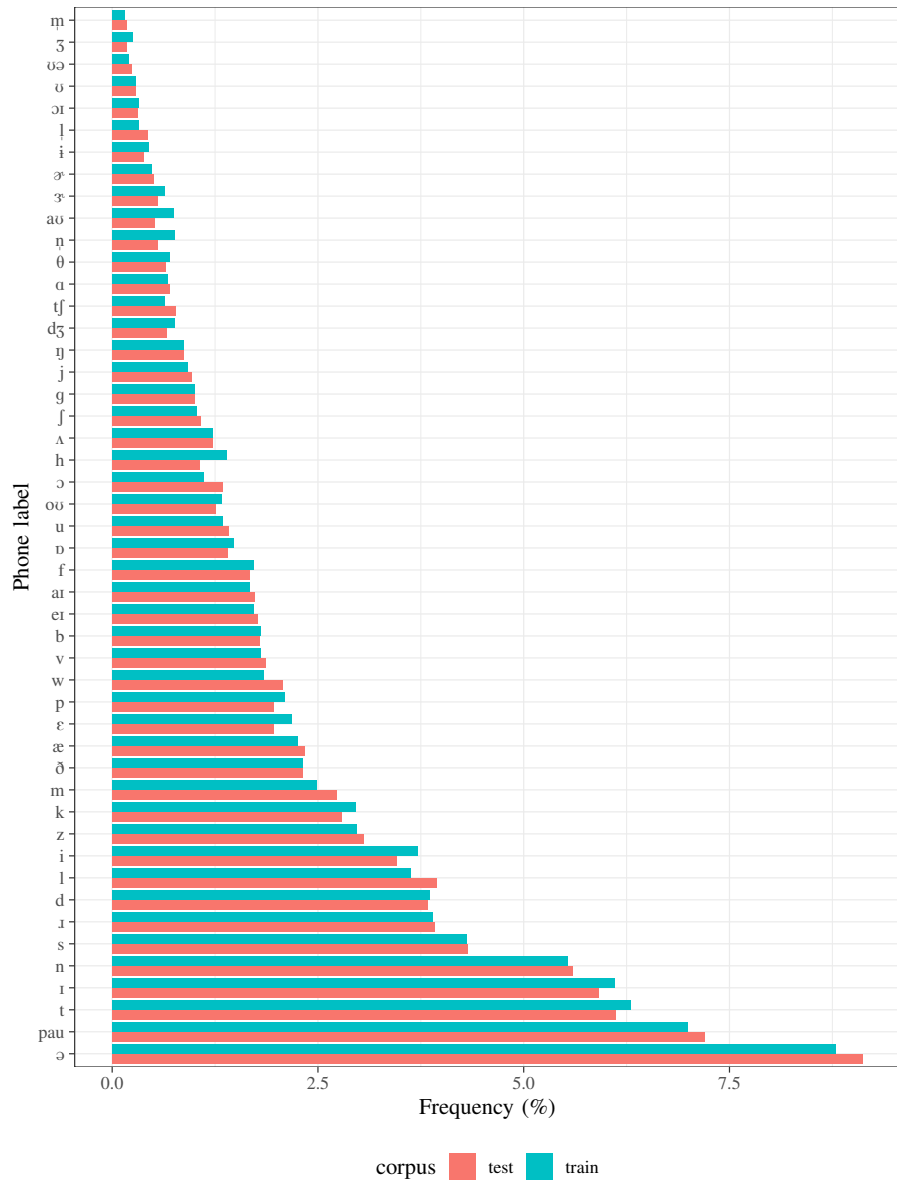


Figure 6.3.: Distribution of phones across the training and test sets. The frequency of the silence intervals, denoted by *pau*, is also shown.

Label	Location
<b>T1</b>	Tongue tip
<b>T2</b>	Tongue body
<b>T3</b>	Tongue dorsum
<b>upperlip</b>	Upper lip
<b>lowerlip</b>	Lower lip
<b>ref</b>	Upper incisor
<b>jaw</b>	Lower incisor

Table 6.1.: EMA coil labels and locations in the used data of the mngu0 corpus.

### 6.3.5. EMA registration approach

In addition to using a prototype version of the tongue model, these studies also used a different strategy for performing the EMA registration. First of all, the following corrections of the EMA data are missing:

- mid-sagittal projection
- rotational correction

Furthermore, the *ref* coil was used as the reference point that was used to map the EMA data to the origin of the tongue model. This position is slightly different from the reference point that was used in the previous chapter: the *ref* coil is located on the front incisors instead of the location where the hard palate ends and the tooth area starts. Additionally, the correspondence optimization between the tongue coils and the tongue mesh still used the randomized approach that was described in Chapter 5. Finally, the mean bias term was missing in the used energy for performing the registration.

Like in the current approach, two iterations of the registration are performed: in the first iteration, all components are optimized. Afterwards, the speaker anatomy is estimated from the results. Finally, the second iteration of the registrations fixes the speaker parameters to the estimated anatomy. The tongue pose parameters  $\mathbf{p}$  obtained from the second iteration are used as the parameterization in the training stage of the systems.

### 6.3.6. Acoustic Synthesis

In the first experiment, a conventional TTS system is built that uses acoustic data only. This system represents a baseline, which served mainly the purpose of validating the voicebuilding process and of ensuring that the transcriptions provided, and labels generated from them, along with the acoustic signal parameters, were able to generate audio of sufficient quality. Accordingly, a formal subjective listening test was omitted, and instead the baseline experiment was evaluated by using objective measures only.

To perform this task, 152 utterances were synthesized in the test set using two conditions. The first condition is the standard synthesis process. This condition allows to evaluate the duration accuracy. For the second condition, the acoustic phone durations

## 6. Multimodal speech synthesis

were imposed from the provided transcriptions of the dataset to allow direct comparison with the natural recordings. For the following experiments, both conditions were synthesized as well. The objective evaluation was conducted based on the following metrics.

For the duration evaluation, the duration root mean square error (RMSE) was calculated at the phone level (in ms) between the reference duration and the one synthesized using the first condition.

Considering the other coefficients, the synthesis result ( $s$ ), achieved using the second condition, was compared to the reference ( $r$ ) present in the test corpus. As the duration was imposed, the same number  $T$  of frames for the produced utterance and the reference one are available. To evaluate the fundamental frequency  $F_0$ , three measures were used: the voiced-unvoiced error rate percentage  $VUV(r, s)$  (Equation (6.2)) to check the prediction of the  $F_0$ , the RMSE in Hz (Equation (6.3)), and the RMSE in cent (Equation (6.4)). The latter measure focuses on the frames which are voiced in both conditions (original and predicted  $F_0$ ). Furthermore, it is a log scale measure adapted to the human perception.

$$v(x, y) = \begin{cases} 0 & x, y \text{ are both voiced/unvoiced} \\ 1 & \text{otherwise} \end{cases} \quad (6.1)$$

$$VUV(r, s) = 100 \sum_{t=1}^T v(r_t, s_t) / T \quad (6.2)$$

$$RMSE_{\text{Hz}}(r, s) = \sqrt{\sum_{t=1}^T (r_t - s_t)^2 / T} \quad (6.3)$$

$$RMSE_{\text{cent}}(r, s) = \sqrt{1200 \sum_{t=1}^T (\log(r_t) - \log(s_t))^2 / T} \quad (6.4)$$

Finally, to evaluate the spectral envelope production, the mel cepstral distortion between the MGC vectors of dimension  $M$  in dB is computed:

$$d(x, y) = \sum_{m=2}^M (x(m) - y(m))^2 \quad (6.5)$$

$$MCD(r, s) = \frac{10\sqrt{2}}{\ln 10} \sqrt{\sum_{t=1}^T d(r_t, s_t) / T} \quad (6.6)$$

Except for the duration, all parameters were evaluated at the frame level. Based on these measures, the results can be compared to previous studies, such as the one presented by Yokomizo et al. (2010).

The results of this evaluation are given in Table 6.2 and comprise the mean, standard deviation, and confidence interval with a  $p$  value at 5%. Compared to Yokomizo et al., 2010, the achieved results are slightly better, notwithstanding the different dataset. Therefore, it is safe to conclude that that the acoustic prediction of the baseline system

	HMM			DNN		
	mean	std. dev.	conf. int.	mean	std. dev.	conf. int.
<b>RMSE <math>F_0</math> (cent)</b>	188.52	76.92	12.33	153.62	67.86	10.91
<b>RMSE <math>F_0</math> (Hz)</b>	10.77	5.47	0.88	8.57	5.06	0.81
<b>VUV (%)</b>	12.03	3.94	0.63	11.38	3.70	0.60
<b>MCD (dB)</b>	2.45	0.22	0.04	2.13	0.20	0.03
<b>RMSE dur. (ms)</b>	42.00	18.29	2.93	43.04	18.65	3.00

Table 6.2.: Global evaluation measures for the acoustic synthesis baseline conditions. Results for the HMM and DNN setup are provided.

is consistent with the state of the art in HTS. Moreover, the results show that the DNN setup is outperforming the HMM one in this scenario, even with a relatively small amount of data. An explanation might be that the decision tree modeling in the HMM case is unable to capture some important correlation in the data.

### 6.3.7. Combined Acoustic and EMA Synthesis

In the second experiment, the paradigm of early multimodal fusion is adopted where the acoustic signal parameters are combined with the 3D positions of the seven EMA coils that are shown in Table 6.1, which increases the vector size by 21, to 97 parameters per frame. Here, the modified HTS framework presented in Figure 6.1a was used to build another TTS system from the present multimodal data.

By synthesizing the test set in this way, synthetic trajectories of predicted EMA coil positions were obtained in addition to the audio. To evaluate the combined acoustic and EMA synthesis, the same objective measures as in Section 6.3.6 were computed. Additionally, the Euclidean distance in space between the observed and predicted positions for the EMA coils were calculated. The results of this evaluation are shown in Table 6.3. The differences in the acoustic measures compared to the acoustic-only synthesis (cf. Table 6.2) are negligible. Again, it can be seen that the DNN setup outperforms the HMM one.

Like in the previous chapter, it is important to also inspect the results on a local scale: the comparison between the observed and predicted trajectories for one test utterance is illustrated in Figure 6.4. The observed and predicted (synthesized) positions of the three tongue coils are shown in each of the three dimensions in the data, along with the Euclidean distance. Silent intervals and consonants classified as coronal [t, d, n, l, s, z, ʃ, ʒ, θ, ð] and dorsal [g, k, ŋ], based on the provided phonetic transcription, have been highlighted. This helps visualize the correspondence between gestures of the tongue tip (coil T1) and tongue back (coils T2 and T3) for coronal and dorsal consonants, respectively, and the phonetic units they produce.

Several points merit discussion. First of all, there are large mismatches between the observed and predicted tongue EMA coil positions during the silent (pause) intervals at the beginning and end of the utterance. This can be attributed to the fact that

## 6. Multimodal speech synthesis

	HMM			DNN		
	mean	std. dev.	conf. int.	mean	std. dev.	conf. int.
<b>RMSE <math>F_0</math> (cent)</b>	188.43	63.70	10.21	158.87	65.25	10.46
<b>RMSE <math>F_0</math> (Hz)</b>	10.66	4.91	0.79	8.83	4.96	0.80
<b>VUV (%)</b>	12.14	3.84	0.62	11.16	3.95	0.63
<b>MCD (dB)</b>	2.45	0.23	0.04	2.11	0.19	0.03
<b>RMSE dur. (ms)</b>	41.93	19.04	3.05	41.58	16.84	2.70
<b>Eucl. dist. T3 (mm)</b>	2.14	1.47	$8.57 \times 10^{-3}$	1.78	1.26	$7.34 \times 10^{-3}$
<b>Eucl. dist. T2 (mm)</b>	2.10	1.54	$9.00 \times 10^{-3}$	1.76	1.31	$7.66 \times 10^{-3}$
<b>Eucl. dist. T1 (mm)</b>	2.17	1.62	$9.44 \times 10^{-3}$	1.79	1.33	$7.76 \times 10^{-3}$
<b>Eucl. dist. ref (mm)</b>	0.22	0.12	$6.97 \times 10^{-4}$	0.19	0.11	$6.25 \times 10^{-4}$
<b>Eucl. dist. jaw (mm)</b>	1.26	0.65	$3.80 \times 10^{-3}$	1.07	0.56	$3.28 \times 10^{-3}$
<b>Eucl. dist. ulip (mm)</b>	0.72	0.38	$2.21 \times 10^{-3}$	0.59	0.32	$1.86 \times 10^{-3}$
<b>Eucl. dist. llip (mm)</b>	1.45	0.93	$5.45 \times 10^{-3}$	1.23	0.78	$4.54 \times 10^{-3}$

Table 6.3.: Global evaluation measures for the combined acoustic and EMA synthesis.

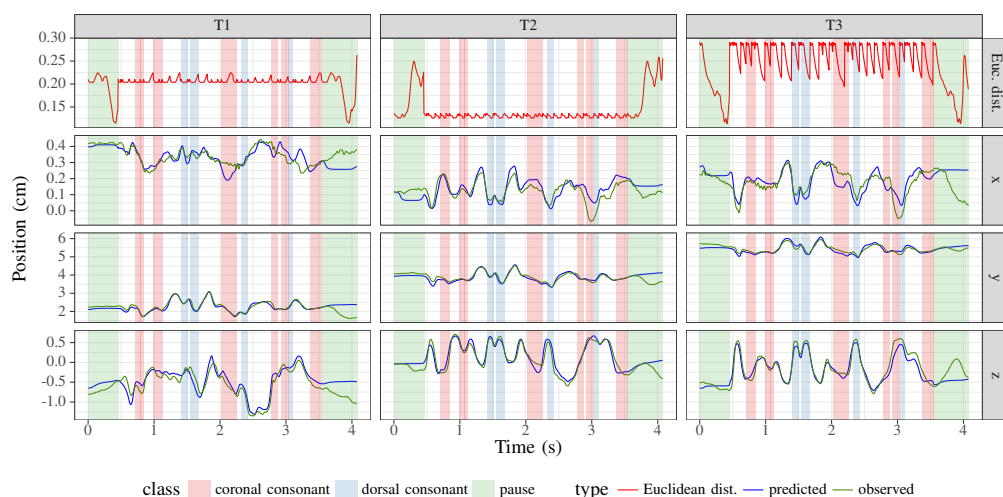


Figure 6.4.: Observed and predicted position trajectories (along the  $x$ ,  $y$ , and  $z$  axis), and Euclidean distance (top), for the tongue EMA coils (T1, T2, T3) for one test utterance, using combined acoustic and EMA synthesis with the HMM setup. The utterance  $s1\_0016$  is shown: “Because these deer are gregarious, they go about in groups”. Based on the provided transcriptions, intervals containing silent (pause) and coronal and dorsal consonants have been highlighted.

the wide range of the speaker’s tongue movements during non-speech intervals are not distinguished in the provided annotations, but invariably labeled with the same pause symbol. However, there are at least two very distinct shapes for the tongue during such silent intervals, including a *rest* and a *ready* position (just before speech is produced), in addition to other complex movements such as swallowing. In the absence of distinct labels corresponding to these positions and movements, none of this silent variation can be captured by the HMM trained on this data; instead, the tongue coils are predicted to hover around global means.

Secondly, there is noticeable oversmoothing and target extrema are not always quite reached. This can typically be attributed to the HMM based synthesis technique, despite the integration of global variance. The dynamics, however, are well represented, and the predicted positional trajectories match the observed reference quite closely.

The  $x$  axis appears to suffer from a greater amount of prediction error than the  $y$  or  $z$  axes. However, it should be noted that the positional variation along the  $x$  axis is an order of magnitude smaller than that along the  $y$  axis. It must also be borne in mind that nearly all of the speech-related movements occur in the mid-sagittal plane, represented by the  $y$  (anterior/posterior) and  $z$  (inferior/superior) axes; variation along the  $x$  axis corresponds to lateral movements, which are infrequent during speech.<sup>2</sup>

The Euclidean distances *during* speech are in the millimeter range, indicating that the predictions of EMA coil positions are accurate to within the precision of the EMA measurements themselves. However, there appears to be a certain amount of fluctuation with a more or less regular range and shape. The peaks of this fluctuation appear to correlate with spikes in the RMS channels of the provided EMA data, which supports the hypotheses that it is either an artifact of the algorithm which calculates the coil positions and orientations from the raw amplitudes (Stella et al., 2012), or measurement noise in the articulograph itself (Kroos, 2012), or, conceivably, a combination of both factors.

### 6.3.8. EMA Synthesis

The results of the previous two experiments provided an interesting observation: the evaluation of the acoustic measures described in subsection 6.3.6 are practically equivalent. This observation actually motivates the idea of decoupling the EMA synthesis completely from the acoustic one as shown in Figure 6.1b. Accordingly, the default HTS framework was used to build another TTS system trained only on the EMA data, without the acoustic parameters in order to find out if the same observation holds true for the articulatory data.

Under this condition, the evaluation of the duration RMSE and Euclidean distances between the predicted and observed EMA coils, computed using the formula given by Equation (6.3), is given in Table 6.4. As it can be seen, the results are nearly identical to those in Table 6.3, which confirms the validity of the assumption that the articulatory synthesis can be decoupled from the acoustic one. As expected, the DNN setup is

---

<sup>2</sup>The EMA simplification step in the previous chapter follows this rationale and flattens the data accordingly. A similar processing was applied to the normalized release variant of the mngu0 EMA dataset.

	HMM			DNN		
	mean	std. dev.	conf. int.	mean	std. dev.	conf. int.
<b>RMSE dur. (ms)</b>	53.73	20.74	3.32	53.79	20.28	3.25
<b>Eucl. dist. T3 (mm)</b>	2.18	1.42	$8.32 \times 10^{-3}$	1.84	1.28	$7.50 \times 10^{-3}$
<b>Eucl. dist. T2 (mm)</b>	2.17	1.54	$9.01 \times 10^{-3}$	1.84	1.37	$7.98 \times 10^{-3}$
<b>Eucl. dist. T1 (mm)</b>	2.26	1.61	$9.44 \times 10^{-3}$	1.89	1.41	$8.24 \times 10^{-3}$
<b>Eucl. dist. ref (mm)</b>	0.22	0.12	$6.80 \times 10^{-4}$	0.19	0.11	$6.19 \times 10^{-4}$
<b>Eucl. dist. jaw (mm)</b>	1.27	0.66	$3.87 \times 10^{-3}$	1.08	0.56	$3.27 \times 10^{-3}$
<b>Eucl. dist. ulip (mm)</b>	0.71	0.37	$2.19 \times 10^{-3}$	0.60	0.32	$1.89 \times 10^{-3}$
<b>Eucl. dist. llip (mm)</b>	1.47	0.92	$5.36 \times 10^{-3}$	1.26	0.82	$4.78 \times 10^{-3}$

Table 6.4.: Global evaluation for the EMA-only synthesis.

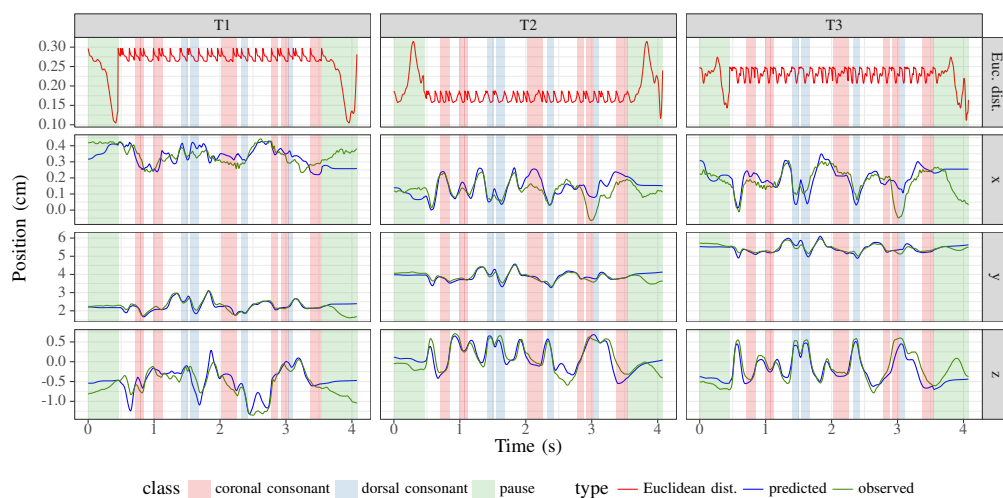


Figure 6.5.: One test utterance produced using EMA-only synthesis; all other details are the same as in Figure 6.4.

again performing better than its HMM counterpart. Figure 6.5 visualizes the comparison between the observed and predicted trajectories for one test utterance. The results of this experiment indicate that a decoupling of the EMA synthesis completely from the acoustic synthesis is a viable option.

### 6.3.9. Tongue-only EMA Synthesis

In order to focus on the tongue in the following section, it is necessary to first investigate how far the tongue coil EMA positions can be predicted in isolation from the remaining EMA coils. To this end, a modified version of the TTS system from the previous section was generated, by including *only* the tongue coils (T1, T2, and T3), and excluding the rest of the EMA data from the training set.

Table 6.5 provides the evaluation results of the EMA synthesis restricted to the three tongue coils. By comparing these results with those in Table 6.4 it can again be observed



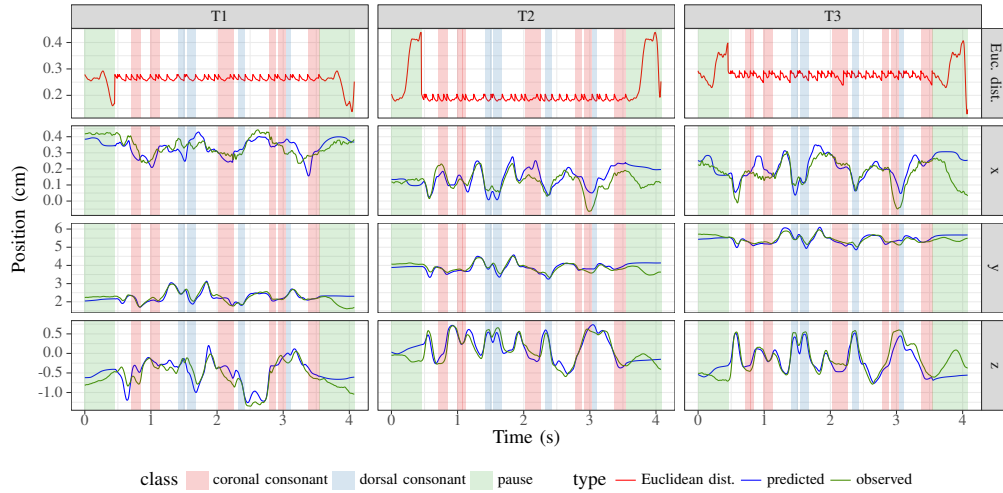


Figure 6.6.: One test utterance produced using EMA-only synthesis restricted to the tongue coils; all other details are the same as in Figure 6.4.

	HMM			DNN		
	mean	std. dev.	conf. int.	mean	std. dev.	conf. int.
<b>RMSE dur. (ms)</b>	61.20	21.64	3.47	59.21	18.37	2.95
<b>Eucl. dist. T3 (mm)</b>	2.21	1.45	$8.46 \times 10^{-3}$	3.84	2.08	0.01
<b>Eucl. dist. T2 (mm)</b>	2.18	1.50	$8.76 \times 10^{-3}$	3.97	2.07	0.01
<b>Eucl. dist. T1 (mm)</b>	2.25	1.56	$9.12 \times 10^{-3}$	3.75	2.41	0.01

Table 6.5.: Global evaluation for the EMA-only synthesis restricted to the tongue coils.

## 6. Multimodal speech synthesis

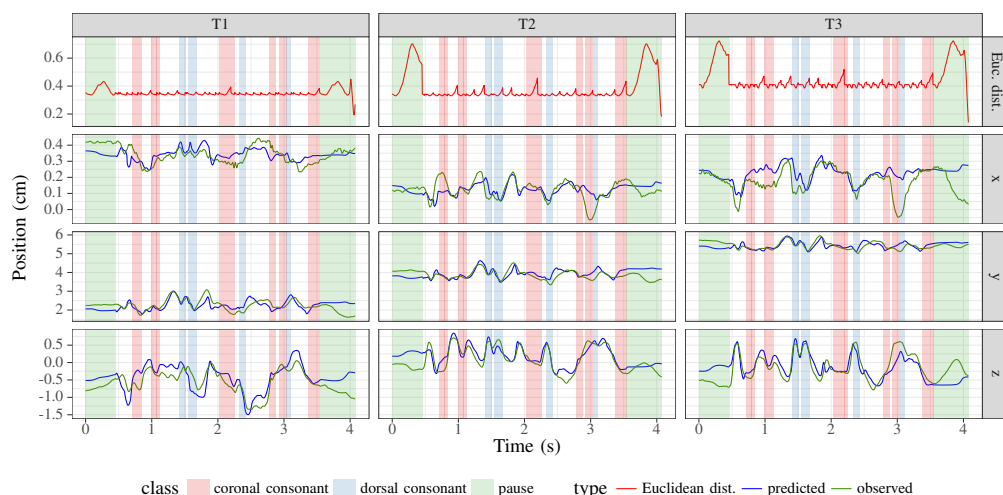


Figure 6.7.: One test utterance produced using the tongue model parameters synthesis; all other details are the same as in Figure 6.4.

	HMM			DNN		
	mean	std. dev.	conf. int.	mean	std. dev.	conf. int.
<b>RMSE dur. (ms)</b>	77.66	26.51	4.25	75.08	23.86	3.82
<b>Eucl. dist. T3 (mm)</b>	2.61	1.61	$9.43 \times 10^{-3}$	2.07	1.35	$7.92 \times 10^{-3}$
<b>Eucl. dist. T2 (mm)</b>	2.80	1.74	0.01	2.33	1.48	$8.66 \times 10^{-3}$
<b>Eucl. dist. T1 (mm)</b>	2.91	1.85	0.01	2.22	1.49	$8.74 \times 10^{-3}$

Table 6.6.: Global evaluation for the tongue model parameters synthesis.

that the values are virtually identical. However, this time, the DNN approach is providing significantly worse results than the HMM variant. A reason for this behavior could be that the available data is insufficient for the DNN to capture the bio-mechanical constraints of the tongue.

As before, the comparison between the observed and predicted trajectories for one test utterance is shown in Figure 6.6 for the HMM setup. It should be noted that despite the removal of the EMA coil on the lower incisor, some residual jaw motion is implicitly retained in the movements of the tongue coils. As stated before in this work, this is due to the fact that the tongue is attached to the lower jaw and therefore includes its motions.

### 6.3.10. Tongue model based tongue motion synthesis

The previous experiment revealed that the HTS framework can be used to synthesize audio and predict the movements of three tongue EMA coils using *separate* models trained on the mngu0 database. However, the DNN variant of the tongue coil prediction framework indicated that the raw EMA data might be an insufficient representation of the

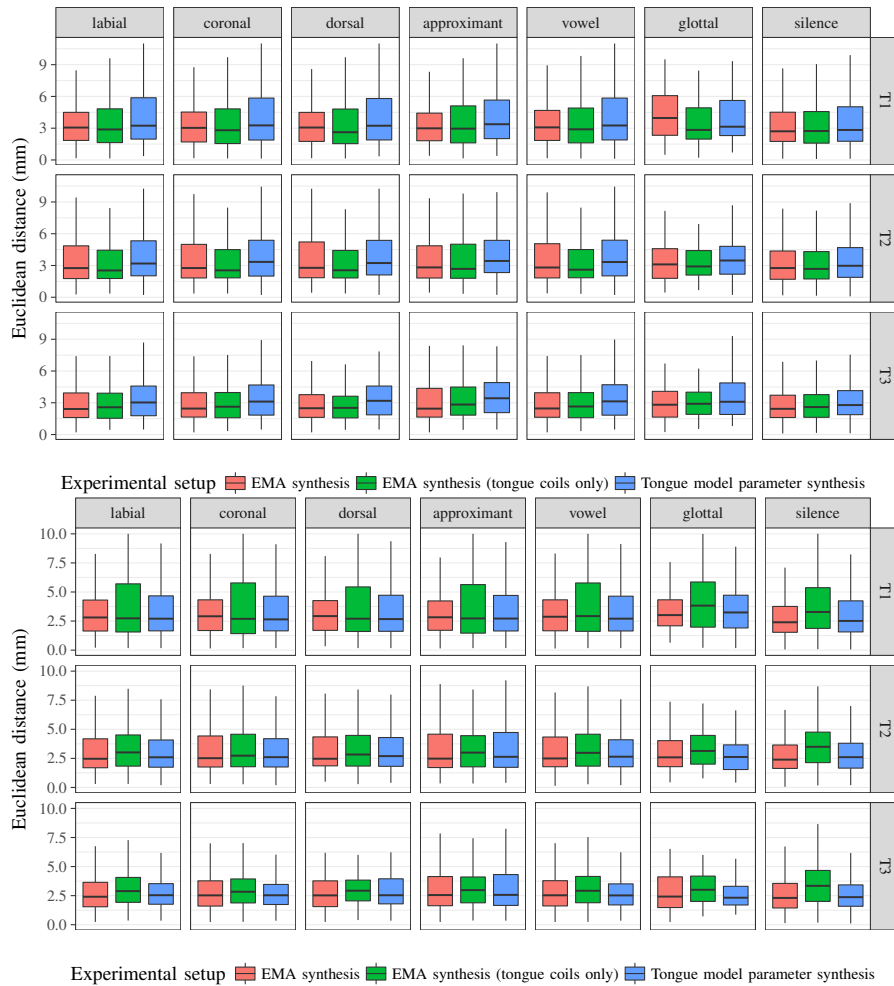


Figure 6.8.: Distributions of Euclidean distances between observed and predicted tongue EMA coil positions for each experimental TTS setup, split by phone class and tongue EMA coil. Results for the HMM (top) and DNN setup (bottom) are shown.

## 6. Multimodal speech synthesis

articulatory data for this purpose. The following experiment explored another kind of TTS system: instead of using the EMA data of the tongue directly, the tongue pose parameters obtained from the registration of this data with a tongue model are used in the training stage. Such a TTS system is then able to predict the tongue pose parameters and thus the whole 3D surface and motion of the tongue. To evaluate the performance of this system against the reference EMA data, the spatial coordinates of the vertices of the tongue mesh assigned during the correspondence optimization are used to produce synthetic trajectories that served as a virtual surrogate for predicted EMA data.

Like before, this synthetic EMA data is evaluated against the reference; Table 6.6 provides the Euclidean distances between the predicted and observed EMA coils, and one test utterance is visualized in Figure 6.7. Surprisingly, the DNN variant is outperforming the HMM counterpart again. Moreover, its performance is also better than the HMM one for the tongue coil only setup. This behavior may be seen as an indicator that the tongue pose parameters help the DNN to learn more about the structure of the underlying tongue motions than the raw positional data of the EMA coils.

As explained in the previous chapter, the obtained tongue pose parameters used for the training phase were subjected to a temporal smoothing during registration. While this step serves the purpose of ensuring a coherence of the shape over time and making the registration robust against noise, this extra smoothing seems to contribute to widespread target undershoot in the comparison.

### 6.3.11. Comparison of the different systems

Finally, in order to compare the three experimental TTS systems (trained without acoustic data), the distribution of Euclidean distances between each system and the observed reference data over the entire test set was analyzed; the results are shown in Figure 6.8. The distances are slightly greater when the non-tongue EMA coils are excluded, and greater still when the EMA prediction is replaced by the direct synthesis of tongue model parameters in the HMM case. However, overall, the distances remain in the same range, which indicates that the latter approach performs no worse than synthesis of EMA data – while adding the full 3D tongue surface into the synthesis process. For the DNN setup, it can be seen that in the case where only the tongue coils are used for the training, the tongue tip coil (T1) is the hardest to predict correctly.

Overall, it may be concluded that the DNN configuration is superior to the HMM one if the tongue pose parameters are used for parameterization. A similar observation holds true for the acoustic synthesis and the one where all EMA coils are involved.

### 6.3.12. Experiment using new model and registration

The evaluation so far showed that the DNN setup was the most reliable one for predicting the whole tongue surface by using the tongue pose parameters as parameterization of the articulatory data. However, the registration of the EMA data was performed by using a prototype tongue model and an earlier version of the EMA registration process. In a new experiment, the final version of the USC tongue model and the tongue pose parameters

	DNN (old)			DNN (new)		
	mean	std. dev.	conf. int.	mean	std. dev.	conf. int.
<b>RMSE dur. (ms)</b>	75.08	23.86	3.82	60.01	20.68	3.31
<b>Eucl. dist. T3 (mm)</b>	2.07	1.35	$7.92 \times 10^{-3}$	1.99	1.33	$7.78 \times 10^{-3}$
<b>Eucl. dist. T2 (mm)</b>	2.33	1.48	$8.66 \times 10^{-3}$	1.94	1.37	$8.03 \times 10^{-3}$
<b>Eucl. dist. T1 (mm)</b>	2.22	1.49	$8.74 \times 10^{-3}$	1.91	1.44	$8.44 \times 10^{-3}$

Table 6.7.: Comparison of results for old and new setup.

obtained from the new registration strategy presented in the previous chapter were used for the DNN system.

Table 6.7 summarizes the results and compares them to the previous DNN setup. It is important to note that the reference EMA trajectories correspond to the results of the preprocessing steps of the new EMA registration strategy. As it can be seen, the performance of the approach improved. Possible explanations for this phenomenon could be: the truncated USC model offers only 6 DoF for the tongue pose, compared to the 13 parameters of the prototype model, which may help the DNN to learn more about patterns in the pose parameter space. Additionally, Chapter 4 revealed that the USC model proved to be superior to the Ultrax one. Furthermore, the preprocessed EMA data could be a cause for the better results.

All in all, the experiment also showed that it is possible to update the articulatory model independently from the acoustical part.

## 6.4. Conclusion

This chapter described a process of synthesizing acoustic speech and synchronized animation of a full 3D surface model of the tongue. In particular, the HTS framework was used in combination with a single-speaker, multimodal articulatory database containing EMA motion capture data. First, a conventional and a fused multimodal approach were shown. Afterwards, the two modalities were separated while ensuring that the objective evaluation measures remained comparable. Finally, a tongue model was integrated into the TTS approach by using the tongue pose parameters as parameterization of the articulatory data. The accuracy of this system was evaluated by comparing the spatial coordinates of vertices on the tongue model surface to the reference EMA data from the original speaker’s tongue movements. Additionally, an objective evaluation between HMM and DNN modelings for the different TTS systems was undertaken. Here, it became apparent that the DNN approach is in general superior to the HMM one. As the original studies used a prototype tongue model and another EMA registration process, the tongue model based DNN TTS was retrained with updated data. This experiment revealed that the performance improved.

As noted before, the acoustic synthesis and predicted phone durations need not come from the same corpus as the one used for training the tongue model parameter synthesis

## 6. *Multimodal speech synthesis*

system. Under certain conditions, it would be straightforward to use a different, conventional TTS system with speech recordings from a different speaker in combination with this tongue model parameter synthesis, perhaps adapting it in the speaker subspace automatically or by hand, to generate a multimodal TTS application with plausible, speech-synchronized tongue motion, without the requirement of having articulatory data available for the target speaker. In this way, it is possible to first synthesize the acoustic speech signal, and to provide the predicted acoustic durations to guide the synthesis of corresponding tongue model parameters, which are then used to render the animation of the 3D tongue model in real time. However, an evaluation of this claim still needs to be performed. Clearly, more work has to be done: for example, the system should be evaluated by using human subjects who will rate it perceptually. Such a study can include intelligibility, such as the contribution of visible tongue movements during degraded, noisy, or absent audible speech. In this regard, it is also important to assess the impact on perceived naturalness by integrating the tongue model into a realistic talking avatar (e.g., Taylor et al. (2012) and Schabus et al. (2014)), and investigating the importance of naturalistic tongue movements for the overall impression of such avatars in multimodal spoken interaction scenarios with artificial characters. This may also lead to ideas on how to model distinct non-speech poses for the tongue, such as separate rest and ready positions.

# 7. Conclusion

## 7.1. Summary

This dissertation was roughly divided into two parts. In the first part (chapters 2 to 4), a framework was developed for estimating tongue meshes from magnetic resonance imaging (MRI) data and deriving tongue models from the results. Initially, a basic approach was constructed in Chapter 2 that combined image processing techniques with a template matching approach. This approach was semi-supervised, but suffered from several drawbacks that made it unsuitable for providing tongue meshes for a thorough statistical analysis. After the issues were addressed in Chapter 4, the framework was able to reliably register a significant amount of tongue shape configurations. Again, the framework was semi-supervised, which means that the user only had to provide a few annotations and tune some settings. The results of the extended approach were used to derive three tongue models. After an in-depth analysis, the tongue model derived from the USC dataset was determined to be the best one.

The next part of this work was concerned about possible applications of such a tongue model. The first application in Chapter 5 was concerned with registering sparse motion capture data by using the tongue model. Here, it was shown that the model could be adapted to different speakers. Furthermore, projecting the motion capture data into the pose parameter space of the tongue model revealed that patterns could be identified in the case of diphones. Additionally, the obtained meshes could be combined with face meshes originating from simultaneously recorded face capture data.

In Chapter 6, a multimodal text-to-speech (TTS) was designed that was able to synthesize audio and synchronized tongue motion. Here, it was discovered that the training of the acoustical model could completely be separated from the articulatory one. Moreover, the tongue model parameters proved to be a suitable parameterization of the articulatory data. Finally, it was discovered that a deep neural network (DNN) modeling outperformed a hidden Markov model (HMM) one.

## 7.2. Future work

Whereas the current results of the presented frameworks look promising, several issues can be identified that could be addressed in the future. Examples are given below:

**Open issue 1:** An objective evaluation of the estimated tongue shapes from the MRI scans is missing, only a subjective one was conducted. Although it was validated in an experiment by consulting speech experts, such an evaluation always suffers from a subjective bias. One way to perform an evaluation would be an analysis-by-synthesis

## 7. Conclusion

approach: here, MRI would be synthesized and the tongue shapes would be estimated from these generated scans. As the MRI data was generated, information about the true shape of the tongue would be available.

**Open issue 2:** Currently, it is still unclear how good the derived model separates speaker anatomy from tongue pose. The results of the fixed speaker registration approach of electromagnetic articulography (EMA) data showed an anatomic coherence over time, which might indicate that the separation is appropriate. However, a sophisticated analysis is needed to address this issue.

**Open issue 3:** So far, only a TTS system was presented that is able to synthesize audio and tongue motion. The experiments in Chapter 5 revealed that it is also possible to combine tongue and facial animation into a single one. The question arises if the presented TTS could be extended to also synthesize the corresponding face shape. To this end, a suitable dataset is needed that is large enough to perform this analysis. Accordingly, also the teeth could be added to the synthesized animations.

**Open issue 4:** The current framework is unable to handle tongue configurations with a large contact area between the soft palate and the tongue. Thus, an extension is needed to solve this issue. Here, it would be worthwhile to explore if volume-based template matching methods that use color information could be used to track the soft palate. The result could be used to reconstruct the corresponding surface area. Moreover, if this type of template matching works for MRI data, it may also be directly applied to the tongue.

**Open issue 5:** Although the objective evaluation of the EMA registration approach produced acceptable results, a subjective evaluation is missing that determines how realistic the produced tongue animations are. To this end, a perceptive study needs to be conducted that asks participants to rate the naturalness of the produced animations.

**Open issue 6:** While the tongue model is able to estimate the full three-dimensional surface from sparse EMA data, it is unclear if the estimated shape really corresponds to the real tongue surface. Like stated earlier, additional data of the speaker from other modalities, like, e.g., real-time magnetic resonance imaging (rtMRI) or ultrasound (US) could be helpful to validate the acquired shape. Here, it is also of interest to determine how many points on the tongue are actually needed to uniquely define a shape in the tongue model.

### 7.3. Source code

The source code of the presented framework is available for public use under an open-source software license. The interested reader can visit

<https://github.com/m2ci-msp/mri-shape-framework>

as a starting point.



# Appendices







# Bibliography

- Ackermann, Herrmann, B.F. Gröne, Gerhard Hoch, and Paul Walter Schönle (1993). “Speech freezing in Parkinson’s disease: a kinematic analysis of orofacial movements by means of electromagnetic articulography”. In: *Folia Phoniatrica et Logopaedica* 45.2, pp. 84–89. DOI: 10.1159/000266222.
- Akdemir, Eren and Tolga Çiloğlu (2008). “The use of articulator motion information in automatic speech segmentation”. In: *Speech Communication* 50.7, pp. 594–604. DOI: 10.1016/j.specom.2008.04.005.
- Allen, Brett, Brian Curless, and Zoran Popović (2003). “The space of human body shapes: reconstruction and parameterization from range scans”. In: *ACM Transactions on Graphics* 22.3, pp. 587–594. DOI: 10.1145/1201775.882311.
- Ananthakrishnan, Gopal, Pierre Badin, Julián Andrés Valdés Vargas, and Olov Engwall (Sept. 2010). “Predicting unseen articulations from multi-speaker articulatory models”. In: *Interspeech*. Makuhari, Japan, pp. 1588–1591. URL: [http://www.isca-speech.org/archive/interspeech\\_2010/i10\\_1588.html](http://www.isca-speech.org/archive/interspeech_2010/i10_1588.html).
- Astrinaki, Maria, Alexis Moinet, Junichi Yamagishi, Korin Richmond, Zhen-Hua Ling, Simon King, and Thierry Dutoit (Aug. 2013). “Mage – reactive articulatory feature control of HMM-based parametric speech synthesis”. In: *ISCA Workshop on Speech Synthesis*. Barcelona, Catalonia, Spain, pp. 207–211. URL: [http://www.isca-speech.org/archive/ssw8/ssw8\\_207.html](http://www.isca-speech.org/archive/ssw8/ssw8_207.html).
- Aubert, Gilles and Pierre Kornprobst (2006). *Mathematical problems in image processing: partial differential equations and the calculus of variations*. Vol. 147. Springer Science & Business Media. DOI: 10.1007/978-0-387-44588-5.
- Badin, Pierre, Gérard Bailly, Monica Raybaudi, and Christoph Segebarth (Nov. 1998). “A three-dimensional linear articulatory model based on MRI data”. In: *ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves House, Blue Mountains, NSW, Australia, pp. 249–254. URL: [http://www.isca-speech.org/archive\\_open/ssw3/ssw3\\_249.html](http://www.isca-speech.org/archive_open/ssw3/ssw3_249.html).
- Badin, Pierre, Gérard Bailly, Lionel Revéret, Monica Baciu, Christoph Segebarth, and Christophe Savariaux (2002). “Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images”. In: *Journal of Phonetics* 30.3, pp. 533–553. DOI: 10.1006/jpho.2002.0166.
- Badin, Pierre, Pascal Borel, Gérard Bailly, Lionel Revéret, Monica Baciu, and Christoph Segebarth (May 2000). “Towards an audiovisual virtual talking head: 3D articulatory modeling of tongue, lips and face based on MRI and video images”. In: *Speech Production Seminar*. Seeon, Germany, pp. 261–264.
- Badin, Pierre, Frédéric Elisei, Gérard Bailly, and Yuliya Tarabalka (2008). “An audiovisual talking head for augmented speech generation: models and animations based

## Bibliography

- on a real speaker's articulatory data". In: *Articulated Motion and Deformable Objects*. Springer, pp. 132–143. DOI: 10.1007/978-3-540-70517-8\_14.
- Badin, Pierre and Antoine Serrurier (Dec. 2006). "Three-dimensional linear modeling of tongue: articulatory data and models". In: *International Seminar on Speech Production*. Ubatuba, SP, Brazil, pp. 395–402. URL: <http://hal.archives-ouvertes.fr/hal-00167379>.
- Baer, Thomas, John C. Gore, L.C. Gracco, and Patrick W. Nye (1991). "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels". In: *Journal of the Acoustical Society of America* 90.2, pp. 799–828. DOI: 10.1121/1.401949.
- Baker, Adam (2011). *A biomechanical tongue model for speech production based on MRI live speaker data*. URL: <http://www.adambaker.org/qmu.php>.
- Baker, Simon and Iain Matthews (2004). "Lucas-Kanade 20 years on: a unifying framework". In: *International Journal of Computer Vision* 56.3, pp. 221–255. DOI: 10.1023/B:VISI.0000011205.11775.fd.
- Bälter, Olle, Olov Engwall, Anne-Marie Öster, and Hedvig Sidenbladh-Kjellström (Oct. 2005). "Wizard-of-Oz test of ARTUR - a computer-based speech training system with articulation correction". In: *International ACM SIGACCESS Conference on Computers and Accessibility*. Baltimore, MD, USA, pp. 36–43. DOI: 10.1145/1090785.1090795.
- Beautemps, Denis, Pierre Badin, and Gérard Bailly (2001). "Linear degrees of freedom in speech production: analysis of cineradio- and labio-film data and articulatory-acoustic modeling". In: *Journal of the Acoustical Society of America* 109.5, pp. 2165–2180. DOI: 10.1121/1.1361090.
- Ben Youssef, Atef (2011). "Control of talking heads by acoustic-to-articulatory inversion for language learning and rehabilitation". PhD thesis. Université Grenoble Alpes. URL: <https://tel.archives-ouvertes.fr/tel-00721957>.
- Benoît, Christian and Bertrand Le Goff (1998). "Audio-visual speech synthesis from French text: eight years of models, designs and evaluation at the ICP". In: *Speech Communication* 26.1, pp. 117–129. DOI: 10.1016/S0167-6393(98)00045-4.
- Bijar, Ahmad, Pierre-Yves Rohan, Pascal Perrier, and Yohan Payan (2016). "Atlas-based automatic generation of subject-specific finite element tongue meshes". In: *Annals of Biomedical Engineering* 44.1, pp. 16–34. DOI: 10.1007/s10439-015-1497-y.
- Blandin, Rémi, Marc Arnela, Rafael Laboissière, Xavier Pelorson, Oriol Guasch, Annemie Van Hirtum, and Xavier Laval (2015). "Effects of higher order propagation modes in vocal tract like geometries". In: *Journal of the Acoustical Society of America* 137.2, pp. 832–843. DOI: 10.1121/1.4906166.
- Blanz, Volker and Thomas Vetter (Aug. 1999). "A morphable model for the synthesis of 3D faces". In: *Annual Conference on Computer Graphics and Interactive Techniques*. Los Angeles, CA, USA, pp. 187–194. DOI: 10.1145/311535.311556.
- Blender Online Community (2018). *Blender - a 3D modelling and rendering package*. Blender Foundation. Blender Institute, Amsterdam. URL: <http://www.blender.org>.
- Bolkart, Timo (2016). "Dynamic and groupwise statistical analysis of 3D faces". PhD thesis. Saarland University. DOI: 10.22028/D291-26660.

- Bolkart, Timo and Stefanie Wuhler (June 2016). “A robust multilinear model learning framework for 3D faces”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA, pp. 4911–4919. DOI: 10.1109/CVPR.2016.531.
- Botsch, Mario, Leif Kobbelt, Mark Pauly, Pierre Alliez, and Bruno Lévy (2010). *Polygon mesh processing*. A K Peters/CRC Press. DOI: 10.1201/b10688.
- Boykov, Yuri and Gareth Funka-Lea (2006). “Graph cuts and efficient ND image segmentation”. In: *International Journal of Computer Vision* 70.2, pp. 109–131. DOI: 10.1007/s11263-006-7934-5.
- Brown, Robert W., Yu-Chung N. Cheng, E. Mark Haacke, Michael R. Thompson, and Ramesh Venkatesan (2014). *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons. DOI: 10.1002/9781118633953.
- Brunner, Jana, Susanne Fuchs, and Pascal Perrier (2009). “On the relationship between palate shape and articulatory behavior”. In: *Journal of the Acoustical Society of America* 125.6, pp. 3936–3949. DOI: 10.1121/1.3125313.
- Buchaillard, Stéphanie, Muriel Brix, Pascal Perrier, and Yohan Payan (2007). “Simulations of the consequences of tongue surgery on tongue mobility: implications for speech production in post-surgery conditions”. In: *International Journal of Medical Robotics and Computer Assisted Surgery* 3.3, pp. 252–261. DOI: 10.1002/rcs.142.
- Buchaillard, Stéphanie, Pascal Perrier, and Yohan Payan (2009). “A biomechanical model of cardinal vowel production: muscle activations and the impact of gravity on tongue positioning”. In: *Journal of the Acoustical Society of America* 126.4, pp. 2033–2051. DOI: 10.1121/1.3204306.
- Burdumy, Michael, Louisa Traser, Bernhard Richter, Matthias Echternach, Jan G. Korvink, Jürgen Hennig, and Maxim Zaitsev (2015). “Acceleration of MRI of the vocal tract provides additional insight into articulator modifications”. In: *Journal of Magnetic Resonance Imaging* 42.4, pp. 925–935. DOI: 10.1002/jmri.24857.
- Cai, Ming-Qi, Zhen-Hua Ling, and Li-Rong Dai (2015). “Statistical parametric speech synthesis using a hidden trajectory model”. In: *Speech Communication* 72, pp. 149–159. DOI: 10.1016/j.specom.2015.05.008.
- Carroll, J. Douglas and Jih-Jie Chang (1970). “Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition”. In: *Psychometrika* 35.3, pp. 283–319. DOI: 10.1007/BF02310791.
- Chan, Tony F. and Luminita A. Vese (2001). “Active contours without edges”. In: *IEEE Transactions on Image Processing* 10.2, pp. 266–277. DOI: 10.1109/83.902291.
- Cignoni, Paolo, Marco Callieri, Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, and Guido Ranzuglia (July 2008). “Meshlab: an open-source mesh processing tool.” In: *Eurographics Italian Chapter Conference*. Salerno, Italy, pp. 129–136. DOI: 10.2312/LocalChapterEvents/ItalChap/ItalianChapConf2008/129-136.
- De Lathauwer, Lieven (1997). “Signal processing based on multilinear algebra”. PhD thesis. Katholieke Universiteit Leuven.
- De Silva, Vin and Lek-Heng Lim (2008). “Tensor rank and the ill-posedness of the best low-rank approximation problem”. In: *SIAM Journal on Matrix Analysis and Applications* 30.3, pp. 1084–1127. DOI: 10.1137/06066518X.

## Bibliography

- Demolin, Didier, Thierry Metens, and Alain Soquet (May 2000). “Real time MRI and articulatory coordinations in vowels”. In: *Seminar on Speech Production*. Seon, Germany, pp. 86–93.
- Dryden, Ian L. and Kanti V. Mardia (1998). *Statistical shape analysis*. Wiley.
- Eldén, Lars and Berkant Savas (2009). “A Newton–Grassmann method for computing the best multilinear rank- $(r_1, r_2, r_3)$  approximation of a tensor”. In: *SIAM Journal on Matrix Analysis and Applications* 31.2, pp. 248–271. DOI: 10.1137/070688316.
- Elie, Benjamin, Yves Laprie, Pierre-André Vuissoz, and Freddy Odille (2016). “High spatiotemporal cineMRI films using compressed sensing for acquiring articulatory data”. In: *European Signal Processing Conference*, pp. 1353–1357. DOI: 10.1109/EUSIPCO.2016.7760469.
- Engwall, Olov (Oct. 2000a). “A 3D tongue model based on MRI data”. In: *International Conference on Spoken Language Processing*. Vol. 3. Beijing, China, pp. 901–904. URL: [http://www.isca-speech.org/archive/icslp\\_2000/i00\\_3901.html](http://www.isca-speech.org/archive/icslp_2000/i00_3901.html).
- (Oct. 2000b). “Are static MRI measurements representative of dynamic speech? Results from a comparative study using MRI, EPG and EMA”. In: *International Conference on Spoken Language Processing*. Vol. 1. Beijing, China, pp. 17–20. URL: [https://www.isca-speech.org/archive/archive\\_papers/icslp\\_2000/i00\\_1017.pdf](https://www.isca-speech.org/archive/archive_papers/icslp_2000/i00_1017.pdf).
- (Sept. 2001). “Making the tongue model talk: merging MRI & EMA measurements”. In: *Eurospeech*. Aalborg, Denmark, pp. 261–264. URL: [https://www.isca-speech.org/archive/archive\\_papers/eurospeech\\_2001/e01\\_0261.pdf](https://www.isca-speech.org/archive/archive_papers/eurospeech_2001/e01_0261.pdf).
- (Sept. 2002). “Evaluation of a system for concatenative articulatory visual speech synthesis”. In: *International Conference on Spoken Language Processing*. Denver, CO, USA, pp. 665–668. URL: [http://www.isca-speech.org/archive/icslp\\_2002/i02\\_0665.html](http://www.isca-speech.org/archive/icslp_2002/i02_0665.html).
- (2003). “Combining MRI, EMA and EPG measurements in a three-dimensional tongue model”. In: *Speech Communication* 41.2, pp. 303–329. DOI: 10.1016/S0167-6393(02)00132-2.
- (2006). “Assessing MRI measurements: effects of sustenation, gravitation and coarticulation”. In: *Speech production: Models, Phonetic Processes and Techniques*. Psychology Press, pp. 301–314.
- (Sept. 2008). “Can audio-visual instructions help learners improve their articulation? - an ultrasound study of short term changes”. In: *Interspeech*. Brisbane, Australia, pp. 2631–2634. URL: [http://www.isca-speech.org/archive/interspeech\\_2008/i08\\_2631.html](http://www.isca-speech.org/archive/interspeech_2008/i08_2631.html).
- Engwall, Olov and Pierre Badin (1999). “Collecting and analysing two- and three-dimensional MRI data for Swedish”. In: *KTH Department for Speech, Music and Hearing Quarterly Progress and Status Report* 40.3-4, pp. 11–38. URL: [http://www.speech.kth.se/prod/publications/files/qpsr/1999/1999\\_40\\_3-4\\_011-038.pdf](http://www.speech.kth.se/prod/publications/files/qpsr/1999/1999_40_3-4_011-038.pdf).
- Engwall, Olov and Olle Bälter (2007). “Pronunciation feedback from real and virtual language teachers”. In: *Computer Assisted Language Learning* 20.3, pp. 235–262. DOI: 10.1080/09588220701489507.
- Eryildirim, Abdulkadir and Marie-Odile Berger (Aug. 2011). “A guided approach for automatic segmentation and modeling of the vocal tract in MRI images”. In: *Euro-*



- pean Signal Processing Conference*. Barcelona, Spain, pp. 61–65. URL: <http://www.eurasip.org/Proceedings/Eusipco/Eusipco2011/papers/1569425007.pdf>.
- Fagel, Sascha and Caroline Clemens (2004). “An articulation model for audiovisual speech synthesis – determination, adjustment, evaluation”. In: *Speech Communication* 44.1-4, pp. 141–154. DOI: 10.1016/j.specom.2004.10.006.
- Fang, Qiang, Yun Chen, Haibo Wang, Jianguo Wei, Jianrong Wang, Xiyu Wu, and Aijun Li (Sept. 2016). “An improved 3D geometric tongue model”. In: *Interspeech*. San Francisco, CA, USA, pp. 1104–1107. DOI: 10.21437/Interspeech.2016-901.
- Foldvik, Arne Kjell, Ulf Kristiansen, and Jorn Kvaerness (Sept. 1993). “A time-evolving three-dimensional vocal tract model by means of magnetic resonance imaging (MRI)”. In: *Eurospeech*. Berlin, Germany, pp. 557–558. URL: [http://www.isca-speech.org/archive/eurospeech\\_1993/e93\\_0557.html](http://www.isca-speech.org/archive/eurospeech_1993/e93_0557.html).
- Förstner, Wolfgang and Eberhard Gülch (June 1987). “A fast operator for detection and precise location of distinct points, corners and centres of circular features”. In: *Intercommission Conference on Fast Processing of Photogrammetric Data*. Interlaken, Switzerland, pp. 281–305.
- Fu, Maojing, Bo Zhao, Christopher Carignan, Ryan K. Shosted, Jamie L. Perry, David P. Kuehn, Zhi-Pei Liang, and Bradley P. Sutton (2015). “High-resolution dynamic speech imaging with joint low-rank and sparsity constraints”. In: *Magnetic Resonance in Medicine* 73.5, pp. 1820–1832. DOI: 10.1002/mrm.25302.
- Fuchs, Susanne, Ralf Winkler, and Pascal Perrier (Dec. 2008). “Do speakers’ vocal tract geometries shape their articulatory vowel space?”. In: *International Seminar on Speech Production*. Strasbourg, France, pp. 333–336. URL: <http://issp2008.loria.fr/Proceedings/PDF/issp2008-77.pdf>.
- Fukada, Toshiaki, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai (Mar. 1992). “An adaptative algorithm for mel-cepstral analysis of speech”. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. San Francisco, CA, USA, pp. 137–140. DOI: 10.1109/ICASSP.1992.225953.
- Garofolo, John S., Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, and Nancy L. Dahlgren (1993). *DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM*. Interagency/Internal Report 4930. National Institute of Standards and Technology. DOI: 10.6028/nist.ir.4930.
- Geng, Christian and Christine Mooshammer (2009). “How to stretch and shrink vowel systems: results from a vowel normalization procedure”. In: *Journal of the Acoustical Society of America* 125.5, pp. 3278–3288. DOI: 10.1121/1.3106130.
- Gonzalez, Rafael C. and Richard E. Woods (2017). *Digital image processing, 4th edition*. Pearson.
- Grady, Leo (2006). “Random walks for image segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.11, pp. 1768–1783. DOI: 10.1109/TPAMI.2006.233.
- Hansen, Glen, Rod W. Douglass, and Andrew Zardecki (2005). *Mesh enhancement: selected elliptic methods, foundations and applications*. Imperial College Press, London. DOI: 10.1142/p351.

## Bibliography

- Harandi, Negar M., Rafeef Abugharbieh, and Sidney Fels (2015). “3D segmentation of the tongue in MRI: a minimally interactive model-based approach”. In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 3.4, pp. 178–188. DOI: 10.1080/21681163.2013.864958.
- Harandi, Negar M., Jonghye Woo, Maureen Stone, Rafeef Abugharbieh, and Sidney Fels (2017). “Variability in muscle activation of simple speech motions: a biomechanical modeling approach”. In: *Journal of the Acoustical Society of America* 141.4, pp. 2579–2590. DOI: 10.1121/1.4978420.
- Harshman, Richard A. (1970). “Foundations of the PARAFAC procedure: models and conditions for an “explanatory” multi-modal factor analysis”. In: *UCLA Working Papers in Phonetics* 16. URL: <http://escholarship.org/uc/item/0410x385>.
- Harshman, Richard A., Peter Ladefoged, and Louis Goldstein (1977). “Factor analysis of tongue shapes”. In: *Journal of the Acoustical Society of America* 62.3, pp. 693–707. DOI: 10.1121/1.381581.
- Hermes, Anne, Jane Mertens, and Doris Mücke (Sept. 2018). “Age-related effects on sensorimotor control of speech production”. In: *Interspeech*. Hyderabad, India, pp. 1526–1530. DOI: 10.21437/Interspeech.2018-1233.
- Herrmann, Leonard R. (1976). “Laplacian-isoparametric grid generation scheme”. In: *Journal of the Engineering Mechanics Division* 102.5, pp. 749–907.
- Hewer, Alexander, Ingmar Steiner, Timo Bolkart, Stefanie Wuhrer, and Korin Richmond (Aug. 2015). “A statistical shape space model of the palate surface trained on 3D MRI scans of the vocal tract”. In: *International Congress of Phonetic Sciences*. Glasgow, Scotland, pp. 0724.1–0724.5. URL: <https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0724.pdf>.
- Hewer, Alexander, Ingmar Steiner, and Korin Richmond (Mar. 2019). “Analysis of coarticulation using EMA data with a statistical shape space model of the tongue”. In: *Conference on Electronic Speech Signal Processing*. Dresden, Germany, pp. 296–303. URL: [http://www.essv.de/pdf/2019\\_296\\_303.pdf](http://www.essv.de/pdf/2019_296_303.pdf).
- Hewer, Alexander, Ingmar Steiner, and Stefanie Wuhrer (Sept. 2014). “A hybrid approach to 3D tongue modeling from vocal tract MRI using unsupervised image segmentation and mesh deformation”. In: *Interspeech*. Singapore, pp. 418–421. URL: [http://www.isca-speech.org/archive/interspeech\\_2014/i14\\_0418.html](http://www.isca-speech.org/archive/interspeech_2014/i14_0418.html).
- Hewer, Alexander, Stefanie Wuhrer, Ingmar Steiner, and Korin Richmond (2016). “Tongue mesh extraction from 3D MRI data of the human vocal tract”. In: *Perspectives in Shape Analysis*. Springer, pp. 345–365. DOI: 10.1007/978-3-319-24726-7\_16.
- (Sept. 2018). “A multilinear tongue model derived from speech related MRI data of the human vocal tract”. In: *Computer Speech & Language* 51, pp. 68–92. DOI: 10.1016/j.csl.2018.02.001.
- Hitchcock, Frank L. (1927). “The expression of a tensor or a polyadic as a sum of products”. In: *Journal of Mathematics and Physics* 6.1-4, pp. 164–189. DOI: 10.1002/sapm192761164.
- Honda, Kiyoshi, Shinji Maeda, Michiko Hashi, Jim S. Dembowski, and John R. Westbury (Oct. 1996). “Human palate and related structures: their articulatory consequences”. In: *International Conference on Spoken Language Processing*. Philadelphia, PA, USA,

- pp. 784–787. URL: [http://www.isca-speech.org/archive/icslp\\_1996/i96\\_0784.html](http://www.isca-speech.org/archive/icslp_1996/i96_0784.html).
- Hoole, Phil and Noël Nguyen (1997). “Electromagnetic articulography in coarticulation research”. In: *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München* 35, pp. 177–184.
- Hoole, Phil, Axel Wismüller, Gerda Leinsinger, Christian Kroos, Anja Geumann, and Michiko Inoue (May 2000). “Analysis of tongue configuration in multi-speaker, multi-volume MRI data”. In: *Seminar on Speech Production*. Seeon, Germany, pp. 157–160.
- Hoole, Phil and Andreas Zierdt (2010). “Five-dimensional articulography”. In: *Speech Motor Control: New Developments in Basic and Applied Research*. Oxford University Press, pp. 331–349. DOI: 10.1093/acprof:oso/9780199235797.003.0020.
- Hoole, Phil, Andreas Zierdt, and Christian Geng (Aug. 2003). “Beyond 2D in articulatory data acquisition and analysis”. In: *International Congress of Phonetic Sciences*. Barcelona, Spain, pp. 265–268. URL: [https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15\\_0265.html](https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15_0265.html).
- HTS Working Group (Dec. 2015). *HTS Document: list of modifications made in HTS (for version 2.3)*. URL: [http://hts.sp.nitech.ac.jp/archives/2.3/HTS\\_Document.pdf](http://hts.sp.nitech.ac.jp/archives/2.3/HTS_Document.pdf).
- International Phonetic Association (1999). *Handbook of the International Phonetic Association*. Cambridge University Press.
- (2018). *IPA chart*. URL: <http://www.internationalphoneticassociation.org/content/ipa-chart>.
- Jackson, Philip J. B. and Veena D. Singampalli (2009). “Statistical identification of articulation constraints in the production of speech”. In: *Speech Communication* 51.8, pp. 695–710. DOI: 10.1016/j.specom.2009.03.007.
- James, Kristy, Alexander Hewer, Ingmar Steiner, and Stefanie Wuhler (Sept. 2016). “A real-time framework for visual feedback of articulatory data using statistical shape models”. In: *Interspeech*. San Francisco, CA, USA, pp. 1569–1570. URL: [http://www.isca-speech.org/archive/Interspeech\\_2016/abstracts/2019.html](http://www.isca-speech.org/archive/Interspeech_2016/abstracts/2019.html).
- Johnson, Keith, Peter Ladefoged, and Mona Lindau (1993). “Individual differences in vowel production”. In: *Journal of the Acoustical Society of America* 94.2, pp. 701–714. DOI: 10.1121/1.406887.
- Kaburagi, Tokihiko (Sept. 2015). “Morphological and acoustic analysis of the vocal tract using a multi-speaker volumetric MRI dataset”. In: *Interspeech*. Dresden, Germany, pp. 379–383. URL: [http://www.isca-speech.org/archive/interspeech\\_2015/i15\\_0379.html](http://www.isca-speech.org/archive/interspeech_2015/i15_0379.html).
- Kass, Michael, Andrew Witkin, and Demetri Terzopoulos (1988). “Snakes: active contour models”. In: *International Journal of Computer Vision* 1.4, pp. 321–331. DOI: 10.1007/BF00133570.
- Katz, William, Thomas F. Campbell, Jun Wang, Eric Farrar, J. Coleman Eubanks, Arvind Balasubramanian, Balakrishnan Prabhakaran, and Rob Rennaker (Sept. 2014). “Opti-Speech: a real-time, 3D visual feedback system for speech training.” In: *Interspeech*. Singapore, pp. 1174–1178. URL: [http://www.isca-speech.org/archive/interspeech\\_2014/i14\\_1174.html](http://www.isca-speech.org/archive/interspeech_2014/i14_1174.html).

## Bibliography

- Kawahara, Hideki, Ikuyo Masuda-Katsuse, and Alain de Cheveigné (1999). “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds”. In: *Speech Communication* 27.3-4, pp. 187–207. DOI: 10.1016/S0167-6393(98)00085-5.
- Kazhdan, Michael, Matthew Bolitho, and Hugues Hoppe (June 2006). “Poisson surface reconstruction”. In: *Eurographics Symposium on Geometry Processing*. Cagliari, Sardinia, Italy, pp. 61–70. DOI: 10.2312/SGP/SGP06/061-070.
- Kiers, Henk A. L. and Wim P. Krijnen (1991). “An efficient algorithm for PARAFAC of three-way data with large numbers of observation units”. In: *Psychometrika* 56.1, pp. 147–152. DOI: 10.1007/BF02294592.
- Kim, Yoon-Chul, Shrikanth S. Narayanan, and Krishna Shrinivas Nayak (2009). “Accelerated three-dimensional upper airway MRI using compressed sensing”. In: *Magnetic Resonance in Medicine* 61.6, pp. 1434–1440. DOI: 10.1002/mrm.21953.
- King, Simon, Joe Frankel, Karen Livescu, Erik McDermott, Korin Richmond, and Mirjam Wester (2007). “Speech production knowledge in automatic speech recognition”. In: *Journal of the Acoustical Society of America* 121.2, pp. 723–742. DOI: 10.1121/1.2404622.
- Kitamura, Tatsuya, Hironori Takemoto, Kiyoshi Honda, Yasuhiro Shimada, Ichiro Fujimoto, Yuko Syakudo, Shinobu Masaki, Kagayaki Kuroda, Noboru Oku-Uchi, and Michio Senda (2005). “Difference in vocal tract shape between upright and supine postures: observations by an open-type MRI scanner”. In: *Acoustical Science and Technology* 26.5, pp. 465–468. DOI: 10.1250/ast.26.465.
- Kittler, Josef and John Illingworth (1985). “On threshold selection using clustering criteria”. In: *IEEE Transactions on Systems, Man, and Cybernetics* 5, pp. 652–655. DOI: 10.1109/TSMC.1985.6313443.
- Kolda, Tamara G. and Brett W. Bader (2009). “Tensor decompositions and applications”. In: *SIAM review* 51.3, pp. 455–500. DOI: 10.1137/07070111X.
- Kröger, Bernd J., Ralf Winkler, Christine Mooshammer, and Bernd Pompino-Marschall (May 2000). “Estimation of vocal tract area function from magnetic resonance imaging: preliminary results”. In: *Seminar on Speech Production*. Seeon, Germany, pp. 333–336.
- Kroos, Christian (May 2012). “Evaluation of the measurement precision in three-dimensional electromagnetic articulography (Carstens AG500)”. In: *Journal of Phonetics* 40.3, pp. 453–465. DOI: 10.1016/j.wocn.2012.03.002.
- Labrunie, Mathieu, Pierre Badin, Dirk Voit, Arun A. Joseph, Jens Frahm, Laurent Lamalle, Coriandre Vilain, and Louis-Jean Boë (2018). “Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning”. In: *Speech Communication* 99, pp. 27–46. DOI: 10.1016/j.specom.2018.02.004.
- Ladefoged, Peter (1982). *A course in phonetics*. Second edition. Harcourt Brace Jovanovich.
- Ladefoged, Peter and Donald Eric Broadbent (1957). “Information conveyed by vowels”. In: *Journal of the Acoustical Society of America* 29.1, pp. 98–104. DOI: 10.1121/1.1908694.

- Le Maguer, Sébastien, Ingmar Steiner, and Alexander Hewer (Aug. 2017). “An HMM/DNN comparison for synchronized text-to-speech and tongue motion synthesis”. In: *Interspeech*. Stockholm, Sweden, pp. 239–243. DOI: 10.21437/Interspeech.2017-936.
- Lee, Junghoon, Jonghye Woo, Fangxu Xing, Emi Z. Murano, Maureen Stone, and Jerry L Prince (Apr. 2013). “Semi-automatic segmentation of the tongue for 3D motion analysis with dynamic MRF”. In: *IEEE International Symposium on Biomedical Imaging*. San Francisco, CA, USA, pp. 1465–1468. DOI: 10.1109/ISBI.2013.6556811.
- Lee, Sang Uk, Seok Yoon Chung, and Rae Hong Park (1990). “A comparative performance study of several global thresholding techniques for segmentation”. In: *Computer Vision, Graphics, and Image Processing* 52.2, pp. 171–190. DOI: 10.1016/0734-189X(90)90053-X.
- Li, Chunming, Chiu-Yen Kao, John C. Gore, and Zhaohua Ding (June 2007). “Implicit active contours driven by local binary fitting energy”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. Minneapolis, MN, USA, pp. 1–7. DOI: 10.1109/CVPR.2007.383014.
- Li, Hao, Bart Adams, Leonidas J. Guibas, and Mark Pauly (2009). “Robust single-view geometry and motion reconstruction”. In: *ACM Transactions on Graphics* 28.5, pp. 1–10. DOI: 10.1145/1618452.1618521.
- Li, Tianye, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero (2017). “Learning a model of facial shape and expression from 4D scans”. In: *ACM Transactions on Graphics* 36.6, pp. 194.1–194.17. DOI: 10.1145/3130800.3130813.
- Lim, Yongwan, Yinghua Zhu, Sajan Goud Lingala, Dani Byrd, Shrikanth S. Narayanan, and Krishna Shrinivas Nayak (2019). “3D dynamic MRI of the vocal tract during natural speech”. In: *Magnetic Resonance in Medicine* 81.3, pp. 1511–1520. DOI: 10.1002/mrm.27570.
- Ling, Zhen-Hua, Korin Richmond, and Junichi Yamagishi (2010a). “An analysis of HMM-based prediction of articulatory movements”. In: *Speech Communication* 52.10, pp. 834–846. DOI: 10.1016/j.specom.2010.06.006.
- (Sept. 2010b). “HMM-based text-to-articulatory-movement prediction and analysis of critical articulators”. In: *Interspeech*. Makuhari, Japan, pp. 2194–2197. URL: [http://www.isca-speech.org/archive/interspeech\\_2010/i10\\_2194.html](http://www.isca-speech.org/archive/interspeech_2010/i10_2194.html).
- (2013). “Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.1, pp. 207–219. DOI: 10.1109/tasl.2012.2215600.
- Ling, Zhen-Hua, Korin Richmond, Junichi Yamagishi, and Ren-Hua Wang (2009). “Integrating articulatory features into HMM-based parametric speech synthesis”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.6, pp. 1171–1185. DOI: 10.1109/tasl.2009.2014796.
- Lingala, Sajan Goud, Asterios Toutios, Johannes Töger, Yongwan Lim, Yinghua Zhu, Yoon-Chul Kim, Colin Vaz, Shrikanth S. Narayanan, and Krishna Shrinivas Nayak (Sept. 2016). “State-of-the-art MRI protocol for comprehensive assessment of vocal tract structure and function”. In: *Interspeech*. San Francisco, CA, USA, pp. 475–479. DOI: 10.21437/Interspeech.2016-559.

## Bibliography

- Lingala, Sajan Goud, Yinghua Zhu, Yoon-Chul Kim, Asterios Toutios, Shrikanth S. Narayanan, and Krishna Shrinivas Nayak (2017). “A fast and flexible MRI system for the study of dynamic vocal tract shaping”. In: *Magnetic Resonance in Medicine* 77.1, pp. 112–125. DOI: 10.1002/mrm.26090.
- Liu, Dong C. and Jorge Nocedal (1989). “On the limited memory BFGS method for large scale optimization”. In: *Mathematical Programming* 45.1-3, pp. 503–528. DOI: 10.1007/BF01589116.
- Liu, Ji, Przemyslaw Musialski, Peter Wonka, and Jieping Ye (2013). “Tensor completion for estimating missing values in visual data”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1, pp. 208–220. DOI: 10.1109/TPAMI.2012.39.
- Liu, Jiamin and Jayaram K. Udupa (2009). “Oriented active shape models”. In: *IEEE Transactions on Medical Imaging* 28.4, pp. 571–584. DOI: 10.1109/TMI.2008.2007820.
- Lucas, Bruce D. and Takeo Kanade (Aug. 1981). “An iterative image registration technique with an application to stereo vision”. In: *International Joint Conference on Artificial Intelligence*. Vancouver, BC, Canada, pp. 674–679.
- Maeda, Shinji (1990). “Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model”. In: *Speech Production and Speech Modelling*. Springer, pp. 131–149. DOI: 10.1007/978-94-009-2037-8\_6.
- Magenat-Thalmann, Nadia, Richard Laperrière, and Daniel Thalmann (June 1988). “Joint-dependent local deformations for hand animation and object grasping”. In: *Graphics Interface*. Edmonton, AB, Canada, pp. 26–33. DOI: 10.20380/GI1988.04.
- Max, Nelson (1999). “Weights for computing vertex normals from facet normals”. In: *Journal of Graphics Tools* 4.2, pp. 1–6. DOI: 10.1080/10867651.1999.10487501.
- McGurk, Harry and John MacDonald (1976). “Hearing lips and seeing voices”. In: *Nature* 264, pp. 746–748. DOI: 10.1038/264746a0.
- Meenakshi, Nisha, Chiranjeevi Yarra, B. K. Yamini, and Prasanta Kumar Ghosh (Sept. 2014). “Comparison of speech quality with and without sensors in electromagnetic articulograph AG 501 recording”. In: *Interspeech*. Singapore, pp. 935–939. URL: [https://www.isca-speech.org/archive/archive\\_papers/interspeech\\_2014/i14\\_0935.pdf](https://www.isca-speech.org/archive/archive_papers/interspeech_2014/i14_0935.pdf).
- Mermelstein, Paul (1973). “Articulatory model for the study of speech production”. In: *Journal of the Acoustical Society of America* 53.4, pp. 1070–1082. DOI: 10.1121/1.1913427.
- Mills, Anne E. (1987). “The development of phonology in the blind child”. In: *Hearing by eye: The psychology of lip-reading*. Lawrence Erlbaum Associates, pp. 145–161.
- Mitra, Vikramjit, Hosung Nam, Carol Y. Espy-Wilson, Elliot Saltzman, and Louis Goldstein (2011). “Articulatory information for noise robust speech recognition”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7, pp. 1913–1924. DOI: 10.1109/TASL.2010.2103058.
- Montgomery, D. (Oct. 1981). “Do dyslexics have difficulty accessing articulatory information?” In: *Psychological Research* 43.2, pp. 235–243. DOI: 10.1007/BF00309832.
- Mount, David M. and Sunil Arya (2010). *ANN: a library for approximate nearest neighbor searching*. URL: <http://www.cs.umd.edu/~mount/ANN/>.

- Musche, Sebastian (2014). “Visualisierung von Artikulationsbewegungen mittels Articulographie und Ultraschall”. MA thesis. Trier University.
- Narayanan, Shrikanth S., Abeer A. Alwan, and Katherine Haker (1995). “An articulatory study of fricative consonants using magnetic resonance imaging”. In: *Journal of the Acoustical Society of America* 98.3, pp. 1325–1347. DOI: 10.1121/1.413469.
- (1997). “Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals”. In: *Journal of the Acoustical Society of America* 101.2, pp. 1064–1077. DOI: 10.1121/1.418030.
- Narayanan, Shrikanth S., Krishna Shrinivas Nayak, Sungbok Lee, Abhinav Sethy, and Dani Byrd (2004). “An approach to real-time magnetic resonance imaging for speech production”. In: *Journal of the Acoustical Society of America* 115.4, pp. 1771–1776. DOI: 10.1121/1.1652588.
- Narayanan, Shrikanth S., Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee, Krishna Shrinivas Nayak, Yoon-Chul Kim, Yinghua Zhu, Louis Goldstein, Dani Byrd, Erik Bresch, Prasanta Ghosh, Athanasios Katsamanis, and Michael Proctor (2014). “Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)”. In: *Journal of the Acoustical Society of America* 136.3, pp. 1307–1311. DOI: 10.1121/1.4890284.
- Niebergall, Aaron, Shuo Zhang, Esther Kunay, Götz Keydana, Michael Job, Martin Uecker, and Jens Frahm (2013). “Real-time MRI of speaking at a resolution of 33 ms: undersampled radial FLASH with nonlinear inverse reconstruction”. In: *Magnetic Resonance in Medicine* 69.2, pp. 477–485. DOI: 10.1002/mrm.24276.
- Otsu, Nobuyuki (1979). “A threshold selection method from gray-level histograms”. In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1, pp. 62–66. DOI: 10.1109/TSMC.1979.4310076.
- Ouni, Slim, Loïc Mangeonjean, and Ingmar Steiner (Sept. 2012). “VisArtico: a visualization tool for articulatory data”. In: *Interspeech*. Portland, OR, USA, pp. 1878–1881. URL: [http://www.isca-speech.org/archive/interspeech\\_2012/i12\\_1878.html](http://www.isca-speech.org/archive/interspeech_2012/i12_1878.html).
- Peng, Ting, Erwan Kerrien, and Marie-Odile Berger (Mar. 2010). “A shape-based framework to segmentation of tongue contours from MRI data”. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Dallas, TX, USA, pp. 662–665. DOI: 10.1109/ICASSP.2010.5495123.
- Perkell, Joseph S., Marc H. Cohen, Mario A. Svirsky, Melanie L. Matthies, Inaki Garabita, and Michel T. Jackson (1992). “Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements”. In: *Journal of the Acoustical Society of America* 92.6, pp. 3078–3096. DOI: 10.1121/1.404204.
- Perkell, Joseph S. and Winston L. Nelson (1985). “Variability in production of the vowels /i/ and /a/”. In: *Journal of the Acoustical Society of America* 77.5, pp. 1889–1895. DOI: 10.1121/1.391940.
- Perona, Pietro and Jitendra Malik (1990). “Scale-space and edge detection using anisotropic diffusion”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.7, pp. 629–639. DOI: 10.1109/34.56205.
- Raeesy, Zeynab, Sylvia Rueda, Jayaram K. Udupa, and John Coleman (Apr. 2013). “Automatic segmentation of vocal tract MR images”. In: *IEEE International Symposium*

## Bibliography

- on *Biomedical Imaging*. San Francisco, CA, USA, pp. 1328–1331. DOI: 10.1109/ISBI.2013.6556777.
- Richmond, Korin, Phil Hoole, and Simon King (Aug. 2011). “Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus”. In: *Interspeech*. Florence, Italy, pp. 1505–1508. URL: [http://www.isca-speech.org/archive/interspeech\\_2011/i11\\_1505.html](http://www.isca-speech.org/archive/interspeech_2011/i11_1505.html).
- Richmond, Korin and Steve Renals (Sept. 2012). “Ultrax: an animated midsagittal vocal tract display for speech therapy”. In: *Interspeech*. Portland, OR, USA, pp. 74–77. URL: [http://www.isca-speech.org/archive/interspeech\\_2012/i12\\_0074.html](http://www.isca-speech.org/archive/interspeech_2012/i12_0074.html).
- Rodrigues, Maria Andréia Formico, Duncan F. Gillies, and Peter Charters (2001). “A biomechanical model of the upper airways for simulating laryngoscopy”. In: *Computer Methods in Biomechanics and Biomedical Engineering* 4.2, pp. 127–148. DOI: 10.1080/10255840008908001.
- Rosset, Antoine, Luca Spadola, and Osman Ratib (2004). “OsiriX: an open-source software for navigating in multidimensional DICOM images”. In: *Journal of Digital Imaging* 17.3, pp. 205–216. DOI: 10.1007/s10278-004-1014-6.
- Rudy, Krista and Yana Yunusova (2013). “The effect of anatomic factors on tongue position variability during consonants”. In: *Journal of Speech, Language, and Hearing Research* 56.1, pp. 137–149. DOI: 10.1044/1092-4388(2012/11-0218).
- Rudzicz, Frank, Aravind Kumar Namasivayam, and Talya Wolff (2012). “The TORGO database of acoustic and articulatory speech from speakers with dysarthria”. In: *Language Resources and Evaluation* 46.4, pp. 523–541. DOI: 10.1007/s10579-011-9145-0.
- Savariaux, Christophe, Pierre Badin, Adeline Samson, and Silvain Gerber (2017). “A comparative study of the precision of Carstens and Northern Digital Instruments electromagnetic articulographs”. In: *Journal of Speech, Language, and Hearing Research* 60.2, pp. 322–340. DOI: 10.1044/2016\_JSLHR-S-15-0223.
- Savran, Arman, Neşe Alyüz, Hamdi Dibeklioglu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun (May 2008). “Bosphorus database for 3D face analysis”. In: *European Workshop on Biometrics and Identity Management*. Roskilde, Denmark, pp. 47–56. DOI: 10.1007/978-3-540-89991-4\_6.
- Schabus, Dietmar, Michael Pucher, and Gregor Hofer (Apr. 2014). “Joint audiovisual hidden semi-Markov model-based speech synthesis”. In: *IEEE Journal of Selected Topics in Signal Processing* 8.2, pp. 336–347. DOI: 10.1109/jstsp.2013.2281036.
- Schenck, John F. (2000). “Safety of strong, static magnetic fields”. In: *Journal of Magnetic Resonance Imaging* 12.1, pp. 2–19.
- Schönle, Paul W., Klaus Gräbe, Peter Wenig, Jörg Höhne, Jörg Schrader, and Bastian Conrad (May 1987). “Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract”. In: *Brain and Language* 31.1, pp. 26–35. DOI: 10.1016/0093-934X(87)90058-7.
- Scott, A.D., R. Boubertakh, M.J. Birch, and M.E. Miquel (2012). “Towards clinical assessment of velopharyngeal closure using MRI: evaluation of real-time MRI sequences at 1.5 and 3 T”. In: *British Journal of Radiology* 85.1019, pp. 1083–1092. DOI: 10.1259/bjr/32938996.



- Serrurier, Antoine, Pierre Badin, Louis-Jean Boë, Laurent Lamalle, and Christiane Neuschaefer-Rube (Aug. 2017). “Inter-speaker variability: speaker normalisation and quantitative estimation of articulatory invariants in speech production for French”. In: *Interspeech*. Stockholm, Sweden, pp. 2272–2276. DOI: 10.21437/Interspeech.2017-1126.
- Shadle, Christine H., Mohammad Mohammad, John N. Carter, and Phillip J. B. Jackson (Aug. 1999). “Multi-planar dynamic magnetic resonance imaging: new tools for speech research”. In: *International Congress of Phonetic Sciences*. San Francisco, CA, USA, pp. 623–626. URL: [https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/p14\\_0623.html](https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/p14_0623.html).
- Somandepalli, Krishna, Asterios Toutios, and Shrikanth S. Narayanan (Aug. 2017). “Semantic edge detection for tracking vocal tract air-tissue boundaries in real-time magnetic resonance images”. In: *Interspeech*. Stockholm, Sweden, pp. 631–635. DOI: 10.21437/Interspeech.2017-1580.
- Sorensen, Tanner, Zisis Iason Skordilis, Asterios Toutios, Yoon-Chul Kim, Yinghua Zhu, Jangwon Kim, Adam Lammert, Vikram Ramanarayanan, Louis Goldstein, Dani Byrd, Krishna Shrinivas Nayak, and Shrikanth S. Narayanan (Aug. 2017). “Database of volumetric and real-time vocal tract MRI for speech science”. In: *Interspeech*. Stockholm, Sweden, pp. 645–649. DOI: 10.21437/Interspeech.2017-608.
- Sorkine, Olga, Daniel Cohen-Or, Yaron Lipman, Marc Alexa, Christian Rössl, and H.-P. Seidel (July 2004). “Laplacian surface editing”. In: *Eurographics Symposium on Geometry Processing*. Nice, France, pp. 175–184. DOI: 10.1145/1057432.1057456.
- Steiner, Ingmar, Peter Knopp, Sebastian Musche, Astrid Schmiedel, Angelika Braun, and Slim Ouni (May 2014). “Investigating the effects of posture and noise on speech production”. In: *International Seminar on Speech Production*. Cologne, Germany, pp. 413–416.
- Steiner, Ingmar, Sébastien Le Maguer, and Alexander Hewer (Dec. 2017). “Synthesis of tongue motion and acoustics from text using a multimodal articulatory database”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12, pp. 2351–2361. DOI: 10.1109/TASLP.2017.2756818.
- Steiner, Ingmar and Slim Ouni (June 2011). “Investigating articulatory differences between upright and supine posture using 3D EMA”. In: *International Seminar on Speech Production*. Montréal, Canada. URL: <https://hal.inria.fr/inria-00602427>.
- (Mar. 2012). “Artimate: an articulatory animation framework for audiovisual speech synthesis”. In: *ISCA Workshop on Innovation and Applications in Speech Technology*. Dublin, Ireland, pp. 57–60. arXiv: 1203.3574.
- Steiner, Ingmar, Korin Richmond, Ian Marshall, and Calum D. Gray (2012). “The magnetic resonance imaging subset of the mngu0 articulatory corpus”. In: *Journal of the Acoustical Society of America* 131.2. Express Letters, pp. 106–111. DOI: 10.1121/1.3675459.
- Steiner, Ingmar, Korin Richmond, and Slim Ouni (Aug. 2013). “Speech animation using electromagnetic articulography as motion capture data”. In: *International Conference on Auditory-Visual Speech Processing*. Annecy, France, pp. 55–60. URL: [http://avsp2013.loria.fr/proceedings/papers/paper\\_52.pdf](http://avsp2013.loria.fr/proceedings/papers/paper_52.pdf).

## Bibliography

- Stella, Massimo, Paolo Bernardini, Francesco Sigona, Antonio Stella, Mirko Grimaldi, and Barbara Gili Fivela (2012). “Numerical instabilities and three-dimensional electromagnetic articulography”. In: *Journal of the Acoustical Society of America* 132.6, pp. 3941–3949. DOI: 10.1121/1.4763549.
- Stone, Maureen and Subhash Lele (Oct. 1992). “Representing the tongue surface with curve fits”. In: *International Conference on Spoken Language Processing*. Banff, AB, Canada, pp. 875–878. URL: [https://www.isca-speech.org/archive/icslp\\_1992/i92\\_0875.html](https://www.isca-speech.org/archive/icslp_1992/i92_0875.html).
- Stone, Maureen and Andrew Lundberg (1996). “Three-dimensional tongue surface shapes of English consonants and vowels”. In: *Journal of the Acoustical Society of America* 99.6, pp. 3728–3737. DOI: 10.1121/1.414969.
- Stone, Maureen, G. Stock, Kevin Bunin, Kausum Kumar, M. Epstein, Chandra Kambhamettu, Min Li, Vijay Parthasarathy, and J. Prince (2007). “Comparison of speech production in upright and supine position”. In: *Journal of the Acoustical Society of America* 122.1, pp. 532–541. DOI: 10.1121/1.2715659.
- Stone, Maureen, Jonghye Woo, Junghoon Lee, Tera Poole, Amy Seagraves, Michael Chung, Eric Kim, Emi Z. Murano, Jerry L. Prince, and Silvia S. Blemker (2018). “Structure and variability in human tongue muscle anatomy”. In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6.5, pp. 499–507. DOI: 10.1080/21681163.2016.1162752.
- Styner, Martin A., Kumar T. Rajamani, Lutz-Peter Nolte, Gabriel Zsemlye, Gábor Székely, Christopher J. Taylor, and Rhodri H. Davies (July 2003). “Evaluation of 3D correspondence methods for model building”. In: *International Conference on Information Processing in Medical Imaging*. Ambleside, UK, pp. 63–75. DOI: 10.1007/978-3-540-45087-0\_6.
- Su, Zhihua, Jianguo Wei, Qiang Fang, Jianrong Wang, and Kiyoshi Honda (Sept. 2018). “Tongue segmentation with geometrically constrained snake model”. In: Hyderabad, India, pp. 3117–3121. DOI: 10.21437/Interspeech.2018-1108.
- Sumby, W. H. and Irwin Pollack (1954). “Visual contribution to speech intelligibility in noise”. In: *Journal of the Acoustical Society of America* 26.2, pp. 212–215. DOI: 10.1121/1.1907309.
- Szeliski, Richard (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media. DOI: 10.1007/978-1-84882-935-0.
- Taylor, Sarah L., Moshe Mahler, Barry-John Theobald, and Iain Matthews (July 2012). “Dynamic units of visual speech”. In: *Eurographics/ACM SIGGRAPH Symposium on Computer Animation*. Lausanne, Switzerland, pp. 275–284. DOI: 10.2312/SCA/SCA12/275-284.
- Tiede, Mark K., Shinobu Masaki, and Eric Vatikiotis-Bateson (May 2000). “Contrasts in speech articulation observed in sitting and supine conditions”. In: *Seminar on Speech Production*. Seeon, Germany, pp. 25–28.
- Tiede, Mark K., Shinobu Masaki, Masahiko Wakumoto, and Eric Vatikiotis-Bateson (1997). “Magnetometer observation of articulation in sitting and supine conditions”. In: *Journal of the Acoustical Society of America* 102.5, pp. 3166–3166. DOI: 10.1121/1.420773.

- Tiede, Mark K., Hani Camille Yehia, and Eric Vatikiotis-Bateson (May 1996). “A shape-based approach to vocal tract area function estimation”. In: *ESCA Tutorial and Research Workshop on Speech Production Modeling*. Aufrans, France, pp. 41–44. URL: [http://www.isca-speech.org/archive/spm\\_96/sps6\\_041.html](http://www.isca-speech.org/archive/spm_96/sps6_041.html).
- Tokuda, Keiichi, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura (June 2000). “Speech parameter generation algorithms for HMM-based speech synthesis”. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 3. Istanbul, Turkey, pp. 1315–1318. DOI: 10.1109/ICASSP.2000.861820.
- Tomaschek, Fabian, Martijn Wieling, Denis Arnold, and Harald Baayen (Aug. 2013). “Word frequency, vowel length and vowel quality in speech production: an EMA study of the importance of experience”. In: *Interspeech*. Lyon, France, pp. 1302–1306. URL: [http://www.isca-speech.org/archive/interspeech\\_2013/i13\\_1302.html](http://www.isca-speech.org/archive/interspeech_2013/i13_1302.html).
- Toutios, Asterios and Shrikanth S. Narayanan (Aug. 2015). “Factor analysis of vocal-tract outlines derived from real-time magnetic resonance imaging data”. In: *International Congress of Phonetic Sciences*. Glasgow, Scotland, pp. 0514.1–0514.5. URL: <https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0514.pdf>.
- Tucker, Ledyard R. (1966). “Some mathematical notes on three-mode factor analysis”. In: *Psychometrika* 31.3, pp. 279–311. DOI: 10.1007/BF02289464.
- Valliappan, C. A., Renuka Mannem, and Prasanta Kumar Ghosh (2018). “Air-tissue boundary segmentation in real-time magnetic resonance imaging video using semantic segmentation with fully convolutional networks”. In: *Interspeech*. Hyderabad, India, pp. 3132–3136. DOI: 10.21437/Interspeech.2018-1939.
- Vargas, Julián Andrés Valdés, Pierre Badin, Gopal Ananthakrishnan, and Laurent Lamalle (June 2012). “Articulatory speaker normalisation based on MRI-data using three-way linear decomposition methods”. In: *Journées d’Études sur la Parole*. Grenoble, France, pp. 529–536. URL: <http://www.aclweb.org/anthology/F12-1067>.
- Vargas, Julián Andrés Valdés, Pierre Badin, and Laurent Lamalle (Sept. 2012). “Articulatory speaker normalisation based on MRI-data using three-way linear decomposition methods”. In: *Interspeech*. Portland, OR, USA, pp. 2186–2189. URL: [http://www.isca-speech.org/archive/interspeech\\_2012/i12\\_2186.html](http://www.isca-speech.org/archive/interspeech_2012/i12_2186.html).
- Vlasic, Daniel, Matthew Brand, Hanspeter Pfister, and Jovan Popović (July 2005). “Face transfer with multilinear models”. In: *Annual Conference on Computer Graphics and Interactive Techniques*. Los Angeles, CA, USA, pp. 426–433. DOI: 10.1145/1186822.1073209.
- Weickert, Joachim (1998). *Anisotropic diffusion in image processing*. Teubner.
- Weickert, Joachim, Andrés Bruhn, Thomas Brox, and Nils Papenberg (2006). “A survey on variational optic flow methods for small displacements”. In: *Mathematical Models for Registration and Applications to Medical Imaging*. Springer, pp. 103–136. DOI: 10.1007/978-3-540-34767-5\_5.
- Weirich, Melanie and Susanne Fuchs (2013). “Palatal morphology can influence speaker-specific realizations of phonemic contrasts”. In: *Journal of Speech, Language, and Hearing Research* 56.6, pp. 1894–1908. DOI: 10.1044/1092-4388(2013/12-0217).

## Bibliography

- Weirich, Melanie, Leonardo Lancia, and Jana Brunner (2013). "Inter-speaker articulatory variability during vowel-consonant-vowel sequences in twins and unrelated speakers". In: *Journal of the Acoustical Society of America* 134.5, pp. 3766–3780. DOI: 10.1121/1.4822480.
- Weismer, Gary and Kate Bunton (1999). "Influences of pellet markers on speech production behavior: acoustical and perceptual measures". In: *Journal of the Acoustical Society of America* 105.5, pp. 2882–2894. DOI: 10.1121/1.426902.
- Wieling, Martijn, Fabian Tomaschek, Denis Arnold, Mark Tiede, Franziska Bröker, Samuel Thiele, Simon N. Wood, and R. Harald Baayen (2016). "Investigating dialectal differences using articulography". In: *Journal of Phonetics* 59, pp. 122–143. DOI: 10.1016/j.wocn.2016.09.004.
- Woo, Jonghye, Junghoon Lee, Emi Z. Murano, Fangxu Xing, Meena Al-Talib, Maureen Stone, and Jerry L. Prince (2015). "A high-resolution atlas and statistical model of the vocal tract from structural MRI". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 3.1, pp. 47–60. DOI: 10.1080/21681163.2014.933679.
- Woo, Jonghye, Fangxu Xing, Junghoon Lee, Maureen Stone, and Jerry L. Prince (June 2015). "Construction of an unbiased spatio-temporal atlas of the tongue during speech". In: *International Conference on Information Processing in Medical Imaging*. Sabhal Mor Ostaig, Isle of Skye, UK, pp. 723–732. DOI: 10.1007/978-3-319-19992-4\_57.
- Wrench, Alan A (Mar. 2000). "A multi-channel/multi-speaker articulatory database for continuous speech recognition research". In: *Workshop on "Phonetics and Phonology in ASR. Parameters and Features, and their Implications"*. Saarbrücken, Germany, pp. 1–13. URL: [http://www.coli.uni-saarland.de/groups/WB/Phonetics/contents/phonus-pdf/phonus5/Wrench\\_PHONUS5.pdf](http://www.coli.uni-saarland.de/groups/WB/Phonetics/contents/phonus-pdf/phonus5/Wrench_PHONUS5.pdf).
- Wu, Xiyu, Jianwu Dang, and Ian Stavness (2014). "Iterative method to estimate muscle activation with a physiological articulatory model". In: *Acoustical Science and Technology* 35.4, pp. 201–212. DOI: 10.1250/ast.35.201.
- Wuhrer, Stefanie, Jochen Lang, Motahareh Tekieh, and Chang Shu (2015). "Finite element based tracking of deforming surfaces". In: *Graphical Models* 77, pp. 1–17. DOI: 10.1016/j.gmod.2014.10.002.
- Yin, Lijun, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale (Sept. 2008). "A high-resolution 3D dynamic facial expression database". In: *International Conference on Automatic Face and Gesture Recognition*. Amsterdam, The Netherlands, pp. 1–6. DOI: 10.1109/AFGR.2008.4813324.
- Yin, Lijun, Xiaozhou Wei, Yi Sun, Jun Wang, and M. J. Rosato (Apr. 2006). "A 3D facial expression database for facial behavior research". In: *International Conference on Automatic Face and Gesture Recognition*. Southampton, UK, pp. 211–216. DOI: 10.1109/FGR.2006.6.
- Yokomizo, Shuji, Takashi Nose, and Takao Kobayashi (Sept. 2010). "Evaluation of prosodic contextual factors for HMM-based speech synthesis". In: *Interspeech*. Makuhari, Japan, pp. 430–433. URL: [http://www.isca-speech.org/archive/interspeech\\_2010/i10\\_0430.html](http://www.isca-speech.org/archive/interspeech_2010/i10_0430.html).

- Yunusova, Yana, Melanie Baljko, Grigore Pintilie, Krista Rudy, Petros Faloutsos, and John Daskalogiannakis (2012). “Acquisition of the 3D surface of the palate by in-vivo digitization with Wave”. In: *Speech Communication* 54.8, pp. 923–931. DOI: 10.1016/j.specom.2012.03.006.
- Yunusova, Yana, Jeffrey S. Rosenthal, Krista Rudy, Melanie Baljko, and John Daskalogiannakis (2012). “Positional targets for lingual consonants defined using electromagnetic articulography”. In: *Journal of the Acoustical Society of America* 132.2, pp. 1027–1038. DOI: 10.1121/1.4733542.
- Zen, Heiga, Andrew Senior, and Mike Schuster (May 2013). “Statistical parametric speech synthesis using deep neural networks”. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vancouver, BC, Canada, pp. 7962–7966. DOI: 10.1109/ICASSP.2013.6639215.
- Zen, Heiga and Tomoki Toda (Sept. 2005). “An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005”. In: *Interspeech*. Lisbon, Portugal, pp. 93–96. URL: [http://www.isca-speech.org/archive/interspeech\\_2005/i05\\_0093.html](http://www.isca-speech.org/archive/interspeech_2005/i05_0093.html).
- Zheng, Yanli, Mark Hasegawa-Johnson, and Shamala Pizza (2003). “Analysis of the three-dimensional tongue shape using a three-index factor analysis model”. In: *Journal of the Acoustical Society of America* 113.1, pp. 478–486. DOI: 10.1121/1.1520538.