From Scan to Speech

Articulation analysis from real-time vocal tract MRI

Kicky van Leeuwen

From Scan to Speech

Articulation analysis from real-time vocal tract MRI

Master's Thesis

by

Kicky van Leeuwen

Master Technical Medicine
University of Twente, Enschede
Netherlands Cancer Institute, Amsterdam

September 2019

UNIVERSITY OF TWENTE.

ANTONI VAN LEEUWENHOEK
NEDERLANDS KANKER INSTITUUT

Abstract

Speech production is a complex process drawing much attention from researchers. When speech is altered due to, for example, cancer in the tongue, lips or palate, it is important to understand how the articulation has changed to provide optimal rehabilitation therapy.

In this thesis, we aimed to develop a methodology that enables objective and replicable assessment of speech articulation. Real-time magnetic resonance imaging (rtMRI) was chosen as a means to acquire articulatory information during speech, due to its good soft tissue contrast and non-invasiveness. The USC Speech and Vocal Tract Morphology MRI Database [1] was used throughout this thesis and contains MR data from seventeen healthy American-English speaking subjects.

A preliminary study was performed to demonstrate that relevant articulatory information is present in the MRI data. We trained a deep learning network to predict from a single MR image the phoneme that was articulated. The network itself was analyzed and revealed that it had learned similar relations between vowels as is known to phoneticians.

During articulation, the vocal tract shape changes through which sound is transformed to speech. To extract quantitative information on the articulation, we segmented the vocal tract from every rtMRI frame in the dataset with the Chan-Vese level set method. We used Bayesian hyperparameter optimization to learn optimal parameters for the level set and image preprocessing. With this method, we showed that all frames could be segmented with the need of a single manual segmentation per subject with a dice score of 95.6 % and a mean surface distance of 1.8 mm.

From these vocal tract segmentations, we subsequently derived the vocal tract distance function. The centerline of the vocal tract was found and a grid was projected from which the width of the vocal tract was deducted for each frame. We

[1] https://sail.usc.edu/span/morphdb/

have proposed several ways of visualizing the vocal tract dynamics, such that the location of articulation for different phonemes can be studied and the differences in articulation space could may be observed.

With this thesis, we developed a methodology to extract the vocal tract distance function of a large rtMRI dataset. Where most studies focus on merely a single time point or a single location within the vocal tract over time, the method proposed here regards both the temporal and spatial dimension. The enrichment of the dataset is made publicly available to support articulatory studies in healthy subjects to broaden our understanding of articulation in speech. The tool itself has the potential to be used in clinical practice by aiding speech therapist with the assessment of the patient's articulation abilities. By performing pre- and postintervention measurements the effect of treatment can be studied and a personalized rehabilitation plan proposed.

Examination committee

Chairman and technical supervisor
Dr. ir. F. van der Heijden, University of Twente and Netherlands Cancer Institute

Medical supervisor
Prof. dr. L.E. Smeele, Netherlands Cancer Institute

Personal development supervisor
Drs. P.A. van Katwijk, University of Twente

Additional supervisors
Dr. R.J.J.H. van Son, Netherlands Cancer Institute and University of Amsterdam
P. Bos, MSc., Netherlands Cancer Institute

External member
Dr. M. Poel, University of Twente

Contents

List of Figures

List of Tables

List of Abbreviations

2D	Two-dimensional
3D	Three-dimensional
CNN	Convolutional neural network
Conv	Convolutional layer
CVC	Consonant-Vowel-Consonant
DSC	Dice score coefficient
EMA	Electromagnetic articulography
HPV	Human papilloma virus
MSD	Mean surface distance
PCA	Principal component analysis
PCC	Pearson correlation coefficient
qHD	quantile Hausdorff distance
ReLU	Rectified linear unit
RMS	Root mean squared
ROI	Region of interest
rtMRI	Real-time magnetic resonance imaging
SMAC	Sequential model-based algorithm configuration
TPE	Tree Parzen estimator
USC	University of Southern California
VCV	Vowel-Consonant-Vowel

1

General introduction

1.1 Oral cancer and consequences

Cancer in the lips and oral cavity is prevalent around the world and had an incidence of almost 19,000 in 2018 in Western Europe alone [4]. The incidence of oral cancer is decreasing consistent with the reduction of tobacco use and alcohol intake. However, HPV related tongue cancer is becoming more common, especially in young people [6, 24]. After oral cancer is diagnosed, many factors influence the clinical decision about whether treatment is an option and, if so, what treatment should be offered. For example, the type, the size, and the location of the tumor play an important role in determining the operability. Also, the expected function after the therapy, as well as the physical state and wishes of the patient, are taken into account.

Increasingly, oral cancer is curable with treatments such as surgery, chemo radiation and/or immunotherapy. However, when an extensive resection is necessary to achieve clear margins, function loss can be so severe that the consequences of this treatment might not be acceptable. The functioning of the tongue and the oral area has a major impact on the quality of life [12, 21, 26]. The tongue, oral cavity, and epiglottis are for example essential for eating (Figure 1.1.1). With these structures, the bolus of food is formed, pushed to the back of the mouth, and directed to the esophagus for swallowing [1, 14]. If the patient is unable to do this, a diet of liquids or a percutaneous endoscopic gastrostomy is inevitable. Another important consequence of changes to the vocal tract organs, due to oral cancer and/or intervention, is the decrease in quality of speech, which is regarded in this research. Research from Suarez-Cunqueiro et al. [26] showed that 64% of patients treated for oral or oropharyngeal cancer reported speech problems. While functionality becomes more important, predicting posttreatment impairment remains difficult [11, 27].

1.2 Speech

Decrease in the quality of speech is mostly caused by surgery of the tongue, palate, and jaw, since these are important structures used to articulate. To under-

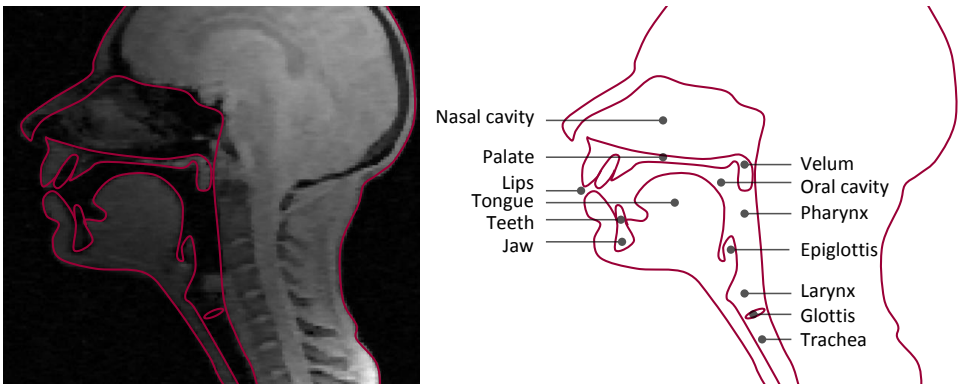


Figure 1.1.1: The anatomy of the head and neck region based on a midsagittal MR slice. The vocal tract ranges from the glottis located in the larynx, including the pharynx and tongue region, to the lips.

stand how speech is altered due to cancer treatment, we use the source-filter model to describe the production of speech [7]. The source-filter model, as illustrated in Figure 1.2.1, assumes the independent production of sound by the 'source' (mainly the glottis region of the larynx) and altering of the spectrum by a 'filter': the vocal tract. The vocal tract can be simplified by envisioning a tube with resonant frequencies determined by the shape of the tube. By changing the shape of the vocal tract, for example, by raising the tongue, the resonance characteristics of the tube change, and a different sound or phoneme is produced. By creating different vocal tract configurations with our articulators, speech is produced [2]. When, due to the tumor or treatment, the patient can no longer make specific constrictions within the vocal tract, speech may be altered [14, 20]. For example, excisions of the lateral side of the tongue may cause lisping with sibilant consonants (/s/, /z/, /ʃ/), as the tongue is no longer able to close off the vocal tract towards the palate sufficiently. Or the quick changes in the articulators' positions cannot be met, required with consonant combinations like /str/ in street.

1.3 Speech therapy

The speech quality is usually assessed subjectively on the intelligibility by speech therapists [17, 22]. Sometimes a more systematic test is performed, trying to iso-

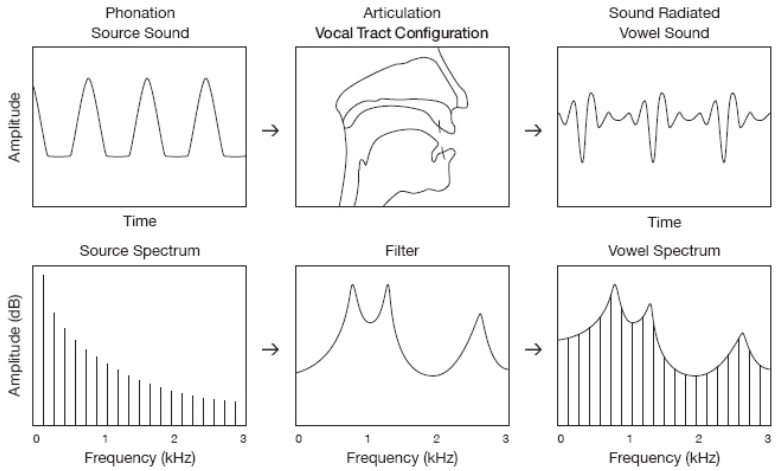


Figure 1.2.1: Illustration of the source-filter model describing the production of speech. The source sound is produced by air flowing through the vocal cords in the glottis. The vocal tract configuration acts as a filter changing the resonance frequencies of the source sound, resulting in speech. This illustration was taken from Maurer, 2016 [16].

late the problematic articulations. Here, the patient reads a text out loud and the therapist keeps score of the words that were not understood or pronounced well. However, these tests rely heavily on the interpretation of the therapist as well, encumbering the rehabilitation process [11, 27].

Gaining awareness of what happens inside the vocal tract may aid in specifying appropriate exercises for speech rehabilitation [27]. There is still much unknown about the relationship between articulation and speech. As an illustration, there are many different articulatory configurations possible for the production of the same sound. This variance causes some patients to be able to compensate for their impairment. Articulatory compensation, however, does not occur with every patient and little is known about what exactly happens with the articulators when compensating [8, 10, 17, 27]. Additionally, compensation occurs on a linguistic level, where patients learn to avoid the sounds they have trouble with and use synonyms instead. This type of compensation may mask articulatory problems, when only judged by listening to free speech.

Additionally, it is currently hard to provide feedback on the progress made, when speech assessment methods are not reproducible or objective. It is known that therapy adherence and patient motivation is improved when the patient receives

feedback on the progression made [9]. Furthermore, patient-specific visualizations of the articulation may function as a therapeutic means to improve speech [3].

Analysis of the articulation will aid in specifying the location and degree of the speech disorder regardless of language. Also, it broadens our understanding of what articulatory compensation mechanisms are possible. In this way, a tool to objectively assess the articulation could contribute to the rehabilitation therapy when speech complaints occur [3].

1.4 Technological challenges

To our knowledge, an automated and objective assessment of articulation has not yet been developed. Although, it is not easy to measure articulation, several techniques are available to acquire articulatory information. Electromagnetic articulography is a method using a magnetic field and electrodes placed in the oral cavity to track the coordinates of these specific locations. The disadvantage is that the setup is complex and electrodes are placed manually thus variety exists between subjects and measurement time points. Furthermore, considering a deformed tongue due to a tumor or surgery, electrode positions cannot be maintained or even placed due to pain. Another methodology is videofluoroscopy, often used for the assessment of swallowing function. Apart from the fact that radiation is needed, videofluoroscopy images have poor soft tissue contrast, making the method not an optimal imaging technique for the vocal tract. [5, 15, 18, 19, 23, 28]

Advances in magnetic resonance imaging (MRI), regarding speed and quality of acquisitions, make it possible to create MR videos, also called real-time MRI (rtMRI). At this point, sufficient quality is reached with single slice imaging (2D), while experiments are being performed to acquire volumetric (3D) data over time [13]. MRI has the advantage of good soft tissue contrast. It gives us information about the complete vocal tract and not a mere selection of points [5, 15, 18, 23]. In this study, we thus chose MRI as the methodology to extract articulatory information.

Additionally, what makes our aim difficult, is the lack of available data, especially pathological speech MR data. Specialized hardware is needed to record speech

during MRI acquisitions, because the large magnetic field distorts electrical signals and the gradient switches cause a lot of noise. For this proof-of-principle study, this hardware was not available, and we decided to use an existing database of healthy subjects to develop the analysis methodology. The database is from the University of Southern California. It contains both 3D static MR imaging during sustained sounds and 2D dynamic MR images of seventeen American English speaking people [25].

1.5 Research aim

In this thesis, we propose a methodology, based on rtMRI, to gain more insights into the articulation of healthy subjects. It is the first time, to our knowledge, that we were able to automatically extract vocal tract information from the segmented images allowing for dynamic analysis of the articulation of subjects. This enables the objective and replicable assessment of speech. Especially intra-subject changes can be monitored and compared, like the impact of an intervention, or the progression due to therapy.

1.6 Research questions and thesis outline

To answer the research aim, we first performed a preliminary study, described in chapter 2, where the goal was to show the power of vocal tract MRI data and deep learning to extract useful information from vocal tract imaging. We train the network to classify 27 different phonemes from 2D MR images. The network's embedding was used to show the relation between these different phonemes that could be distorted when speech is pathological.

The third chapter demonstrates a methodology for segmenting the vocal tract from rtMRI data. The aim was to develop a methodology that needed limited manual labeling effort, while maintaining good segmentation quality. The Chan-Vese level set method was used with Bayesian hyperparameter optimization. With only one manual segmentation per subject, all frames of all videos could be segmented (about 35,000 images per subject). The segmentation works as a feature extractor for further vocal tract analysis.

In the fourth chapter, we explore the question how to use rtMRI data to gain objective and replicable insights into the articulation of a subject. We use the segmentation results of chapter 3 to extract the vocal tract distance function and show characteristics of articulation in different ways.

The thesis finishes with a general discussion in chapter 5, where the research goal is answered. Additionally, suggestions on how to proceed in this field of research are described.

References

[1] (2018). World Cancer Research Fund/American Institute for Cancer Research. Continuous Update Project Expert Report 2018. Diet, nutrition, physical activity and cancers of mouth, pharynx and larynx. Technical report.

[2] Arai, T. (2007). Education system in acoustics of speech production using physical models of the human vocal tract. *Acoustical Science and Technology*, 28(3):190–201.

[3] Beskow, J. (2003). *Talking heads-Models and applications for multimodal speech synthesis*. PhD thesis, Institutionen för talöverföring och musikakustik.

[4] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424.

[5] Bresch, E., Kim, Y.-C., Nayak, K., Byrd, D., and Narayanan, S. (2008). Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]. *IEEE Signal Processing Magazine*, 25(3).

[6] Chaturvedi, A. K., Anderson, W. F., Lortet-Tieulent, J., Curado, M. P., Ferlay, J., Franceschi, S., Rosenberg, P. S., Bray, F., and Gillison, M. L. (2013). Worldwide trends in incidence rates for oral cavity and oropharyngeal cancers. *Journal of clinical oncology*, 31(36):4550.

[7] Fant, G. (1970). *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. Number 2. Walter de Gruyter.

[8] Georgian, D. A., Logemann, J. A., and Fisher, H. B. (1982). Compensatory Articulation Patterns of a Surgically Treated Oral Cancer Patient. *Journal of Speech and Hearing Disorders*, 47(2):154–159.

[9] Govender, R., Smith, C. H., Taylor, S. A., Barratt, H., and Gardner, B. (2017). Swallowing interventions for the treatment of dysphagia after head and neck cancer: A systematic review of behavioural strategies used to promote patient adherence to swallowing exercises. *BMC Cancer*, 17(1):1–15.

[10] Hagedorn, C., Lammert, A., Zu, Y., Sinha, U., Goldstein, L., and Narayanan, S. S. (2013). Characterizing post-glossectomy speech using real-time magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 134(5):4205.

[11] Imai, S. and Michi, K.-i. (1992). Articulatory Function After Resection of the Tongue and Floor of the Mouth. *Journal of Speech, Language, and Hearing Research*, 35(1):68–78.

[12] Kreeft, A., Tan, I., Van Den Brekel, M., Hilgers, F., and Balm, A. (2009). The surgical dilemma of 'functional inoperability'in oral and oropharyngeal cancer: current consensus on operability with regard to functional results. *Clinical Otolaryngology*, 34(2):140–146.

[13] Lim, Y., Zhu, Y., Lingala, S. G., Byrd, D., Narayanan, S., and Nayak, K. S. (2019). 3d dynamic mri of the vocal tract during natural speech. *Magnetic resonance in medicine*, 81(3):1511–1520.

[14] Logemann, J. A., Pauloski, B. R., Rademaker, A. W., and Colangelo, L. A. (1997). Speech and swallowing rehabilitation for head and neck cancer patients. *Oncology-Huntington*, 11(5):651–658.

[15] Mády, K., Sader, R., Zimmermann, A., Hoole, P., Beer, A., Zeilhofer, H. F., and Hannig, C. (2001). Use of real-time MRI in assessment of consonant articulation before and after tongue surgery and tongue reconstruction. In *International Conference on Speech Motor Control*, volume 2001, pages 142–145.

[16] Maurer, D. (2016). *Acoustics of the Vowel-Preliminaries*. PETER LANG LTD International Academic Publishers.

[17] Michi, K. (2003). Functional evaluation of cancer surgery in oral and maxillofacial region: speech function. *International journal of clinical oncology*, 8(1):1–17.

[18] Narayanan, S., Bresch, E., Ghosh, P., Goldstein, L., Katsamanis, A., Kim, Y., Lammert, A., Proctor, M., Ramanarayanan, V., and Zhu, Y. (2011). A multimodal real-time MRI articulatory corpus for speech research. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 837–840.

[19] Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y.-C., Zhu, Y., Goldstein, L., Byrd, D., Bresch, E., Ghosh, P., Katsamanis, A., and Proctor, M. (2014). Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *The Journal of the Acoustical Society of America*, 136(3):1307–1311.

[20] Nicoletti, G., Soutar, D. S., Jackson, M. S., Wrench, A. A., Robertson, G., and Robertson, C. (2004). Objective assessment of speech after surgical treatment for oral cancer: experience from 196 selected cases. *Plastic and reconstructive surgery*, 113(1):114–125.

[21] Schliephake, H., Schmelzeisen, R., Schönweiler, R., Schneller, T., and Altenbernd, C. (1998). Speech, deglutition and life quality after intraoral tumour resection: A prospective study. *International journal of oral and maxillofacial surgery*, 27(2):99–105.

[22] Schuster, M. and Stelzle, F. (2012). Outcome measurements after oral cancer treatment: Speech and speech-related aspects-an overview. *Oral and Maxillofacial Surgery*, 16(3):291–298.

[23] Silva, S. and Teixeira, A. (2015). Unsupervised segmentation of the vocal tract from real-time MRI sequences. *Computer Speech and Language*, 33(1):25–46.

[24] Simard, E. P., Torre, L. A., and Jemal, A. (2014). International trends in head and neck cancer incidence rates: differences by country, sex and anatomic site. *Oral oncology*, 50(5):387–403.

[25] Sorensen, T., Skordilis, Z., Toutios, A., Kim, Y. C., Zhu, Y., Kim, J., Lammert, A., Ramanarayanan, V., Goldstein, L., Byrd, D., Nayak, K., and Narayanan, S. (2017). Database of volumetric and real-time vocal tract MRI for speech science. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2017, pages 645–649.

[26] Suarez-Cunqueiro, M.-M., Schramm, A., Schoen, R., Seoane-Leston, J., Otero-Cepeda, X.-L., Bormann, K.-H., Kokemueller, H., Metzger, M., Diz-Dios, P., and Gellrich, N.-C. (2008). Speech and swallowing impairment after treatment for oral and oropharyngeal cancer. *Archives of Otolaryngology–Head & Neck Surgery*, 134(12):1299–1304.

[27] Sun, J., Weng, Y., Li, J., Wang, G., and Zhang, Z. (2007). Analysis of Determinants on Speech Function After Glossectomy. *Journal of Oral and Maxillofacial Surgery*, 65(10):1944–1950.

[28] Toutios, A. and Narayanan, S. S. (2016). Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research. *APSIPA Transactions on Signal and Information Processing*, 5.

This chapter has been accepted for the proceedings of Interspeech 2019, Graz.

K.G. van Leeuwen, P. Bos, S. Trebeschi, M.J.A. van Alphen, L. Voskuilen, L.E. Smeele, F. van der Heijden, R.J.J.H. van Son (2019), CNN-based phoneme classifier from vocal tract MRI learns embedding consistent with articulatory topology. In *Interspeech 2019*. ISCA.

2

Phoneme classification from vocal tract MRI with a convolutional neural network

2.1 Abstract

Recent advances in real-time magnetic resonance imaging (rtMRI) of the vocal tract provide opportunities for studying human speech. rtMRI together with acquired speech may enable the mapping of articulatory configurations to acoustic features. In this study, we have taken the first step by training a deep learning model to classify 27 different phonemes from midsagittal MR images of the vocal tract.

An American English database was used to train a convolutional neural network to classify vowels (13 classes), consonants (14 classes) and all phonemes (27 classes) of 17 subjects. Classification top-1 accuracy of the test set was 57% for all phonemes. Errors were mostly made between related voiced and unvoiced sounds. We performed principal component analysis on the network's embedding and observed topological similarities between the network's learned representation and the vowel diagram. Saliency maps gave insight into the anatomical regions most important for classification and showed congruence with known regions of articulatory importance.

We demonstrate the feasibility of deep learning to distinguish between phonemes from MRI. Network analysis can be used to improve understanding of normal articulation and speech and impaired speech in the future. This study brings us a step closer to the articulatory-to-acoustic mapping from MR imaging.

2.2 Introduction

Within speech research, it has been a long-standing challenge to be able to estimate the acoustic features corresponding to a specific vocal tract configuration, also called articulatory-to-acoustic mapping. This is not a trivial problem since there is much variability between subjects. Also during speech production, the fast transitions of the articulators are difficult to capture with current measurement methods.

X-ray is one of the methods to extract articulatory information. However, it has the disadvantages of bad soft tissue contrast and potentially hazardous radiation. Electropalatography can only measure when and where the tongue touches

the palate and requires a customized electropalate. In electromagnetic articulography (EMA), sensor coils are placed on the tongue which can cause heterogeneity among speakers and interference with natural articulation. Advances in real-time MRI make it possible to image the whole vocal tract and soft tissue articulators at a sufficient frame rate needed for speech analysis. The advantages of this technique are that no potentially hazardous radiation is needed, and nothing is placed in the mouth that could interfere with the articulators' movement. This technique is also suitable for patients with vocal tract pathology, e.g. patients who have had a (partial) tongue resection or experience pain in these areas. The advantages go at the expense of the increased complexity to acquire the data and patient-level disadvantages such as the high noise level inside an MRI scanner [1, 6, 7, 10]. The University of Southern California gathered a dataset with rtMRI and MRI from sustained sounds (USC Speech and Vocal Tract Morphology MRI Database)[12]. In this study, we use deep learning and single frame MR images of the sustained phonemes to model the relation between the vocal tract configuration and the phoneme.

Saha et al. [9] have previously attempted to classify vowel-consonant-vowel (VCV) combinations from the USC speech database using rtMRI [8, 12]. Image features of multiple frames were extracted and combined with a long short-term memory network to form a general prediction of the VCV-'video'. They reached an accuracy of 42% for 51 different VCV combinations.

This study aims to predict the corresponding phoneme from a static MR image using a convolutional neural network (CNN). We trained a neural network for three tasks: classification of 13 vowels, classification of 14 consonants, and classification of all the 27 phonemes. Secondly, we perform an extensive analysis of what the neural network has learned to gain insights into the relation between vocal tract configurations, speech, and phonetics. These techniques may help us to gain a better understanding in what makes pathological speech abnormal. This study is a preliminary work to demonstrate the feasibility of using rtMRI and deep learning for articulatory-to-acoustic mapping. The larger goal we have is to individualize the approach for people with impaired speech. We want to predict the impact of interventions like surgery and radiotherapy on functional outcomes, such as speech and swallowing, for a more personalized treatment plan. A methodology

Table 2.3.1: Sustained phonemes (in bold) in USC Speech and Vocal Tract Morphology MRI Database.

Sustained vowels	bi:t (beet), bɪt (bit), beit (bait), bɛt (bet), bæt (bat), pa:t (pot), bʌt (but), bɔ:t (bought), boʊt (boat), bu:t (boot), pʊt (put), bɻd (bird), æbʌt (abb**o**t)
Sustained consonants	afɑ, avɑ, aθɑ, aðɑ, asɑ, azɑ, aʃɑ, aʒɑ, ahɑ, amɑ, anɑ, aŋɑ, alɑ, aɹɑ

that is able to perform an articulatory-to-acoustic mapping is essential in order to reach this goal.

2.3 Methods

2.3.1 Data

A published database, the USC Speech and Vocal Tract Morphology MRI Database from the University of Southern California, was used for the classification tasks [12]. The database consists of 2D rtMRI data (1.5 Tesla) including the recorded speech of the vocal tract and 3D volumetric MRI (3 Tesla) while subjects utter sustained vowels and continuant consonants. The latter set was used here for phoneme classification. Data of 17 subjects (8 male, 9 female) were present with 13 vowels (234 images) and 14 consonants (255 images). Further details on the different phonemes are given in Table 2.3.1. For some subjects, a phoneme was missing or a duplicate was present. Overall the dataset was balanced with a mean of 18 ± 0.8 samples per class. For the classification task, only the middle sagittal slice was extracted and used. Each image was unity-based normalized and resampled to 32 by 32 pixels. Data were randomly split on subject level between a train (14 subjects) and a test set (3 subjects).

The Cifar10 dataset [5], existing of 60,000 color images from 10 different classes, was used to pretrain the network for improved image feature extraction. The three color channels were averaged to create grayscale images.

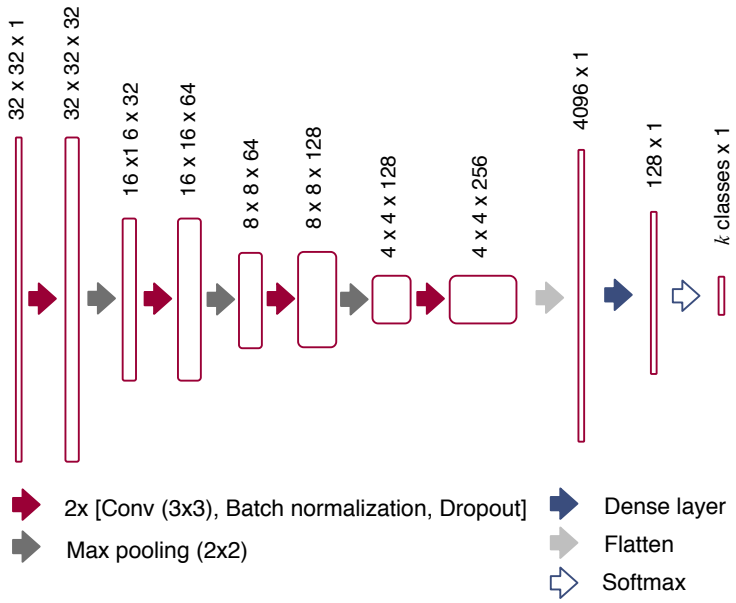


Figure 2.3.1: Classification architecture existing of convolutional layers with kernel size 3 by 3 and max pooling of 2 by 2 for extracting meaningful image features, followed by a dense and softmax layer for phoneme classification with k classes. ReLU activations were present after each convolutional layer.

2.3.2 Architecture and training

The neural network architecture consists of four convolutional blocks. Each block is made up of two convolutional layers with ReLU activations ending with a max pooling layer, as shown in Figure 2.3.1. In the last block, the pooling layer is replaced by flattening to transform the feature maps to vectorial feature space. A softmax classifier is used for the prediction of the phoneme resulting in a probability score for each class.

Because of the limited data in the speech dataset used, the network was pretrained with the Cifar10 dataset to learn general image filters. The training was performed with a batch size of 10 images, early stopping and restoring the best performing model. For the speech classification tasks, the dense layer and softmax layer were replaced by newly initialized layers with the number of output nodes (k) corresponding to the number of classes to predict. The number of possible classes was 13 for the vowel task, 14 for the consonants, and 27 for the vowel+consonant

task. All layers remained trainable and able to be fine-tuned to the MRI data. For both the Cifar10 and the speech dataset loss was defined by the categorical cross-entropy. The network was optimized with Adam optimizer [3].

Six-fold cross-validation was performed for grid hyperparameter tuning with the training set (training on 12 subjects, testing on 2). Hyperparameters were drop-out (0.1, 0.3, 0.5, 0.7), batch-size (1, 4, 8, 16), and amount of data augmentation (with varying amounts of zoom, rotation, shift and shear). Early-stopping with a latency of 30 epochs was applied. For the test phase, the hyperparameter set with the minimum average loss over all folds and over all tasks (vowels, consonants, vowels+consonants) was chosen. Because of the varying number of classes, the number of epochs to train was differed per task and was set to the mean number of epochs of all folds times the factor of 1.3. As the size of training data increases during the test phase, this scaling factor ensures the convergence of the network.

The final performance was determined on each test subject, by retraining the network on all subjects except for the test subject (leave-one-out cross-validation). With this method, as close to all the data could be exploited for the training, while retaining a strict division between train and test samples. This process was repeated ten times to minimize the variance caused by random initialization.

2.3.3 Analysis

Top-1, top-3, and top-5 classification accuracy with standard deviations were computed for each task of the train and test set. The Welch's t-test was performed in order to compare train and test set performance. Probability confusion matrices were computed for the test set with all iterations combined. Principal component analysis (PCA) was performed on the embedded space (output of the flattened layer) to visualize the mapping learned by the network's image feature extractor. Saliency maps highlight regions in the input image contributing most towards the predicted class. They were created by computing the change in the prediction to a small change in the input image, resulting in a sensitivity heatmap over the input image [4, 11].

Table 2.4.1: Leave-one-out cross-validation accuracy of train and test set with batch-size of 4, drop-out of 0.3, and data augmentation with max zoom of factor 0.2, rotation of max 20 degrees, max shift of 0.2 and max shear of 0.2. Average of 10 iterations is given.

		Vowels	Consonants	Vowels + Consonants
top-1 accuracy (stdev) %	train	51.6 (12.7)	52.1 (13.8)	43.0 (10.9)
	test	70.7 (14.1)	61.7 (16.7)	57.0 (8.4)
top-3 accuracy (stdev) %	train	87.3 (7.5)	85.7 (10.6)	76.3 (9.6)
	test	96.2 (2.9)	93.6 (7.6)	89.2 (5.8)
top-5 accuracy (stdev) %	train	97.0 (3.4)	94.2 (7.0)	89.2 (6.0)
	test	100.0 (0.0)	99.1 (1.3)	97.4 (2.2)

2.4 Results

2.4.1 Performance

During hyperparameter tuning, the mean loss was minimized when a batch-size of 4 and drop-out of 0.3 were used, and the data augmentation was performed with a maximum zoom of factor 0.2, a maximum rotation of 20 degrees, a maximum shift of 0.2 and a maximum shear of 0.2.

Table 2.4.1 shows the mean accuracy and variance of all iterations of the different tasks over the train and test subjects. The vowel classification task shows the highest accuracy, 70.7%, on the test set. Correctly classifying the dataset with both vowels and consonants was the most difficult task with 27 classes but, with 57.0% accuracy in the test set, performs well above random chance (±4%). Surprisingly, for all tasks, the test set performance is better than the train set performance, though not significant for most metrics. Only for the top-5 accuracies and the top-3 accuracy for the vowel-task, the difference between the two sets is significant (Welch's t-test, p-value <0.05).

In Figure 2.4.1, the confusion matrix of the consonants shows how voiced and unvoiced consonants with similar articulation are most easily confused, such as /ɑfɑ/-/ɑvɑ/ and /ɑsɑ/-/ɑzɑ/.

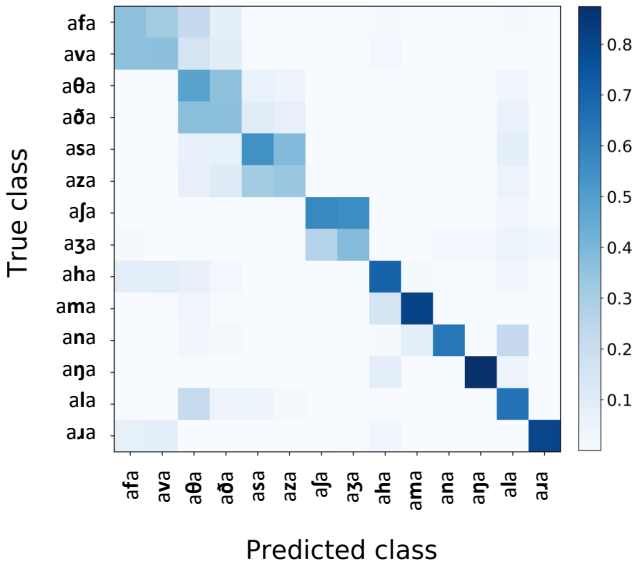


Figure 2.4.1: Confusion matrix of the consonants classification task. Predicted probabilities of the three test subjects of all iterations are combined. Especially where a voiced and unvoiced variant of the phoneme exist confusion can be seen.

2.4.2 Embedding

In Figure 2.4.2a, the first and second principal components of the embedded space of all samples of a vowel model are plotted. We mapped the vowels from the USC speech dataset to the well-studied vowel diagram serving as the legend [2] (Figure 2.4.2b). The PCA plot is rotated to match the orientation of the vowel diagram. Visually, the embedding shows congruence with the orientation of vowels in the vowel diagram, demonstrating that the neural network learned a similar relation between samples as known to phoneticians. It can be seen that the lower vowels, like /*bat*/ and /*pot*/, are oriented at the bottom, as opposed to the higher vowels /*beet*/ and /*boot*/ that are projected at the top.

2.4.3 Saliency

Figure 2.4.3 shows the saliency map of six examples from the same subject derived from the consonants+vowels classification model. Regions in the image light up

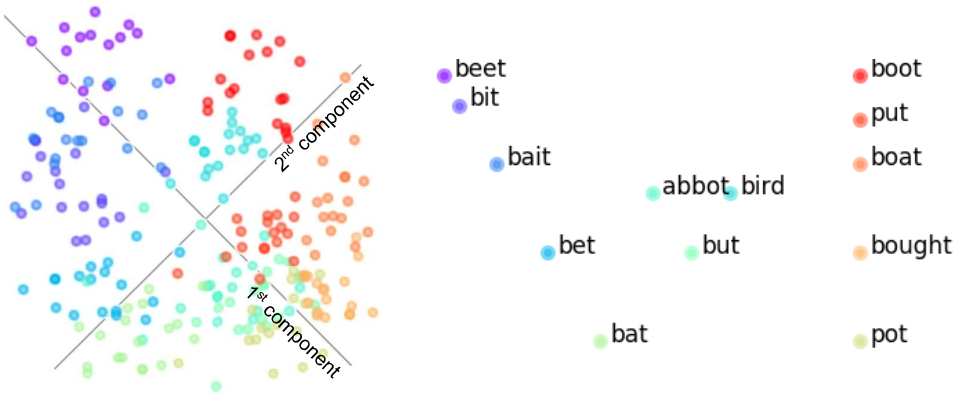


Figure 2.4.2: a. First and second principal component of the output of the flattened layer of all samples based on one of the vowel models. Axes are oriented to align with **b**. **b.** The legend of **a** with the samples mapped in the vowel diagram space. Spatial relations can be seen between the principal components **a** and the well-studied vowel diagram in **b**. Best viewed in color.

where changes in the input have the most impact on the prediction.

For continuant consonants /ɑθɑ/, /ɑfɑ/ and /ɑvɑ/ the lips and tongue are of high importance, and for /ɑmɑ/, /ɑnɑ/ and /ɑŋɑ/ the oropharynx. Vocal tract configurations that are less differentiable from the other samples, like /*bit*/ and /*bat*/, show more widespread attention maps, opposed to well differentiable images like, /*bird*/, /ɑmɑ/ and /ɑhɑ/. Most vowels show a more widespread field between the tongue and palate. Figure 2.4.3e and 2.4.3f show the saliency map of the same input image /*pot*/ at different iterations, which is once classified correctly (2.4.3e) and once misclassified as /ɑhɑ/ (2.4.3f). The saliency maps differ accordingly and help explain why misclassification took place as the sensitivity in figure 2.4.3f is very local and not most meaningful for predicting the true class /*pot*/.

The confusion matrices and embedding for the classification tasks not shown here are included in Appendices A and B.

2.5 Discussion

In this study, we demonstrate that 27 sustained phonemes can be classified from

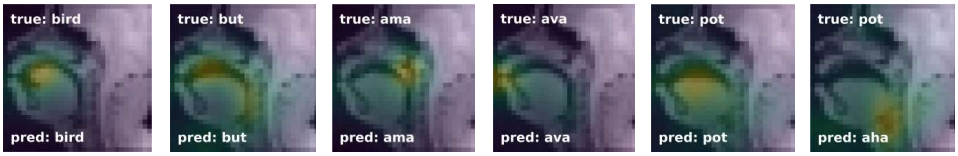


Figure 2.4.3: Saliency maps of one of the subjects. Different input samples are shown with the true class (true) and predicted class (pred). A yellow shade indicates high sensitivity, thus small changes in these pixels in the input have a large effect on the predicted class. Best viewed in electronic form.

MR images by a convolutional neural network with an accuracy of 57.0% on the test set. Our findings suggest that deep learning represents a viable tool for articulatory-to-acoustic mapping from rtMRI.

Apart from the top-1 accuracy, we considered the top-3 and top-5 accuracy. Accuracy increased between 26% and 33% depending on the task for the top-3 accuracy, which indicates that related classes are confused more often. The confusion matrix confirmed this effect. The classification task of both vowels and consonants showed the lowest performance, which can be explained by the fact that twice as many classes were included making the task more difficult.

Most noteworthy, the test set consistently outperformed the train set results. During modeling, it is expected that the dataset, on which hyperparameter tuning was performed, will result in a better performance. One-versus-all cross-validation was performed on each subject to analyze the differences between subjects, while being trained on all other subjects. It appears that the random train/test split gave us three test subjects that outperformed the training set on average. With only 17 subjects in the dataset the risk of having a biased test set is not negligible. The Welch's test shows that the differences in the top-1 accuracy do not significantly differ between the train and test set, thus they come from the same distribution.

The image features learned and visualized in 2D using PCA, demonstrate that the network has indeed learned "sensible" information that resembles the vowel diagram. Furthermore, the saliency maps reveal that the network has learned to focus on the parts of the image that represent the crucial articulatory positions needed to distinguish the different phonemes, such as the lips, tongue and the

oropharynx. The saliency maps were not always similar between subjects since the vocal tract configurations differ with each subject. Moreover, it seems that mistakes were made more often when the saliency maps showed places of sensitivity that were not expected to be important for classification. It would be interesting to apply similar methodology to data of subjects with impaired speech to compare to the healthy subjects. The vowel embedding might reveal insight in the way the different phonemes are related to each other. Saliency maps could aid in explaining which articulators are involved in the impairment of phoneme production.

It is expected that the addition of data of more subjects will improve this research. The risk of getting a biased test set from a random split could be avoided and the model would generalize better. Furthermore, limited experiments have been done on the model architecture due to endless options. The network architecture used in this paper is simple and can be trained with limited computing resources and time. Improvements are possible by using other pre-trained image classification networks as Xception or ResNet.

This research aims to use the speech model in combination with a biomechanical tongue model to better understand and predict the changes in articulation due to oral surgery or radiotherapy. The results of this study give us the confidence to proceed in this direction. The next steps are to develop a method for vocal tract segmentation and use the output to train an articulatory-to-acoustic model.

2.6 Conclusion

The results of this study show the potential of MRI and deep learning as a viable methodology to create a speech model. Analyses of the network provide new insights into what it is that the neural network has learned and 'sees'. This can be used to gain a better understanding of articulation in general and impaired speech in particular.

References

[1] Bresch, E., Kim, Y.-C., Nayak, K., Byrd, D., and Narayanan, S. (2008). Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging

[Exploratory DSP]. *IEEE Signal Processing Magazine*, 25(3).

[2] Decker, D. M. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

[3] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[4] Kotikalapudi, R. (2017). keras-vis.

[5] Krizhevsky, A., Nair, V., and Hinton, G. (2014). The CIFAR-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*.

[6] Mády, K., Sader, R., Zimmermann, A., Hoole, P., Beer, A., Zeilhofer, H. F., and Hannig, C. (2001). Use of real-time MRI in assessment of consonant articulation before and after tongue surgery and tongue reconstruction. In *International Conference on Speech Motor Control*, volume 2001, pages 142–145.

[7] Narayanan, S., Bresch, E., Ghosh, P., Goldstein, L., Katsamanis, A., Kim, Y., Lammert, A., Proctor, M., Ramanarayanan, V., and Zhu, Y. (2011). A multimodal real-time MRI articulatory corpus for speech research. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 837–840.

[8] Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y.-C., Zhu, Y., Goldstein, L., Byrd, D., Bresch, E., Ghosh, P., Katsamanis, A., and Proctor, M. (2014). Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *The Journal of the Acoustical Society of America*, 136(3):1307–1311.

[9] Saha, P., Srungarapu, P., and Fels, S. (2018). Towards automatic speech identification from vocal tract shape dynamics in real-time mri. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1249–1253.

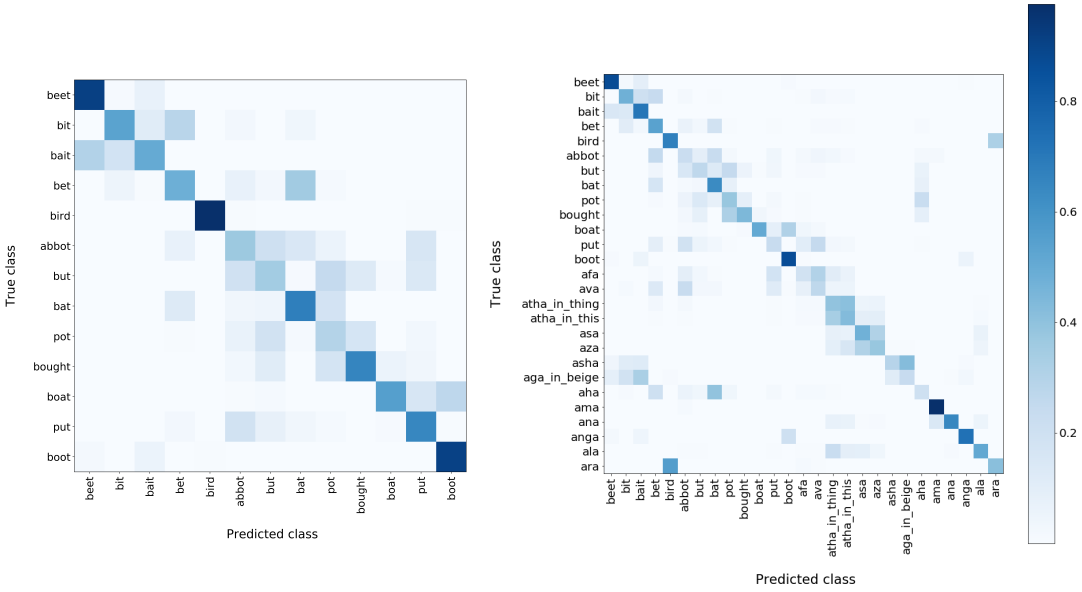
[10] Silva, S. and Teixeira, A. (2015). Unsupervised segmentation of the vocal tract from real-time MRI sequences. *Computer Speech and Language*, 33(1):25–46.

[11] Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

[12] Sorensen, T., Skordilis, Z., Toutios, A., Kim, Y. C., Zhu, Y., Kim, J., Lammert, A., Ramanarayanan, V., Goldstein, L., Byrd, D., Nayak, K., and Narayanan, S. (2017). Database of volumetric and real-time vocal tract MRI for speech science. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2017, pages 645–649.

2.7 Appendices

A Confusion matrices



(a) **(b)**

Figure 2.7.1: Confusion matrix of **a.** the vowel classification task and **b.** vowel+consonant classification task. Predicted probabilities of the three test subjects of all iterations are combined.

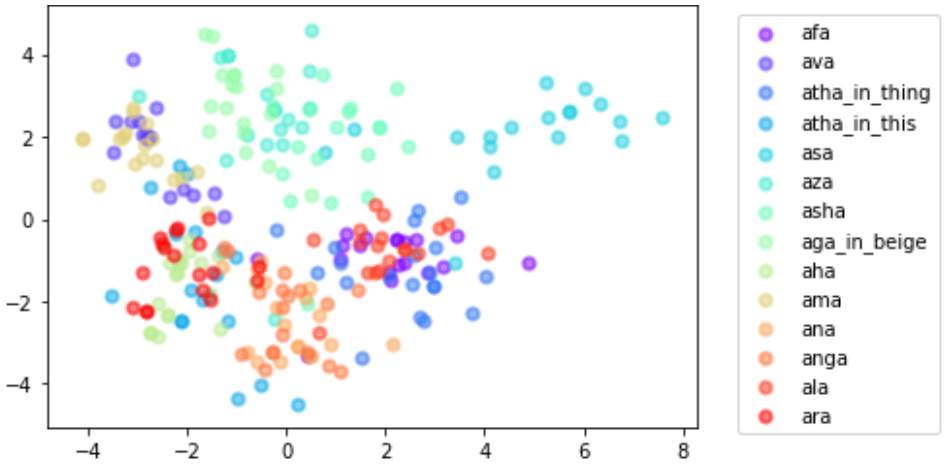


Figure 2.7.2: First and second principal component of the output of the flattened layer of all samples based on one of the consonant models. Best viewed in color.

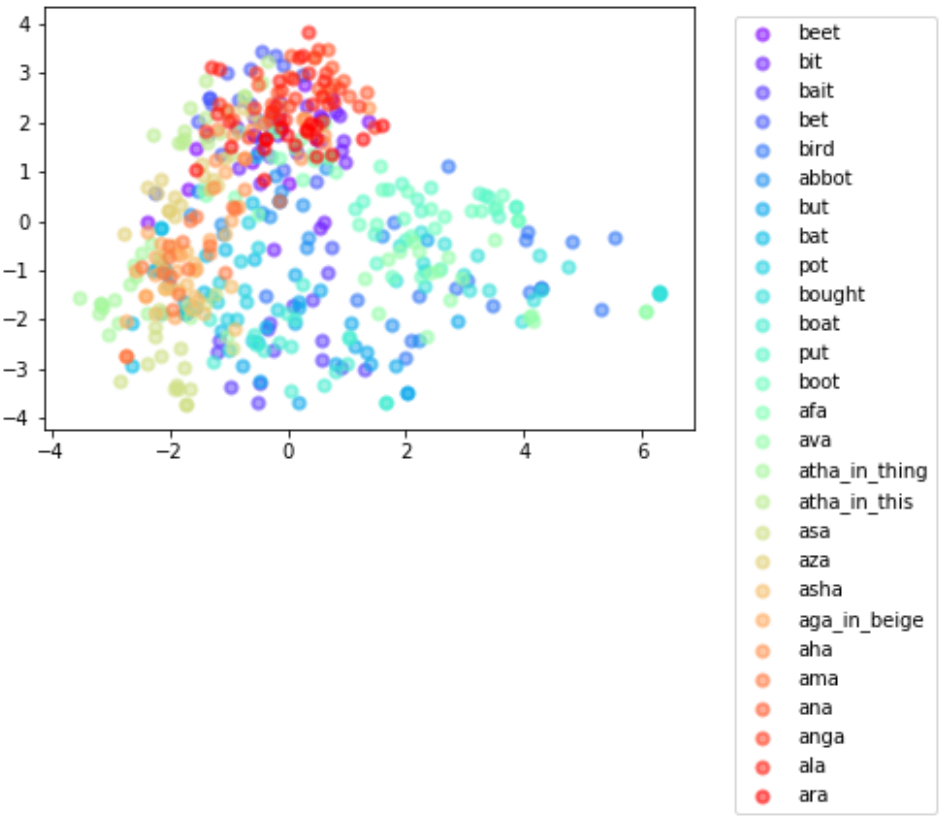


Figure 2.7.3: First and second principal component of the output of the flattened layer of all samples based on one of the vowel+consonant models. Best viewed in color.

3

Segmentation of the vocal tract from real-time MRI with level set method and Bayesian hyperparameter optimization

3.1 Abstract

Real-time MRI (rtMRI) can be used to acquire articulatory information during speech. To enable objective analysis of speech articulation, segmentation of the vocal tract is essential. Here, we aimed to develop a semi-automatic segmentation method that required limited manual labor.

We used the Chan-Vese level set method to segment the images. Bayesian optimization was applied to choose optimal hyperparameters for the image preprocessing and the level set method. Level sets need an initial contour to start evolving from. Experiments were performed to explore the effect of different initialization methods based on manual segmentations. We show that a single manual segmentation per subject results in minimal performance loss compared to a manual segmentation per video. This reduces the manual segmentation work 45-fold, while having a dice similarity coefficient of 95.8 % and a mean surface distance of 1.8 mm.

The method enabled us to segment the data of seventeen subjects with each containing around 40,000 MR frames. The Bayesian approach makes the methodology easily adjustable and transferable to different datasets.

3.2 Introduction

The vocal tract is essential to speech production and therefore the region of interest for studies on speech, articulation and speech pathology. A non-invasive and non-hazardous methodology for dynamically visualizing the vocal tract during speech is real-time MRI (rtMRI). However, to extract quantitative information on the articulators' shape and dynamics, segmentation of the vocal tract is crucial.

Speech production is often described by the source filter model, where the vocal tract acts as a filter on the sound source generated by the glottis [10]. The vocal tract shape, thus, entails a lot of information on the produced speech. Extracting this shape by segmentation of the vocal tract contributes to the long-standing research on articulatory-to-acoustic mapping, and vice versa, acoustic-to-articulatory mapping. It could enable speech simulation, synthesis and aid in rehabilitation therapy for speech disorders and in tracking their progression.

A full-automatic or semi-automatic method for the segmentation of the vocal tract is desired, due to the large amount of data generated by rtMRI. Additionally, an automated method may mitigate high intra- and inter-observer variability caused by the noisy nature of the images [19].

Semi-automatic segmentation methods have been proposed using methods such as active appearance models, level set method, and deep learning. Some studies focus on only the tongue, which limits their usefulness for speech studies [9, 17]. Other researchers have provided evidence for the use of their segmentation method for data of a single subject. However, evidence is lacking for the extrapolation of the method to multiple subjects [5, 18, 27]. Also the amount of manual segmented images needed for the development of the segmentation method is often large. For example, Labrunie et al. [14] use 59 labeled images per subject to create an active shape model. This can lead up to thousands of images when regarding deep learning based segmentation [22, 25]. A more elaborate overview can be found in the Appendix A.

Although many contributions have been made to the research field of vocal tract segmentation, this task remains a challenge to perform with minimal manual segmentation effort. This is caused by the high noise and artifact level of rtMRI and great variation in vocal tract configurations regarding sounds and subjects.

The aim of this study was to segment the vocal tract of each video frame (about 40,000 frames/subject) from a rtMRI database of 17 subjects using the Chan-Vese level set method with only a single manually segmented frame for initialization per subject. To 'learn' the optimal values for the hyperparameters, Bayesian hyperparameter optimization was performed on a subset.

3.3 Methods

3.3.1 Data

The freely-available database from the University of Southern California, the USC Speech and Vocal Tract Morphology MRI Database, was used for the vocal tract segmentation task [23]. It contains 1.5T rtMRI 2D images of the midsagittal slice recorded during speech with a frequency of 23.18 frames per second and resolu-

tion of 68 by 68 pixels. Data of 17 subjects (8 male, 9 female) are present, with a total of 776 videos of varying length ranging from 130 to 2027 frames each (mean 777 frames). The tasks entailed vowel-consonant-vowel (VCV) combinations, consonant-vowel-consonant (CVC) combination, sentences, phrases and free speech. For the method development, only the CVC and VCV videos were used. After the optimization, the method was applied to the complete dataset.

3.3.2 Preprocessing

The signal-to-noise ratio of the image decreases towards the cranium, making it difficult to manually segment the images in this region. Cropping was performed to reduce the impact of the hypo-intense regions towards the cranium. We automatically detected the nose tip through peak detection in the summed intensities over the x- and y-axis. A positive side effect of the per-video cropping is that the anatomies become similarly positioned. All frames were cropped to the same size of 44 by 44 pixels, with the nose tip as reference, and resampled to 98 by 98 pixel images with bilinear interpolation.

An observer performed manual segmentation of three areas, with 3D Slicer [11], denoting the air-tissue boundaries. The first area comprises the tongue, lower lip, jaw and frontal wall of the pharynx (inferior). The second region is bounded by the nose, upper lip, palate and frontal wall of the nasopharynx (superior). The last region is the pharyngeal region (posterior) (Figure 3.3.1). For the method validation, manual segmentations were performed for five subjects, two videos for each subject, and four frames of each video.

3.3.3 Segmentation method

For this study, new methods were explored based on optical flow, image registration, level sets, and an automated deep learning system. The first two were highly dependent of sequential information and previous output, which resulted in the explosion of errors over frames. The last option suffered from the low resolution of the images resulting in sub-optimal segmentations. The level set method was most promising in these experiments and therefore the methodology of choice.

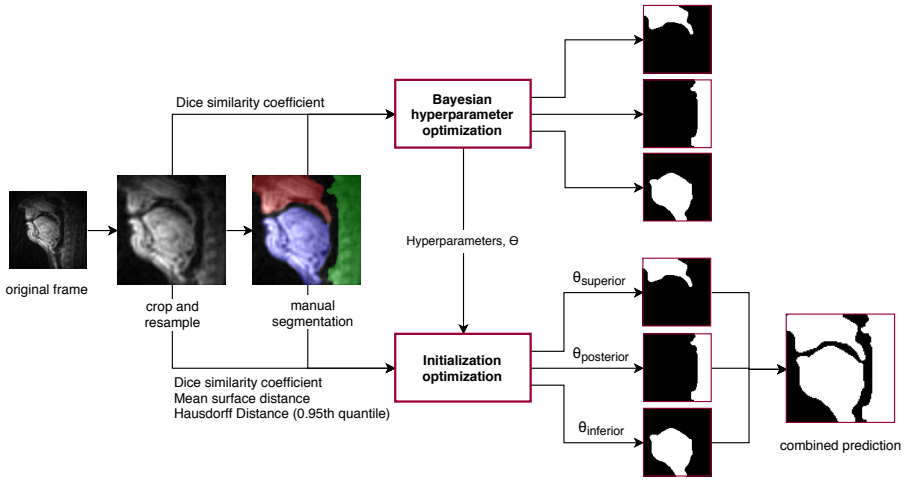


Figure 3.3.1: Overview of Chan-Vese level set method development for vocal tract segmentation. MRI frames were cropped and upsampled. Manual segmentation was performed for a subset of the frames, by delineating the air-tissue boundaries of the superior, posterior and inferior region of the vocal tract. Half of the labeled set was used to find optimal hyperparameters (θ) of the level set method for each region. These hyperparameters were used to test different initialization strategies as explained in Figure 3.3.2. Results were combined to obtain the final vocal tract segmentation.

An overview of the method optimization process is described in Figure 3.3.1.

The Chan-Vese level set based method was used for segmenting the three different regions independently. This algorithm is based on the Mumford-Shah functional and particularly good at segmentation of images without clearly defined boundaries [6]. Level sets evolve in an iterative fashion to minimize an energy (E) dependent on the closed contour (C) and the color intensities inside (c_1) and outside (c_2) the contour (Formula 3.1).

$$E^{CV}(C, c_1, c_2) = \mu L(C) + \lambda_1 \int_{in(c)} (I(x) - c_1)^2 dx + \lambda_2 \int_{out(c)} (I(x) - c_2)^2 dx \quad (3.1)$$

With I being the image and x a point in I. The first term $(\mu L(C))$ is a regularizing term and describes the length of the contour. The terms are weighted with constants μ, λ_1, and λ_2. [2, 6, 18, 26]

Each level set process starts with an initial segmentation. It was attempted to make use of the temporal information and use each predicted segmentation as initial segmentation for the consecutive frame. This approach, however, lead to an unstable system, where errors accumulated and the segmentation quickly blew up, even with hyperparameter tuning. Therefore, it was chosen to use a single segmented frame as initial segmentation for each image independently.

3.3.4 Bayesian hyperparameter optimization

The segmentation methodology has multiple hyperparameters to be set. The value can be chosen manually or with brute force optimization methods like random search or grid search. However, with each extra parameter to be tuned the complexity of the problem scales exponentially, making these methods tedious and very expensive to compute. An alternative way for finding optimal parameter settings is Bayesian hyperparameter optimization [3, 21]. This method is superior due to its capacity of efficiently finding hyperparameters resulting in the optimal model performance. Bayesian hyperparameter optimization is explained in detail in Appendix B.

Sequential Model-based Algorithm Configuration (SMAC) was used for op-

timizing parameters of the segmentation method. SMAC is based on random forests to form a probabilistic model that is used to select subsequent parameter configurations to evaluate [8, 13, 16]. The SMAC Python implementation created by AutoML Freiburg was used [16].

Optimization was performed based on five random videos from five random subjects, considered as the training set. From each video, four random frames were manually segmented. The segmentation of one of the middle frames was taken as the initialization. The other three frames were used for validation. Hyperparameter optimization was performed for each region separately. The dice similarity coefficient (DSC) is an overlap metric further described in Section 3.3.5. The $(1 - DSC)$ was chosen as metric to minimize [7, 24]. The maximum amount of iterations to evaluate different hyperparameter sets was set at 250.

Level set hyperparameters included were μ, λ_1, λ_2, and dt (a multiplication factor to accelerate the algorithm). Also hyperparameters related to adaptive histogram equalization altering the input image were optimized (clip limit and grid size) [4]. The initial segmentation was filtered with an averaging kernel with size k. The maximum number of level set iterations was set at 150 and the tolerance level of the level set method at 10^{-3}. In Appendix C the default values, limits and optimized values are reported.

3.3.5 Initialization optimization

Four experiments were performed to explore the trade-off between the effort of manual segmentations for initializing the level set process and performance. As we deal with over 700 videos to be segmented, we wish to do this with as little manual segmentations possible, with minimal performance loss.

For each subject represented in the training set used for hyperparameter optimization, another random video was picked for validation of the initialization experiments. From the validation video also four random frames were manually segmented. This resulted in a set of five subjects, with two videos each (training and validation), where four frames were manually segmented per video (40 frames in total). Figure 3.3.2 illustrates what frames were used for the initialization and validation at each experiment with one subject as an example.

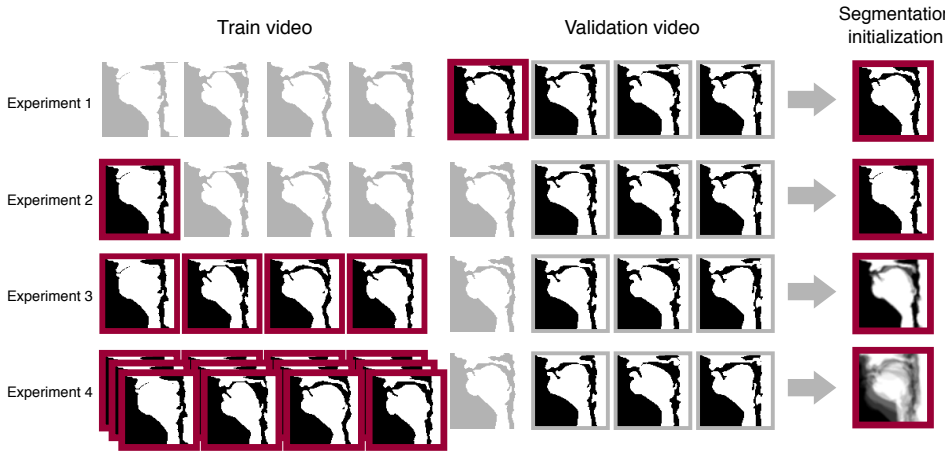


Figure 3.3.2: The figure shows how the initialization frame for the different experiments was determined. An example of one subject is given, copied on each row. The method was applied to the five subjects used for developing the methodology. The train videos represented here were the frames also used for the Bayesian hyperparameter optimization. The red box shows what the initialization frame is comprised of. The grey boxed frames are the validation frames and are constant for each experiment. The light-grey segmentations were disregarded in that experiment.

- **Experiment 1 - Single manual segmentation per video**
  Per validation video, one of the manual segmentations was randomly picked, representing the initial segmentation for that video. This initialization frame was used to segment the other three validation frames. The validation frames are kept constant in each experiment for fair comparison. This is our benchmark method and shows the result if we would choose to manually segment a single frame per video.

- **Experiment 2 - Single manual segmentation per subject**
  Here, we use a manual segmentation from the training video as initial segmentation to segment the validation videos of the same subject. This is done to show the feasibility of having a single initial segmentation per subject that is able to function as an initialization frame for all videos of that subject.

- **Experiment 3 - Multiple manual segmentations per subject**
  We take the four manually segmented frames from the train video and aver-

age them to one grey-scale segmentation map blurred with an average filter (3 by 3 kernel) to evaluate added value of an initialization with different images of the same subject, combining information of different vocal tract configurations.

- **Experiment 4 - Multiple manual segmentations of multiple subjects for all subjects**
  The segmentations from all five subjects from the train video are combined to one initial segmentation map used to segment all validation videos. This map was also convolved with an average filter of kernel 3 by 3 to blur the result. In this experiment, we test the possibility of having a single initialization map to be applied to all videos of all subjects.

The segmentation performances are reported with the overlap metric dice similarity coefficient (DSC) (Formula 3.2)

$$DSC(A_{pred}, A_{true}) = \frac{2\,|A_{pred} \cap A_{true}|}{|A_{pred}| + |A_{true}|} \tag{3.2}$$

with A_{true} and A_{pred} as the manual segmentation areas and predicted areas per subject. When given as a percentage, the DSC was multiplied by 100. Also two distance based metrics were used: the mean surface distance (MSD)(Formula 3.3) and bi-directed quantile Hausdorff distance (qHD) (Formula 3.4)[7, 24]. The MSD and qHD were calculated per frame.

$$MSD\,(U, V) = \frac{1}{2n} \sum_{i=1}^{n} (\min_j d_{euc}(v_i, u_j) + \min_j d_{euc}(u_i, v_j)) \tag{3.3}$$

$$qHD\,(U, V, f) = \max\,(f_i^{th}\,(\min_j d_{euc}(v_i, u_j)),\ f_i^{th}\,(\min_j d_{euc}(u_i, v_j))) \tag{3.4}$$

Here, $U = [u_1, u_2, \ldots, u_n]$ and $V = [v_1, v_2, \ldots, v_n]$ are the predicted and manual labeled point set of the segmentation contours, where the Euclidean distance is calculated over all n points, with i and j being the point indices. The f^{th} quantile was selected as $f = 0.95$ to exclude outliers.

Table 3.4.1: Segmentation results of four experiments, showing the performance on the inferior, superior and posterior region of the vocal tract as well as the total segmentation performance based on the dice similarity coefficient (DSC), mean surface distance (MSD) and 0.95^{th} quantile Hausdorff Distance (qHD).

		1) Single per video	2) Single per subject	3) Multiple per subject	4) Multiple of different subjects
DSC (%)	Inferior	95.6	95.0	95.0	86.1
	Superior	91.2	92.5	89.3	81.7
	Posterior	97.5	97.0	97.2	94.1
	Total	**95.8**	**95.6**	**94.9**	**90.2**
MSD (mm)	Inferior	1.8	2.2	2.2	7.2
	Superior	2.9	2.6	4.0	9.3
	Posterior	1.37	1.7	1.6	4.3
	Total	**1.7**	**1.8**	**2.2**	**5.1**
qHD (mm)	Inferior	10.6	11.4	11.4	29.3
	Superior	15.1	17.6	20.4	44.1
	Posterior	6.46	7.4	7.1	15.0
	Total	**9.1**	**8.8**	**10.4**	**20.6**

3.4 Results

The outcomes of the four experiments are reported in Table 3.4.1. It shows the mean of the DSC, mean of the MSD and the mean of the 0.95^{th} qHD of the different regions independently (inferior, superior, posterior) and the same metrics for the overall binary segmentation map (total). There is only a small performance loss between experiment 1 and 2, with a DSC of 95.8% and 95.6% and a MSD of 1.7mm and 1.8mm respectively for the overall segmentation map. Combining multiple frames as an initial frame (experiment 3) is more manual labor intensive, however, does not result in improved performance over experiment 2. The attempt to use a single initialization map for all subjects (experiment 4) resulted in large deterioration of performance. The standard deviations are given in Appendix D and generally increase with each subsequent experiment.

Regarding the differences between regions, we see that the superior region has

the lowest performance for all experiments and the posterior region the highest. When calculating the metrics over the total segmentation map, the results are better than a mere average. This is due to the fact that regions are independently predicted and can overlap when combined diminishing the error. An example of this situation is demonstrated in Figure 3.4.1d.

Figure 3.4.1 shows example images of the results of experiment 2, comparing the manual segmentation contour (white) with the predicted segmentation contours (red, green and blue dashed line). The areas most prone to error are those where regions touch (e.g. fig 3.4.1d: palate and tongue), especially around the velum (also called soft palate) (fig 3.4.1h), areas of low contrast (fig 3.4.1e, 3.4.1g), and where sharp corners occur (fig 3.4.1f: teeth socket).

An example of the segmentation results of one of the subjects and tasks not used for method optimization is given in Figure 3.4.2. The subject was reading the *Rainbow passage* aloud (4.7.1) of which the first sentence is displayed by every ninth frame.

3.5 Discussion

In this study, we describe a methodology for semi-automated vocal tract segmentation based on level set methods and Bayesian hyperparameter optimization. We find that a single manual initialization segmentation per subject results in limited performance loss opposed to a manual segmentation per video. The DSC reduced only 0.2% going from 95.8% to 95.6%, and the MSD increased 0.1mm from 1.7mm to 1.8mm. Also, the results of these two experiments are within the bounds of each other's standard deviation. This means there is limited added value in providing a manual segmentation for each video opposed to a single manual segmentation per subject, which diminishes the manual segmentation workload 45-fold. The Bayesian approach makes the methodology suitable for transfer to different datasets. With a small reference set, the parameters can easily be tuned for the method to function optimally for the dataset at hand.

It is difficult to compare previous work to this study directly as researchers have

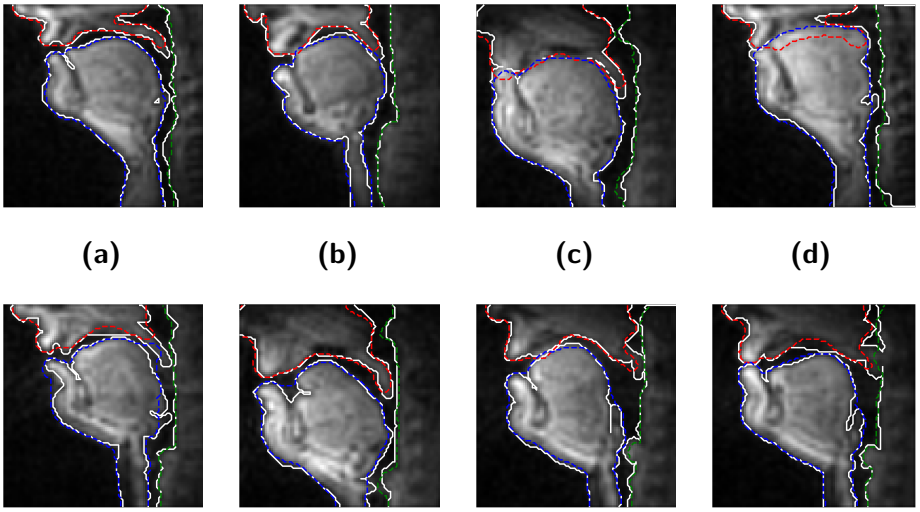


Figure 3.4.1: Examples of different subjects and sequences of the results of experiment 2 are shown. The white line represents the manual segmentation and the dotted colored lines the prediction. **a** and **b** show high quality segmentations. **c** and **d** show similar configurations with different results. In **c** the level set method managed to find the boundary between palate and tongue, which in **d** was not the case for the superior region. This error does not influence the performance of the vocal tract segmentation considered as a whole instead of the separate regions. The bottom row shows how segmentation quality can differ around the area of the velum. Best viewed electronically.

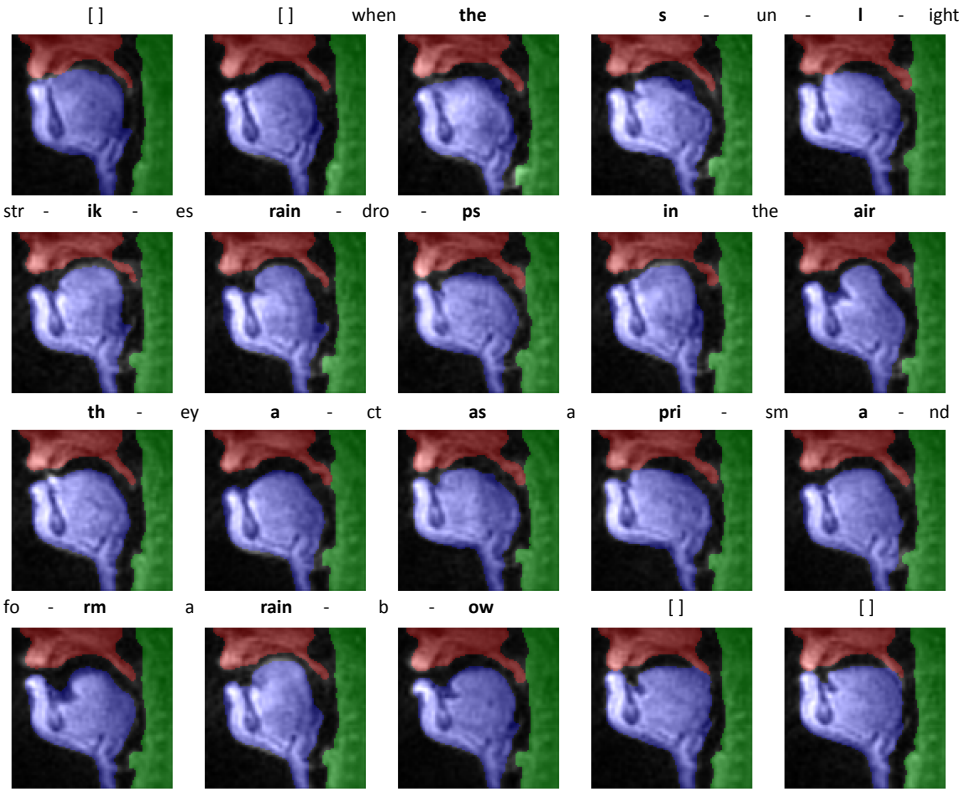


Figure 3.4.2: Example of predicted segmentations of every ninth frame of the first sentence of the *Rainbow passage* read aloud by a subject not used for method optimization. In bold and central above the image is the sound articulated at that frame. The sentence is: 'When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow.' Best viewed electronically.

used different data, approaches, metrics and reference areas. The work of Sampaio et al.[18] also uses DSC as and a distance method (mean Hausdorff distance) related to MSD used here. Results are in a similar range, with DSCs between 87% and 99%. In their method, however, a manual segmentation per video is used. Labrunie et al.[14] showed a high performance with a MSD of less than 1 mm, but used 59 manually segmented images per subject to reach this result opposed to the single frame per subject in this study.

One of the difficulties in segmenting this vocal tract dataset is that the images are of low quality. This is caused by a trade-off when recording rtMRI at a frame rate of 23.18 images per second. The resolution is low, motion artifacts are present, and signal intensity differs heavily over the image. This does not only complicate the process of manually segmenting the images, but also in predicting the right contours with the methodology described here. With the continuous development of MRI hardware and sequences, we believe image quality will improve over time.

Segmentation of the superior region had lowest performance of the three regions, which is mostly caused by the velum. The velum itself is thin and has poor contrast to the oropharyngeal cavity. Furthermore, it has a widely varying position. This made it more difficult for the level set to converge to the right position. For the final application, we manually selected an initialization frame in which the velum had a neutral position instead of randomly appointing one, to improve the segmentation in this region. An alternative approach to mitigate this problem is to pick two initialization frames: one where the velum is open and one where the velum is closed similar to [19]. Together with a classifier determining which initialization to use for which frame, it is likely to get a more accurate segmentation. It must be considered however that this adds complexity and doubles the number of manual segmentations to be performed.

Another improvement that could be made to reduce the segmentation error, is a more robust method for the cropping of the images as part of the preprocessing. As we only use a single manual segmentation per subject, it is important that the orientation of the different rtMRI videos are in line with the manual segmentation used for initializing the level set method. Here, we use a threshold based method to locate the nose tip and crop based on that location. An alternative could be to manually crop the initialization frame and perform registration allowing transla-

tion minimizing an image similarity metric to find the optimal location and translation of the other videos to match the manual segmentation. Also, sensitivity maps of the receiver coils could be estimated to get a more homogeneous signal-to-noise ratio over the image and between different acquisitions.

There is a broad interest for segmented vocal tract data. It can aid in solving the long-standing problem of articulatory-to-acoustic mapping and acoustic-to-articulatory mapping, leading to speech simulation. Recently, researchers from the University of California San Francisco [1], have taken this one step further. They work towards the goal of developing a speech neuroprosthetic where one could 'speak' with the brain. To reach this goal they trained a model to decode neural activity to articulatory information and another model to generate speech from the predicted articulation. For the articulatory information EMA data was used. Vocal tract segmentation from rtMRI data could be used as an alternative means to gain more detailed spatial information on the articulators and potentially an improved model representation.

For the quantitative assessment of speech disorders, segmentation of the vocal tract is essential. It can aid in rehabilitation and progression tracking. In the context of head and neck oncology, knowing the functioning of the articulators before and after interventions, could enable the prediction of the effect of intervention and therapy on speech quality. This may lead to useful information when counseling patients by physicians and speech therapists.

3.6 Conclusion

In this study, we proposed a vocal tract segmentation method where only a single user defined initialization for each speaker is needed for unsupervised operation thereafter for all the remaining frames or images. The resulting method should be easily transferable to other datasets with the Bayesian hyperparameter optimization process in place.

References

[1] Anumanchipalli, G. K., Chartier, J., and Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498.

[2] Avendi, M. R., Kheradvar, A., and Jafarkhani, H. (2016). A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Medical Image Analysis*, 30:108–119.

[3] Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems (NIPS)*, pages 2546–2554.

[4] Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

[5] Bresch, E. and Narayanan, S. (2009). Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE transactions on medical imaging*, 28(3):323–338.

[6] Chan, T. and Vese, L. (1999). An Active Contour Model without Edges. *Scale-Space'99*, LNCS 1682:141–151.

[7] DeepMind (2018). Surface distance based metrics.

[8] Eggensperger, K., Feurer, M., Hutter, F., Bergstra, J., Snoek, J., Hoos, H., and Leyton-Brown, K. (2013). Towards an empirical foundation for assessing bayesian optimization of hyperparameters. In *NIPS workshop on Bayesian Optimization in Theory and Practice*, volume 10, page 3.

[9] Eryildirim, A. and Berger, M.-O. (2011). A guided approach for automatic segmentation and modeling of the vocal tract in MRI images. In *Proceedings of European Signal Processing Conference*, volume 2011, pages 61–65. IEEE.

[10] Fant, G. (1970). *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. Number 2. Walter de Gruyter.

[11] Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., and Sonka, M. (2012). 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic resonance imaging*, 30(9):1323–1341.

[12] Frazier, P. I. (2018). A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*.

[13] Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In *International Conference on Learning and Intelligent Optimization*, pages 507–523. Springer.

[14] Labrunie, M., Badin, P., Voit, D., Joseph, A. A., Frahm, J., Lamalle, L., Vilain, C., and Boë, L. J. (2018). Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning. *Speech Communication*, 99(March):27–46.

[15] Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2016). Hyperband: A novel bandit-based approach to hyperparameter optimization. *arXiv preprint arXiv:1603.06560*.

[16] Lindauer, M., Eggensperger, K., Feurer, M., Falkner, S., Biedenkapp, A., and Hutter, F. (2017). SMAC v3: Algorithm Configuration in Python.

[17] Peng, T., Kerrien, E., and Berger, M.-O. (2010). A shape-based framework to segmentation of tongue contours from MRI data. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 662–665. IEEE.

[18] Sampaio, R. D. A. and Jackowski, M. P. (2017). Vocal Tract Morphology Using Real-Time Magnetic Resonance Imaging. volume 2017, pages 359–366.

[19] Silva, S. and Teixeira, A. (2015). Unsupervised segmentation of the vocal tract from real-time MRI sequences. *Computer Speech and Language*, 33(1):25–46.

[20] Snoek, J. (2012). Spearmint source code.

[21] Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in neural information processing systems*, pages 2951–2959. Curran Associates, Inc.

[22] Somandepalli, K., Toutios, A., and Narayanan, S. S. (2017). Semantic edge detection for tracking vocal tract air-tissue boundaries in real-time magnetic resonance images. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2017, pages 631–635.

[23] Sorensen, T., Skordilis, Z., Toutios, A., Kim, Y. C., Zhu, Y., Kim, J., Lammert, A., Ramanarayanan, V., Goldstein, L., Byrd, D., Nayak, K., and Narayanan, S. (2017). Database of volumetric and real-time vocal tract MRI for speech science. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2017, pages 645–649.

[24] Taha, A. A. and Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15:29.

[25] Valliappan, C. A., Mannem, R., and Kumar Ghosh, P. (2018). Air-tissue boundary segmentation in real-time magnetic resonance imaging video using semantic segmentation with fully convolutional networks. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2018, pages 3132–3136.

[26] van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., and Yu, T. (2014). scikit-image: image processing in Python. *PeerJ*, 2:e453.

[27] Vasconcelos, M. J. M., Ventura, S. M. R., Freitas, D. R. S., and Tavares, J. M. R. (2011). Towards the automatic study of the vocal tract from magnetic resonance images. *Journal of Voice*, 25(6):732–742.

3.7 Appendices

A Literature overview

Table 3.7.1: A non-exhaustive literature overview is given mentioning the segmentation task performed, what type of data was used, on how many subjects, with what segmentation method, how many manual labeled frames were needed, and the performance reported.

	Segmentation target	Data	Number of subjects	Methodology	Manually labeled data needed	Performance
Bresch & Narayanan, 2009	Vocal tract	rtMRI	1	Fourier contour descriptor	1/subject	Not reported
Labrunie et al., 2018	Vocal tract	Static and rtMRI	3	Active Shape Models	59/subject	0.43- 0.65mm MSD
Silva & Teixeira, 2015	Vocal tract	rtMRI	3	Active Appearance Model	51/3 subjects	82% DSC of vocal tract region
Peng et al., 2010	Tongue	Static MRI	4	Chan-Vese level set	1/sound	Not reported
Eryildirim & Berger, 2011	Tongue	Static MRI	4	Chan-Vese level set	1/sound	2.33-3.9 MSD in pixels
Sampaio & Jackowski, 2017	Vocal tract	rtMRI	1	Chan-Vese level set	1/subject/video	87%-99% DSC 2.31mm mean HD
Vasconcelos et al., 2011	Vocal tract	Static MRI	1	Active Shape/Appearance Models	21/subject	4.35-14.23 Euclidean distances (unit not mentioned)
Valliappan et al., 2018	Vocal tract	rtMRI	4	Deep learning (fully convolutional networks)	±6000/4 subjects	±99% pixel accuracy, ±1 pixel Dynamic Time Warping distance

B Bayesian hyperparameter optimization

Bayesian optimization can be applied to find optimal hyperparameters for a given function. The methodology itself is function agnostic and can be applied to all sorts of problems. In this research it was used to find optimal hyperparameters of the level set method for segmenting the vocal tract.

Consider the segmentation method S that is dependent of hyperparameters $\theta_1, \ldots, \theta_n$ drawn from domain $\Theta_1, \ldots, \Theta_n$, with the hyperparameter space defined by $\Theta = \Theta_1 \times \ldots \times \Theta_n$. We try to minimize a metric or loss (L) of S dependent of θ with respect to a constant dataset (D) as seen in Formula 3.5 and 3.6.

$$f(\theta) = L\left(S_\theta, D\right) \tag{3.5}$$

$$\theta^* = \arg\min f(\theta) \tag{3.6}$$

The idea of Bayesian hyperparameter optimization is to search the hyperparameter space Θ in a non-random fashion to find a global optimum resulting in the best performance. A probabilistic model $p_M(f \mid \theta)$ is constructed by point evaluating the function with different hyperparameter configurations (Figure 3.7.1a). What configuration is chosen next to evaluate, is based on an acquisition function using the probabilistic model (Figure 3.7.1b). This acquisition function defines how useful the evaluation of the hyperparameter space is expected to be. The next hyperparameter configuration to evaluate is where the acquisition function is maximal (Figure 3.7.1c). The acquisition functions aim to trade-off between exploration of the hyperparameter space and exploitation of known areas of good performance.[8, 12, 21]

One of the most applied acquisition functions is the expected improvement, $EI_{f_{min}}(\theta)$ in Formula 3.7, which tries to maximize the probability of improvement as well as maximize the size of the improvement at the next evaluation.

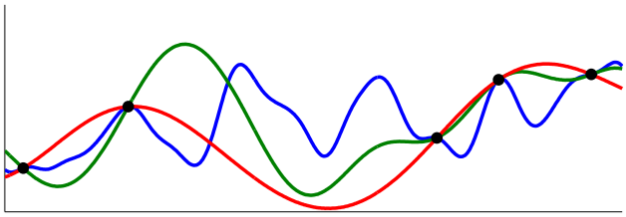
$$EI_{f_{min}}(\theta) = \int_{-\infty}^{f_{min}} \max\left\{f_{min} - f, 0\right\} \cdot p_M(f \mid \theta)\, df \tag{3.7}$$

There are different implementations available of Bayesian hyperoptimization algorithms that mostly differ in the type of model they use to predict $p_M(f \mid \theta)$.

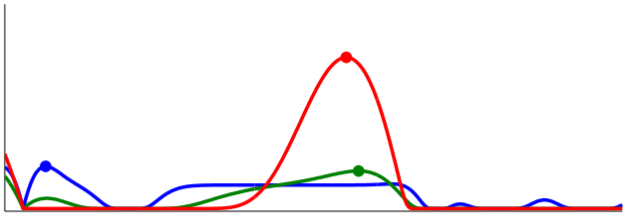
Sequential Model-based Algorithm Configuration (SMAC), Spearmint, and Tree Parzen Estimator (TPE) are some popular examples.

- **SMAC:** SMAC makes use of random forests to model the probability distribution on f as a Gaussian distribution with mean and variance based on the predictions of the forest's trees [8, 13, 16].
- **Spearmint:** this method uses a Gaussian process to model $p_M(f \mid \theta)$ and performs slice sampling over the Gaussian process' hyperparameters [8, 13, 20].
- **TPE:** TPE assumes hierarchy in the hyperparameter space, making it tree-like, hence its name. Also where SMAC and spearmint model $p_M(f \mid \theta)$ directly, the TPE approach models $p_M(\theta \mid f)$ and $p_M(f)$ [3, 8].

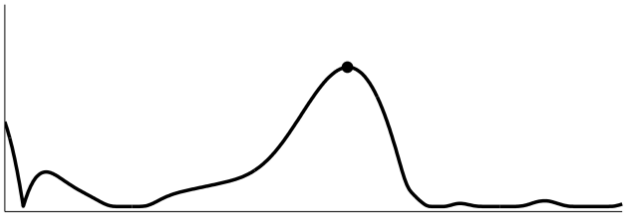
SMAC was used for this study as it showed an overall superior performance on different use-cases. Furthermore, it has an interpretable implementation and supports continuous, categorical and conditional parameters [8, 15].



(a) Posterior samples under varying hyperparameters



(b) Expected improvement under varying hyperparameters



(c) Integrated expected improvement

Figure 3.7.1: Simplified illustration of Bayesian hyperparameter optimization process. **a.** Shows the loss (y-axis) of five different samples with certain hyperparameter (x-axis). The different colored lines represent samples from the probabilistic model $p_M(f \mid \theta)$. The acquisition function is in this example the expected improvement and visualized in **b.** The marked point in **c** shows the next hyperparameter value to evaluate next as it represents the maximal expected improvement. Illustration taken from Snoek et al., 2012 [21].

C Chan-Vese level set hyperparameters

Table 3.7.2: All hyperparameters set or tuned for the Chan-Vese level set segmentation method. Lower and upper limits are given as well as the optimized values for each region to be segmented.

		Default	Lower limit	Upper limit	Superior	Inferior	Posterior
Level set parameters							
μ	Weighing constant length term.	1.2	0.01	1.5	1.34	0.35	0.08
λ_1	Weighing constant for difference of the inner region intensity from average.	1.0	0.1	2.0	1.68	0.97	0.34
λ_2	Weighing constant for difference of the outer region intensity from average.	1.0	0.1	2.0	1.55	0.69	0.21
dt	A multiplication factor applied at each iteration to accelerate the algorithm	0.5	0.5	2.0	1.87	0.60	0.97
tolerance	Stopping criterion: $L^2norm(\varphi_1 - \varphi_0) <$ tolerance	10^{-3}	-	-	10^{-3}	10^{-3}	10^{-3}
iter	Maximum iterations	150	-	-	150	150	150
Image parameters							
clip limit	Contrast limit to prevent noise amplification in adaptive histogram equalization	0.5	0	15	13.7	10.6	14.4
grid size	Tile size over which histograms are equalized for adaptive histogram equalization	4	2	10	8	6	8
Initial segmentation parameters							
k	Kernel size averaging filter (a value of 1 means no filtering takes place)	1	1	10	3	4	1

Table 3.7.3: Segmentation results of four experiments, showing the performance on the inferior, superior and posterior region of the vocal tract as well as the total segmentation performance based on the dice similarity coefficient (DSC), mean surface distance (MSD) and 0.95th quantile Hausdorff Distance (qHD). Standard deviation of the validation frames is given between brackets.

		1) Single per video	2) Single per subject	3) Multiple per subject	4) Multiple of different subjects
DSC (%)	Inferior	95.6 (1.1)	95.0 (2.3)	95.0 (2.3)	86.1 (5.5)
	Superior	91.2 (3.5)	92.5 (2.5)	89.3 (4.8)	81.7 (8.2)
	Posterior	97.5 (1.0)	97.0 (1.0)	97.2 (0.9)	94.1 (4.0)
	Total	**95.8** (0.9)	**95.6** (1.1)	**94.9** (1.4)	**90.2** (3.5)
MSD (mm)	Inferior	1.8 (0.4)	2.2 (1.1)	2.2 (1.1)	7.2 (3.7)
	Superior	2.9 (1.1)	2.6 (1.2)	4.0 (2.1)	9.3 (6.9)
	Posterior	1.37 (0.6)	1.7 (0.6)	1.6 (0.6)	4.3 (3.6)
	Total	**1.7** (0.3)	**1.8** (0.5)	**2.2** (0.8)	**5.1** (2.5)
qHD (mm)	Inferior	10.6 (3.1)	11.4 (4.1)	11.4 (3.5)	29.3 (16.2)
	Superior	15.1 (6.6)	17.6 (10.9)	20.4 (7.8)	44.1 (20.8)
	Posterior	6.46 (2.5)	7.4 (2.6)	7.1 (2.6)	15.0 (7.3)
	Total	**9.1** (2.3)	**8.8** (2.9)	**10.4** (3.4)	**20.6** (9.0)

4

Objective analysis of articulation in speech from real-time MRI

4.1 Abstract

Sound is transformed into speech through articulation with the vocal tract. Oral cancer and its treatments may change the ability to articulate. To objectively assess articulation, we propose a method based on real-time magnetic resonance imaging (rtMRI) of the head and neck region.

Vocal tract segmentations of the rtMRI were used to extract the vocal tract distance function on a frame-to-frame basis. The centerline through the vocal tract was computed on which a grid was projected to find the width of the vocal tract. Different representations are proposed to explain the dynamics and local characteristics of articulation. For example, a heatmap representation demonstrates the dynamics of the vocal tract configuration over time. A text read aloud in different styles could be differentiated from each other through the articulatory space.

With the articulation analysis tool and enrichment of the rtMRI dataset, we can extend our knowledge in the field of speech articulation. Furthermore, it has the potential to be used by speech therapists to aid in speech rehabilitation after oral cancer treatment.

4.2 Introduction

Data that captures articulation is an important source for studying speech production, both in healthy as in pathological situations. Real-time magnetic resonance imaging (rtMRI) is a technique that is able to capture the vocal tract over time with a good spatial resolution. This data allows for the exploration of phonetic principles, like the impact of a specific phoneme on the articulation of the following consonant or vowel, also called coarticulation [17]. Another example is the study of constrictions at specific vocal tract locations to compare different phonemes [5, 10].

Gaining more insights into articulation is of interest clinically to improve speech quality and intelligibility with more targeted therapy. Hagedorn et al. [4] demonstrated the articulatory characteristics of apraxic speech and increased the understanding of the pathological mechanisms underlying apraxia of speech. Some studies have also been performed on patients after partial glossectomy (surgery where

the tongue was partly removed). Researchers observed articulatory compensation mechanisms to maintain speech intelligibility[6, 16]. Mady et al. (2003) [12] showed the differences in the vocal tract distance function between the speech production of /s/ pre- and postoperative.

With the increased amount of available data and growing interest in the clinical application of articulatory analysis, a methodology is needed to analyze and assess rtMRI of speech. This method should enable more objective and reproducible evaluation of articulation, with limited manual annotations involved. Different methodologies have been used to assess articulation so far, which can broadly be defined in four classes: (1) basis decomposition or matrix factorization based techniques, (2) pixel- or region-of-interest (ROI)-based, (3) grid-based, and (4) contour-based [15].

An example of method 1 is the study of Carignan et al. [3]. They applied principal component analysis directly on image data to extract articulatory information, including the movement of short sequences, to explore the differences in articulation between similar vowels. However, the results were difficult to interpret.

The second, ROI-based method, is a common method that uses (changing) pixel intensities in a certain region to show opening and constriction of the vocal tract [2, 5, 10, 13, 18]. No segmentation of the images is necessary, however, ROIs are typically manually defined for a set of locations along the vocal tract having a negative impact on the replicability [15].

The grid-based method superimposes a grid on the image frame based on manually defined landmarks [1, 8, 14, 20]. The landmarks are defined in a way that the gridlines approximate perpendicularity to the vocal tract. The pixel intensities on these gridlines are then used to extract vocal tract distance information. It is a popular method due to its interpretability [15].

With the fourth contour-based method, ROIs in the image are first segmented and sometimes labeled with the different articulator structures. An overview of the contour-based methodologies is given in Chapter 3.2 of this thesis. From the segmentations, vocal tract and articulator representations can be extracted.

One of the main problems with the current methodologies is that they heavily rely on manual annotations. Either regions-of-interests have to be picked, or air-tissue boundaries contoured, or reference points need to be placed to construct

gridlines[15]. Apart from minimizing manual labor to be able to digest the large amount of data and improve reproducibility, we also wish to use all dimensions of the data, both spatial as temporal. In previous research, either the vocal tract cross distance at a particular location is studied over time, or the complete vocal tract distance function is regarded, however only for a single time point. If we want to study the relationship between the articulators over time, both the spatial and temporal dimensions need to be taken into account, which has to our knowledge not yet been done. Last, it is important that the results are interpretable to enable pre- and postoperative comparison by physicians and speech therapists.

We developed an objective methodology that aids in assessing articulation. This tool has been developed taking into account eventual clinical use by speech therapists to improve speech intelligibility for patients treated for oral cancer. We, therefore, focus on the ability to compare pre- and postoperative articulation. The proposed methodology is a combination of the contour- and grid-based method and was used to analyze data from 17 subjects with 25 rtMRI 'videos' each [21]. Data will be made available for further research on the resources page of https://sail.usc.edu/span.

4.3 Methods

4.3.1 Data

For the analysis of articulation, we used the real-time MRI data from the USC Speech and Vocal Tract Morphology MRI Database [21]. All frames (776 videos with a mean of 777 frames per video of in total 17 subjects) were automatically segmented according to the methodology explained in Chapter 3. The database contains videos of vowel-consonant-vowel (VCV) and consonant-vowel-consonant (CVC) combinations, stories and free speech. In this chapter, we demonstrate the articulation analysis methodology based on the repeated recordings of the VCV and CVC combinations and the *Rainbow passage* read aloud in different styles (*normal, loud, whisper, fast, slow*). Table 4.7.1 in Appendix A shows the used recordings.

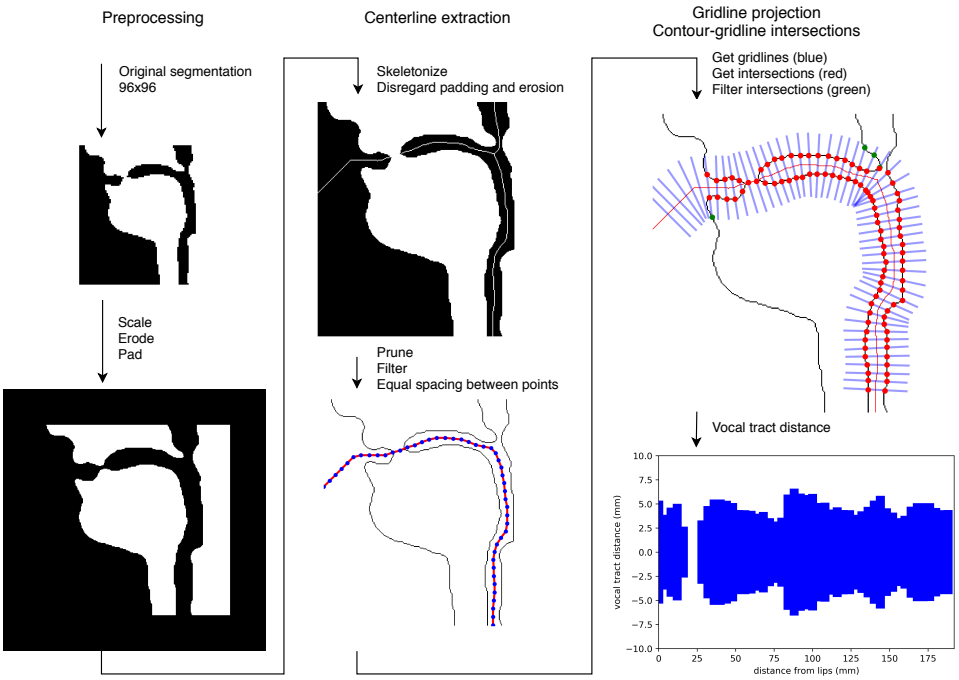


Figure 4.3.1: Schematic overview of the automated frame-by-frame vocal distance extraction methodology. Scaling, padding and spacing parameters were adjusted for improved visualization.

4.3.2 Vocal tract distance extraction

For each frame the vocal tract distance function was extracted. The process, as illustrated in Figure 4.3.1 can be split up in: preprocessing, centerline extraction, projection of gridlines, finding the intersections of the gridlines with the contours, and lastly, calculating the vocal tract distances.

Preprocessing

The segmentation was upscaled 300% with cubic spline interpolation and smoothed with a Gaussian kernel with a standard deviation of two pixels. The three different segmentation areas (superior, posterior and inferior of the vocal tract) were eroded individually to cause a small space between contacting areas, e.g. when the tongue touches the palate as in Figure 4.3.1. This is necessary to find a continuous centerline. The segmentation was zero-padded with 300 pixels on each side to

make the centerline extraction more robust on the edges. The padding and eroding were made undone after the centerline extraction.

Centerline extraction

The segmentation was inverted followed by thinning the vocal tract to reduce the area to a single-pixel wide skeleton [11]. All endpoints of the skeleton, except where it exited the lips and glottis at the edge of the image, were pruned to remove redundant lines, such as the line going through the nasopharynx. A breadth-first search was performed to find the shortest route from the lips to the glottis. This further removed inconsistencies and orders the skeleton pixels to a path. An order-two Savitzky–Golay filter smoothed the line. The centerline was resampled to have equal spacing between every subsequent point with a distance of $0.01mm$, to ensure equal spacing between the gridlines projected in the next step.

Gridline projection

The Frobenius normal on the centerline was computed every k mm, with k set to 1 to have a fairly high resolution at the cost of longer computational time. The normal was calculated using the x and y slope with a window size, $w = 6mm$, to make the normal less sensitive to noise in the centerline. The normals on the centerline form an equally spaced grid from which the vocal tract distances were extracted.

Contour-gridline intersections

The contours from the segmentation and the gridlines were vectorized to find the intersections of the two line types. For the contours, a vector was created between every subsequent point. We iterate over the gridlines and contour segments and search for intersections. To calculate a distance, two intersections of the gridline with the contour are needed, representing the bounds of the vocal tract. When more than two intersections on a single gridline were found, filtering took place by creating vectors between any combination of intersection points. The two intersection points remained if the combination had the shortest distance of all combinations, while the vector between the two points crossed the centerline. In Figure

4.3.1 this situation is demonstrated with the green intersection points around the velum. When no intersection was found, the space was filled with distance 0*mm*, as seen in the Figure at the tongue tip at 25*mm* from the lips. When only a single intersection was detected on a gridline (missing value), the distance value was estimated by linear interpolation based on the neighbouring vocal tract distances. Distance values of gridlines directly next to a missing value were prone to error and therefore also replaced by interpolation. The vocal tract distance sequence starts at the first gridline, considered from the lips, where two intersections were found. All gridlines before that were disregarded, as also seen in Figure 4.3.1 before the lips. The Euclidean distance was calculated between the two intersection points of the gridline and the centerline, representing the vocal tract distance function.

4.3.3 Analysis and visualizations

Articulation dynamics

The power of the dataset and method used here is in the fact that dynamic articulation can be analyzed. To visualize the articulatory dynamics of each VCV and CVC task, they had to be isolated from the video sequence containing multiple tasks. We used peak detection of the collected sound data to label each VCV/CVC task for the start and end frame in the video. In this way, we could extract a specific VCV/CVC combination. The vocal tract distance functions from the correlating frames to the VCV/CVC task were then plotted as a function of distance from the lips and time, with vocal tract distance magnitude in color. For color coding consistency, the magnitude values were clipped at 20*mm*. This methodology was also applied to the *Rainbow passage*.

Articulatory space

The articulatory space was calculated and visualized by aggregating the vocal tract distance function (D) of all frames in a certain set. The mean (mD) and 90% confidence interval were calculated of the *Rainbow passage* for each speaking *style* (*normal, loud, whisper, fast, slow*) and repetition j (Equation 4.1). D_i denotes the vocal

tract distance function of frame i, with the total amount of frames n.

$$mD^{style_j} = \frac{1}{n}\sum_{i=1}^{n} D_i^{style_j} \tag{4.1}$$

To zoom in on the differences with respect to the average of the two normally read passages, we calculated the difference according to

$$\Delta mD^{style_j} = mD^{style_j} - \frac{mD^{normal_1} + mD^{normal_2}}{2} \quad . \tag{4.2}$$

The articulatory space of the *Rainbow passage* was compared for the different speaking styles quantitatively by computing the root mean squared difference (RMS),

$$RMS(mD_p, mD_q) = \sqrt{\frac{1}{2}(mD_p{}^2 + mD_q{}^2)} \quad , \tag{4.3}$$

with $p = q = [1, 2, \ldots, k]$ and k being the number of styles times the number of repetitions, in this case 10. Similarly the Pearson correlation coefficients (PCC) were computed.

$$PCC(mD_p, mD_q) = \frac{k\sum(mD_p\, mD_q) - (\sum mD_p)(\sum mD_q)}{\sqrt{[k\sum mD_p{}^2 - (\sum mD_p)^2][k\sum mD_q{}^2 - (\sum mD_q)^2]}} \tag{4.4}$$

To take a closer look at the constriction and opening of the vocal tract at certain phonemes, we use the data from the CVC and VCV tasks. The VCV and CVC tasks were grouped by consonant, vowel (/ɑ/, /u/, /i/), and articulation location according to the groups described in Figure 4.3.2. The mD was calculated for each group. The differences from the mean of all mD together are given in a heatmap showing where the vocal tract is closed or opened more than average.

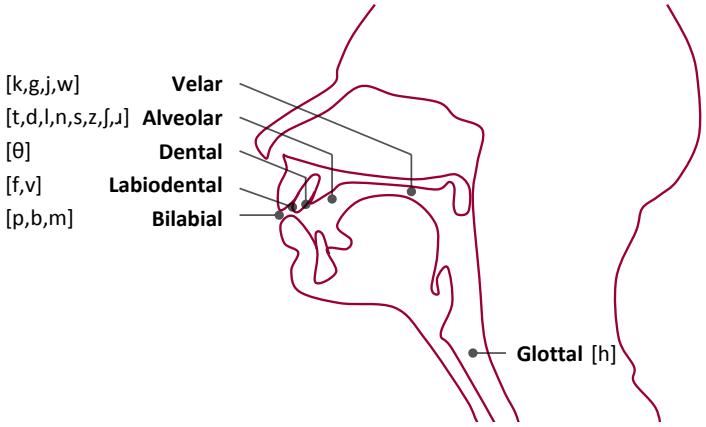


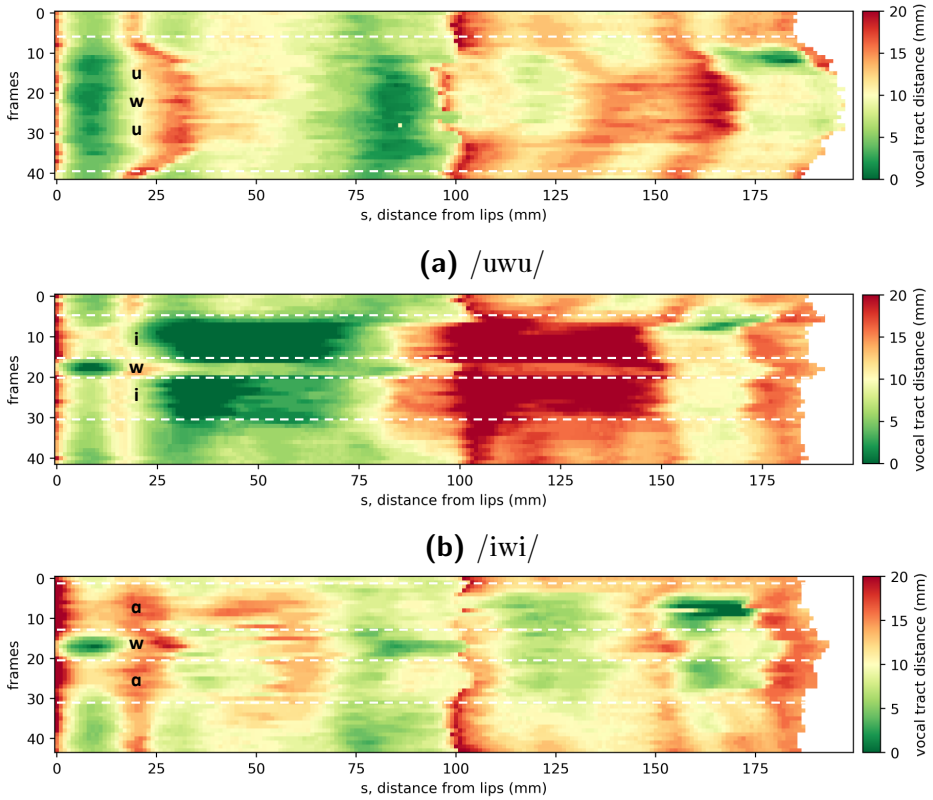
Figure 4.3.2: Different locations of consonant articulation. Examples of consonants for each category are given for English.

4.4 Results

4.4.1 Articulation dynamics

The changes in the vocal tract during speech were visualized with a heatmap as seen in Figure 4.4.1, 4.4.2, and 4.4.3. Figure 4.4.1 shows as an example the VCV combination of /uwu/, /iwi/, /ɑwɑ/. The consonant /w/ was chosen as an example as it illustrates how such data and its visualization can be used to study coarticulation. In Figure 4.4.1a, we see /uwu/ being said between frame 10 and 40 with a stable change throughout the VCV combination, while /u/ and /w/ have a similar vocal tract configuration. In Figure 4.4.1b and 4.4.1c, a clear distinction is visible when the /w/ is pronounced in between the vowels /i/ and /ɑ/. Also, in Figure 4.4.1b, the second /i/ is less profound than the first /i/. This can be explained by the fact that /w/ has a very different configuration than /i/, influencing the subsequent vowel.

Figure 4.4.2 shows another example from two different subjects for the CVC task /rɑr/. The /r/ has many different pronunciations and manners of articulation. In Figure 4.4.2a, the subject uses a bunched postalveolar approximant (/ɹ/) and in Figure 4.4.2c another subject shows a retroflex approximant (/ɻ/). Additionally, Figure 4.4.2a shows how the initial /r/ is articulated differently from the final /r/,



(a) /uwu/

(b) /iwi/

(c) /ɑwɑ/

Figure 4.4.1: A heatmap representation of the vocal tract distance function from the lips (x-axis) over time (y-axis) of the VCV tasks /uwu/, /iwi/, and /ɑwɑ/. The color illustrates the vocal tract distance magnitude. Green represents a small vocal tract distance (constriction) and red a large distance (opening). The effect of coarticulation is demonstrated. **a.** The /u/ and /w/ are similar sounds, and little change in articulation is needed to produce the sounds. **b.** The /i/ and /w/ need different configurations and the second /i/ is less well articulated as observed by a larger vocal tract distance in the velar region. **c.** More symmetric articulation is seen in /ɑwɑ/.

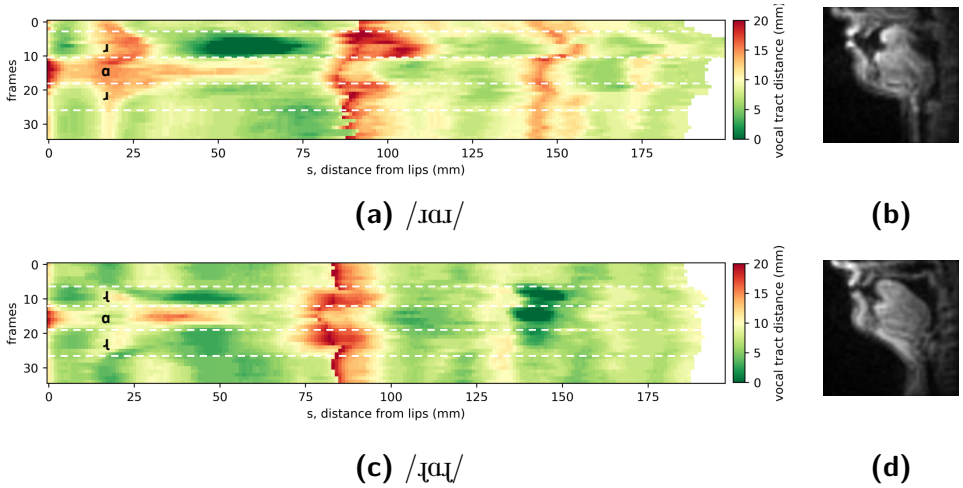


Figure 4.4.2: A heatmap representation of the vocal tract distance function from the lips (x-axis) over time (y-axis) of the CVC task /rɑr/ of two different subjects. The color illustrates the vocal tract distance magnitude. Green represents a small vocal tract distance (constriction) and red a large distance (opening). **a** shows a bunched postalveolar approximant (/ɹ/) with in **b** the frame corresponding to frame 8 in **a**. **c** shows a retroflex approximant (/ɻ/) with in **d** the frame corresponding to frame 9 in **c**.

with increased constriction in the alveolar and velar region for the first /r/.

Figure 4.4.3 illustrates a similar heatmap of the first sentence of the *Rainbow passage*. Different articulatory phenomena can be observed, like the assimilation of the place of the articulation. This happens when phonemes are articulated in a different position, because of the influence of surrounding phonemes.

4.4.2 Articulatory space

Aggregating the vocal tract distance functions from each frame results in an articulatory space profile given (mD) in Figure 4.4.4a. Figure 4.4.4b shows how the style in which the passage was read aloud changes the mean articulatory profile. Figure 4.4.4d shows the differences from the passage read normally (ΔmD) with the standard deviation of the two *normal* repetitions in shaded red. For the largest part along the vocal tract, the articulatory space is larger for the passage read loudly, whispered and slow. Reading the passage fast resulted in a smaller articulatory space than normal. The root mean squared difference (Figure 4.4.4c) and correla-

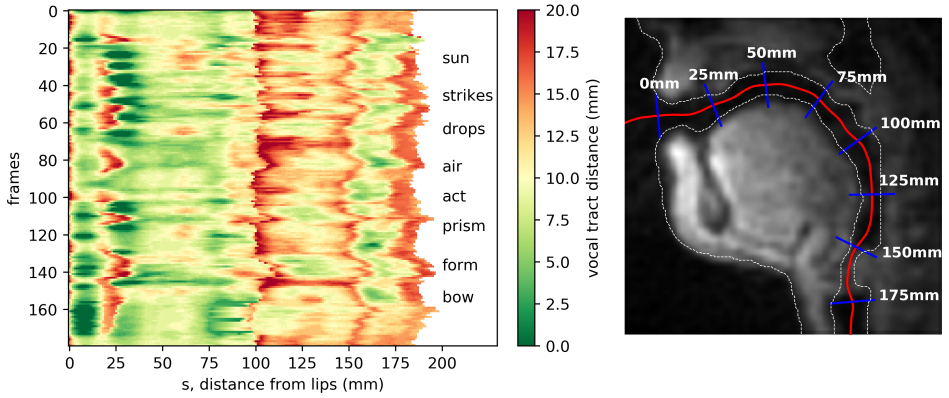


Figure 4.4.3: a. Vocal tract distance function as a function of video frames from the first sentence of the *Rainbow passage* spoken normally. **b.** Illustrates the corresponding anatomy from frame 147 in **a**. The red centerline represents the x-axis in **a**, with in blue the gridlines over which the vocal tract distance was calculated, and in dashed white, the contours of the segmentation. The subject, video sequence and frames correspond to the frames displayed in Figure 3.4.2.

tion coefficients (Figure 4.4.4e) show a similarity between the two repetitions in each style as well as large similarity between *fast* and *normal*. Especially, *whisper* differs largely from normal as well as from the other styles.

Figure 4.4.5 visualizes the articulation space profile of the VCV and CVC tasks grouped in different ways. At every location *s*, the mean of all rows is zero. The color shows the deviation from the mean with red being more open than average, and green signifying more closed than average. In this example, it can be seen that for the bilabial and labio-dental sounds, constriction (green) is present at the beginning of the vocal tract, close to the lips. The constriction moves posteriorly with alveolar and velar consonants. Figure 4.4.5b shows each consonant present in the dataset separately, revealing a clear sequential pattern of constriction amongst the consonants. The bilabial and nasal consonant /m/ shows constriction around the velum (soft palate), caused by the lowering of the velum to let air through to the nasal cavity. In Figure 4.4.5c the vowels are presented. The /ɑ/ and /i/ show inverse behavior, with a more open vocal tract before the velum and more closed vocal tract behind the velum for /ɑ/ and vice versa for /i/.

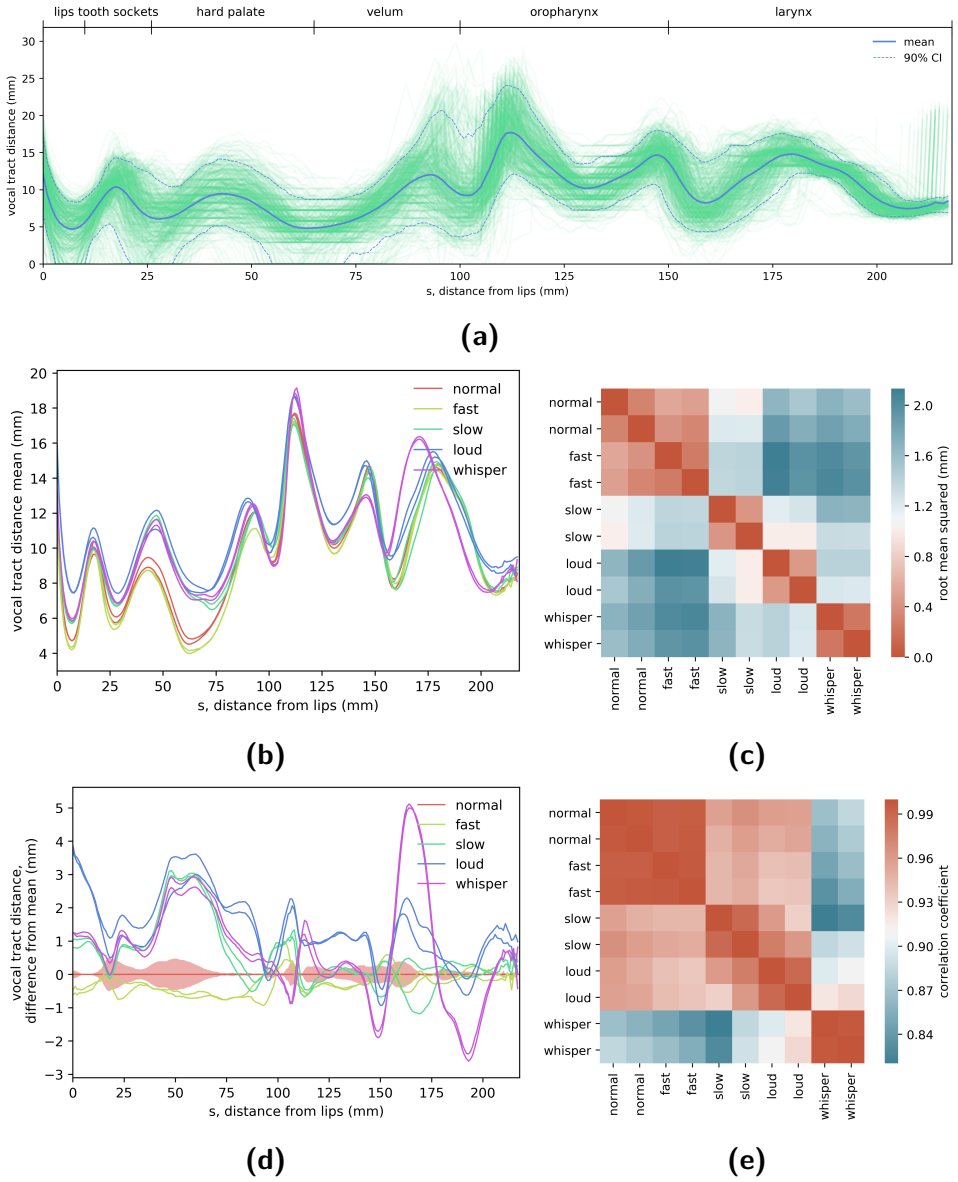
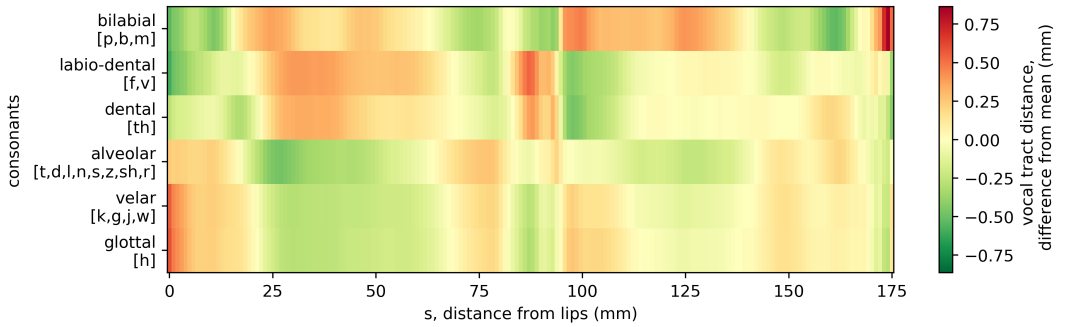
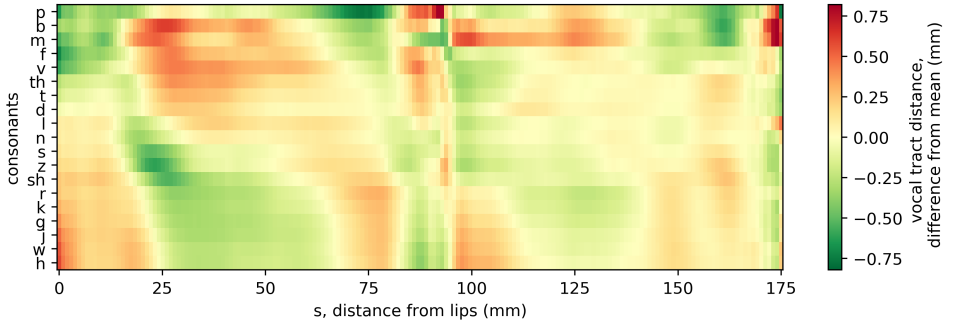


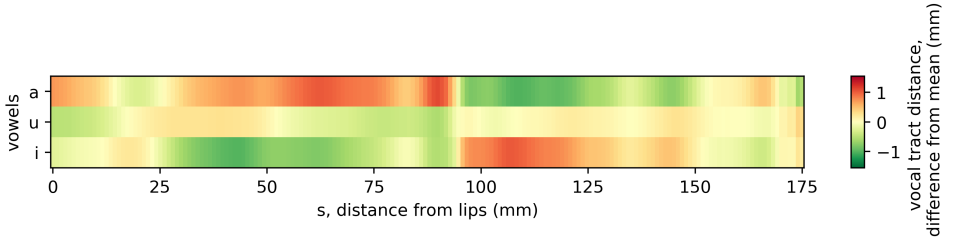
Figure 4.4.4: a. Vocal tract distance functions of each frame of the *Rainbow passage* (read normally), with on the x-axis the distance from the lips towards the glottis, and on the y-axis the vocal tract distance. Each 'line' represents one frame. The blue line is the mean (mD) of all frames, with its 0.9 confidence bounds in dashed lines. **b.** Shows the mean of all frames of the *Rainbow passage* read aloud in the different styles (mD^{normal}, mD^{loud}, $mD^{whisper}$, mD^{slow}, mD^{fast}). There were two repetitions for each style. **d.** Here, the articulatory space means of the *Rainbow passage* from the different repetitions are shown with respect to the *normal* passage (ΔmD^{style}), with the standard deviation of the two *normal* repetitions shaded in red. **c** and **e** show the root mean squared and correlation coefficients of the ten repetitions.



(a)



(b)



(c)

Figure 4.4.5: From a single subject, all VCV and CVC combinations were grouped according to the classes on the y-axis. The mean articulatory space was computed per group. The difference of each mean from all groups aggregated is visualized as a vocal tract distance larger than average (red) and a vocal tract distance smaller than average. **a.** Tasks were grouped according to the location of the consonant articulation. **b.** Tasks were grouped according to the consonant present. **c.** Tasks were grouped according to the vowel present. All frames from the VCV or CVC task were used.

4.5 Discussion

In this study, we described a method that can be used for the objective assessment of speech articulation. The method allows for both holistic analysis of the articulatory space within a complete story as well as detailed information on the vocal tract dynamics of a certain phrase or phoneme. We have shown that subtle changes in the articulatory space can be observed. Here, a similar task spoken in different styles was used to demonstrate this. However, the principle should be translatable to a pathological situation. In such a manner, changes observed in the articulatory space can provide insights into the effect of the therapeutic intervention and/or rehabilitation. Furthermore, we demonstrate that the dynamics of the articulation can well be visualized by a heatmap representation of the vocal tract distance function revealing known articulatory phenomena. The vocal tract distance functions extracted here will become available opening up opportunities to explore characteristics of articulation in multiple healthy subjects, with a diverse set of tasks.

The advantage of the vocal tract distance extraction method described here is that no manual input is needed for the analysis of such a large dataset. Furthermore, the representations proposed here are interpretable, which is important for clinical use. The heatmap visualization was inspired by high-resolution manometry measurements for the assessment of swallowing disorders. Researchers showed that with only a single 20-minute training session, even novice users showed high inter-rater reliability for the analysis of the pressure-based manometry heatmaps [7, 9]. We expect, with some training and regular use, that speech therapists will also be able to interpret and use the articulatory data in clinical practice.

Previously, Silva and Teixeira [19] have also proposed a representation method of the vocal tract configurations. They developed a method, where the data was normalized allowing for not only intra-speaker but also inter-speaker comparison. However, their representation is abstract and especially when visualizing articulation dynamics, only very small time periods can be represented and results are difficult to interpret. Here, we did not aim to allow for comparison between subjects and applied limited normalization techniques. It could be interesting for further research to extend the method to allow inter-speaker comparison. However, the

wide variety of subjects makes solving this problem non-trivial [15]. Apart from the differences in speaker anatomy, the continuously changing shape and length of the vocal tract, and the differences in speed at which speech is produced, encumber normalization.

In articulation studies, it is common to represent the vocal tract distance from the glottis to the lips. Here, for several reasons, we chose to deviate from this commodity by starting from the lips. First of all, it is the practical reason that the glottis is not easy to automatically detect in the rtMRI frames. The resolution of the images is not of sufficient quality to detect the glottis on a frame-to-frame basis. The anatomical structure could be manually labeled, but as the glottis moves as well during speech, this should be done for each frame making it very labor-intensive. Secondly, when considering articulation the focus mostly lies on the tongue area which comprises about the first two-thirds of the vocal tract distance function. Placing this part first draws attention to the area of interest. Important to realize is that neither the glottis or the lips are a static point of the vocal tract and that the length of the vocal tract changes due to the extrusion of both structures. When interpreting the data, it should be taken into consideration that, for example, the lengthening of the vocal tract for the /u/ sound, is shown as an extension at the side of the glottis, while it actually is a lengthening of the vocal tract through extrusion of the lips. Preferably, we would like to have a static reference point along the vocal tract, so the extension of the vocal tract can be acknowledged in both the lip and glottic directions. The hard palate is fairly static over time and could be considered as such a reference point in future research.

Artifacts may occur in the vocal tract distance extraction when the lips are pressed together and the first vocal tract distance calculated from the lips is o*mm*. The first non-zero value is considered the start, meaning the first gridlines are skipped when no intersection or a single intersection only is found. However, no intersection may be found when the lips are pressed together, as the gridline that should be first, might be located within the segmentation completely. In the visualizations, these artifacts are easy to detect and ignore, as suddenly the vocal tract is shorter than the measurements of the surrounding time frames. Using the three individual segmentation contours instead of the combined contour to find gridline intersections might mitigate this problem.

Further improvements may be made in the way interpolation was performed when no intersections between the vocal tract contour and gridlines were found. Vocal tract distances were linearly interpolated on a frame-to-frame basis. Higher-order interpolation methods could lead to overshooting, thus linear interpolation was chosen as a more reliable option. The vocal tract distance functions from surrounding frames were not used for interpolation as we wanted the method to be applicable for single frames only as well, though the interpolation quality would most probably benefit its use.

The vocal tract distance data resulting from the methodology could be a starting point for many types of speech and articulatory research. Think of articulatory-to-acoustic mapping, speech synthesis, studying the articulatory proximity between sounds, different accents, and speakers. To prove its value for clinical use, it is important to acquire pre- and postoperative data of oral cancer patients and apply the methodology in this context.

4.6 Conclusion

Here, we presented an automated methodology for extracting articulatory information from rtMRI during speech. We see the extraction of the vocal tract distance functions as an enrichment of the existing rtMRI dataset that can be further explored for analysis. Here, we presented several ways the data can be further analyzed and visualized, however, possibilities are ample. This methodology for assessing articulation may not only broaden our understanding of speech production but also aid in improving speech intelligibility in oral cancer patients.

References

[1] Asadiabadi, S. and Erzin, E. (2018). A Deep Learning Approach for Data Driven Vocal Tract Area Function Estimation. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 167–173. IEEE.

[2] Bresch, E. and Narayanan, S. (2009). Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE transactions on medical imaging*, 28(3):323–338.

[3] Carignan, C., Shosted, R. K., Fu, M., Liang, Z.-P., and Sutton, B. P. (2015). A real-time mri investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel system of french. *Journal of phonetics*, 50:34–51.

[4] Hagedorn, C., Lammert, A., Zu, Y., Sinha, U., Goldstein, L., and Narayanan, S. S. (2013). Characterizing post-glossectomy speech using real-time magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 134(5):4205.

[5] Hagedorn, C., Proctor, M., and Goldstein, L. (2011). Automatic analysis of singleton and geminate consonant articulation using real-time magnetic resonance imaging. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2011, pages 409–412.

[6] Hagedorn, C., Proctor, M., Goldstein, L., Wilson, S. M., Miller, B., Gorno-Tempini, M. L., and Narayanan, S. S. (2017). Characterizing articulation in apraxic speech using real-time magnetic resonance imaging. *Journal of Speech, Language, and Hearing Research*, 60(4):877–891.

[7] Jones, C. A., Hoffman, M. R., Geng, Z., Abdelhalim, S. M., Jiang, J. J., and McCulloch, T. M. (2014). Reliability of an automated high-resolution manometry analysis program across expert users, novice users, and speech-language pathologists. *Journal of speech, language, and hearing research : JSLHR*, 57(3):831–836.

[8] Kim, Y.-C., Kim, J., Proctor, M., Toutios, A., Nayak, K., Lee, S., and Narayanan, S. S. (2013). Toward automatic vocal tract area function estimation from accelerated three-dimensional magnetic resonance imaging. In *Proceedings ISCA Workshop on Speech Production in Automatic Speech Recognition (SPASR)*.

[9] Knigge, M. A., Thibeault, S., and McCulloch, T. M. (2014). Implementation of high-resolution manometry in the clinical practice of speech language pathology. *Dysphagia*, 29(1):2–16.

[10] Lammert, A. C., Proctor, M. I., and Narayanan, S. S. (2010). Data-driven analysis of realtime vocal tract MRI using correlated image regions. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2010, pages 1572–1575.

[11] Lee, T.-C., Kashyap, R. L., and Chu, C.-N. (1994). Building skeleton models via 3-D medial surface axis thinning algorithms. *CVGIP: Graphical Models and Image Processing*, 56(6):462–478.

[12] Mády, K., Sader, R., Beer, A., Hoole, P., Zimmermann, A., and Hannig, C. (2003). Consonant articulation in glossectomee speech evaluated by dynamic MRI. In *Proceedings of International Congress of Phonetic Sciences*, pages 3233–3236.

[13] Niebergall, A., Zhang, S., Kunay, E., Keydana, G., Job, M., Uecker, M., and Frahm, J. (2013). Real-time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction. *Magnetic Resonance in Medicine*, 69(2):477–485.

[14] Proctor, M. I., Bone, D., Katsamanis, A., and Narayanan, S. S. (2010). Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2010, pages 1576–1579.

[15] Ramanarayanan, V., Tilsen, S., Proctor, M., Töger, J., Goldstein, L., Nayak, K. S., and Narayanan, S. (2018). Analysis of speech production real-time MRI. *Computer Speech and Language*, 52:1–22.

[16] Rastadmehr, O., Bressmann, T., Smyth, R., and Irish, J. C. (2008). Increased mid-sagittal tongue velocity as indication of articulatory compensation in patients with lateral partial glossectomies. *Head & neck*, 30(6):718–726.

[17] Shadle, C., Proctor, M. I., and Iskarous, K. (2008). An MRI study of the effect of vowel context on English fricatives. *Journal of the Acoustical Society of America*, 123(5):3735.

[18] Shosted, R. K., Sutton, B. P., and Benmamoun, A. (2012). Using magnetic resonance to image the pharynx during Arabic speech: Static and dynamic aspects. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2012, pages 2179–2182.

[19] Silva, S. and Teixeira, A. (2016). Quantitative systematic analysis of vocal tract data. *Computer Speech and Language*, 36:307–329.

[20] Skordilis, Z. I., Toutios, A., Töger, J., and Narayanan, S. (2017). Estimation of vocal tract area function from volumetric Magnetic Resonance Imaging. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 924–928. IEEE.

[21] Sorensen, T., Skordilis, Z., Toutios, A., Kim, Y. C., Zhu, Y., Kim, J., Lammert, A., Ramanarayanan, V., Goldstein, L., Byrd, D., Nayak, K., and Narayanan, S. (2017). Database of volumetric and real-time vocal tract MRI for speech science. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2017, pages 645–649.

4.7 Appendices

A Speech tasks

Table 4.7.1: Data from the USC Speech and Vocal Tract Morphology MRI Database focused on in this chapter.

Reference	Task	Repetitions
CVC 1	sɑs-sus-sis, zɑz-zuz-ziz, ʃɑʃ-ʃuʃ-ʃiʃ, θɑθ-θuθ-θiθ	3x
CVC 2	fɑf-fuf-fif, vɑv-vuv-viv, lɑl-lul-lil, ɹɑɹ-ɹuɹ-ɹiɹ	3x
VCV 1	ɑpɑ-upu-ipi, ɑtɑ-utu-iti, ɑkɑ-uku-iki, ɑbɑ-ubu-ibi, ɑdɑ-udu-idi, ɑgɑ-ugu-igi	3x
VCV 2	ɑθɑ-uθu-iθi, ɑsɑ-usu-isi, ɑʃɑ-uʃu-iʃi, ɑmɑ-umu-imi, ɑnɑ-uni-ini, ɑlɑ-ulu-ili	3x
VCV 3	ɑfɑ-ufu-ifi, ɑvɑ-uvu-ivi, ɑɹɑ-uɹu-iɹi, ɑhɑ-uhu-ihi, ɑwɑ-uwu-iwi, ɑjɑ-uju-iji	3x
Rainbow passage	When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long, round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.	2x normal 2x loud 2x whisper 2x fast 2x slow

5

General discussion

5.1 Relevance

The goal of this study was to develop a methodology that enables objective and replicable assessment of articulation in healthy subjects. Real-time MRI (rtMRI) data acquired during speech was analyzed in an automated fashion. We have used deep learning to demonstrate that important information is present in the images that enable phoneme classification, without manually extracting features from the image. Also, the relation between the different vowels became apparent. Next, we developed a methodology based on level set methods to segment the vocal tract from each frame of the USC dataset [13], with only a single manual segmentation per subject. From the vocal tract segmentations, we automatically extracted the vocal tract distance function for each frame. We propose several interpretable ways of visualizing and analyzing the dynamic vocal tract distance function data. This method can be used to broaden our understanding of articulation in speech and to aid speech rehabilitation.

The data enrichment that resulted from this thesis, is a stand-alone result, as well as, a start for further articulatory studies. We are the first, to our knowledge, to propose a method and way of visualization, to extensively analyze articulation over time. Within the field of speech science, this is of great interest to for example understand the influence of certain phonemes on other phonemes (coarticulation)[11], to explore how people apply intonation to voiceless (whispering) speech, or to understand the articulation differences between people. In this thesis, we do not provide answers to these questions, however, by making the segmentation and vocal tract distance function data available, we hope to stimulate other researchers to use the data to support them finding the answers.

Clinically, this thesis is deemed relevant, as there is currently a limited understanding of the impact of oral cancer therapy on the articulation. In the near future, such methodology could be used to acquire knowledge on pathological articulation and the way patients are sometimes able to compensate for their impairment. On a patient-level, having insights into the articulation pre- and postintervention may aid the speech therapists in creating a patient-specific rehabilitation strategy and tracking the progression more objectively. In the long-term, when more pathological speech data has been acquired, we hope we can get a better un-

derstanding of the relation between therapeutic intervention and the quality of speech. Currently, it is hard to predict what the functional consequences are after, for example, a partial glossectomy (tongue resection). Being able to predict the functional outcome, would provide the treating physicians with an indicator for functional inoperability and may thus influence the clinical decision-making process of oral cancer treatment.

5.2 Limitations of current work

The aim was to develop an objective methodology to describe the quality of articulation. In this research, we showed how to extract more quantitative information from the MR images, though, we were not able to come up with a way to objectively score articulation. We have seen that a large variety exists between subjects and it is difficult to compare these subjects. Also, we did not have any pathological speech MRI acquisitions to compare the healthy situation with. Accordingly, we were able to objectively extract information on the vocal tract distance over time, however, the quality of the articulation remains, for now, up to the interpretation of e.g. a speech therapist.

The methodology and data presented here are not yet without error and improvements can be made, as described more extensively in the discussions of the corresponding chapters. The segmentation method, described in Chapter 3, may fail when the orientation of a video deviates too much from the single manual segmentation performed on a different video. The method of cropping and similarly orientating the videos should be made more robust to avoid these artifacts occurring. As for the vocal tract extraction (Chapter 4), artifacts may occur when the lips are pressed together and the first vocal tract distance measurements are skipped as no intersections between the gridline and contours are found. Using the contours of the segmentations of the three regions separately would probably alleviate this problem.

A limitation of rtMRI regarding vocal tract imaging is that the teeth are not visible in the images as they contain very few hydrogen nuclei. The teeth, however, do influence the vocal tract shape and the sound produced, especially for anterior fricative consonants [15]. One way to deal with this is merely to realize the fact

that the teeth are missing when interpreting the data. The segmentation can also be modified in retrospect by modeling the teeth [4, 7], and superimpose them on the segmentation contours. Best would be to avoid the problem, by visualizing the teeth when acquiring the MR images. Researchers [9, 14] have proposed to use MRI contrast agents like ferric ammonium citrate in a bite splint leaving the teeth as signal voids. Superimposition remained necessary as the negative of the teeth was imaged with by the contrast agent. Also, advances in bone MRI are made using, for example, ultrashort echo time imaging [6]. These techniques are, however not compatible with rtMRI.

5.3 Future perspectives

5.3.1 Data

To continue this line of work and prove its clinical applicability, it is important to start gathering patient data pre- and posttreatment of oral cancer. We have experimented with acquiring our own Dutch database of rtMRI using a specialized head-and-neck coil [17] and sequence [16], however, lacked appropriate hardware for the speech acquisition. It is advised to use a fiber-optic microphone with noise-cancelling specialized for speech acquisition in the MRI. A system popular among researchers in this field [1–3, 8, 10] is the FOMRI from Optoacoustics (Optoacoustics Ltd., Moshav Mazor, Israel). A protocol for Dutch speakers was developed and can be found in Appendix 5.4.1.

Another interesting development regarding data acquisition is the research done towards volumetric rtMRI [5]. Currently, we use only the midsaggital view of the tongue and ignore the lateral dynamics of the articulators. However, especially in pathological situations, the vocal tract and its movements might not be symmetric, and a midsaggital view may not be sufficient to assess the articulation. Once, this technology becomes more widespread, the methodology described in this thesis could be extrapolated to a 4D situation, expanding the view of the vocal tract and its articulators.

5.3.2 Inter-subject comparison

In this thesis, we focused on the aim to enable comparing speech within the same subject before and after intervention as well as to monitor the progression of therapy. It has been proposed, to also compare between subjects [12]. To be able to truly quantify the quality of speech compared to a group of healthy speakers, the method described here needs to be extended. Normalization procedures should be improved as subjects vary heavily by their vocal tract anatomy and manner of articulation. Also, variation is present in the timing and speed of the speech produced when considering the dynamics of speech. However, what should be considered before developing this further, is whether being able to compare subjects amongst each other is relevant for the purpose described here. The perception of the quality of speech is very personal. Some patients might have a slight lisp after surgery and are dissatisfied with their speech, while others are satisfied with their speech quality as long as they are understandable to other people. The patient's judgment of speech will, in clinical practice, most probably overrule an objective speech quality score, when deciding on the rehabilitation plan.

5.3.3 Validation

In this study, we chose rtMRI as the means to measure articulation. This technology has many advantages, elaborated on in Chapter 1.4, like a good soft tissue contrast, its non-invasiveness, painless and radiation-free. However, as for clinical application, there are also downsides. MRI acquisitions are costly and time-consuming, moreover many hospitals experience waiting lists due to lack of personnel or hardware. As this method is primarily aimed for rehabilitation, its cost-effectiveness, and therefore clinical feasibility, remains a question.

Apart from demonstrating the cost-effectiveness of the technology itself, the method should also be extensively clinically validated for its added value. Several sessions were held with speech therapists to gather feedback. When pathological speech data has been acquired, I believe the collaboration with the speech therapists should be intensified to make sure, we work to a solution that is relevant and well interpretable. We have, for example, also worked on visualizing the speed of the movement of the vocal tract by calculating the power spectrum over time and

vocal tract location. However, these results were difficult to interpret and therefore most probably of less value in the clinic.

Noteworthy is that the method and proposed application in this thesis is an innovative idea and there is no gold standard or current comparable clinical practice. Assessment of speech is in current practice subjective and based on the observations of the speech therapist and complaints of the patient. Introducing a new methodology to judge the quality of speech, influences the usual clinical practice. We should realize that demonstrating the added value of the method will be important as well as difficult and integration in clinical practice will most probably require time and patience.

References

[1] Fu, M., Barlaz, M. S., Holtrop, J. L., Perry, J. L., Kuehn, D. P., Shosted, R. K., Liang, Z. P., and Sutton, B. P. (2017). High-frame-rate full-vocal-tract 3D dynamic speech imaging. *Magnetic Resonance in Medicine*, 77(4):1619–1629.

[2] Iltis, P. W., Frahm, J., Voit, D., Joseph, A. A., Schoonderwaldt, E., and Altenmüller, E. (2015). High-speed real-time magnetic resonance imaging of fast tongue movements in elite horn players. *Quantitative imaging in medicine and surgery*, 5(3):374–81.

[3] Kryshtopava, M., Van Lierde, K., Defrancq, C., De Moor, M., Thijs, Z., D'Haeseleer, E., Meerschman, I., Vandemaele, P., Vingerhoets, G., and Claeys, S. (2017). Brain activity during phonation in healthy female singers with supraglottic compression: an fMRI pilot study. *Logopedics Phoniatrics Vocology*, 0(0):1–10.

[4] Labrunie, M., Badin, P., Voit, D., Joseph, A. A., Frahm, J., Lamalle, L., Vilain, C., and Boë, L. J. (2018). Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning. *Speech Communication*, 99(March):27–46.

[5] Lim, Y., Zhu, Y., Lingala, S. G., Byrd, D., Narayanan, S., and Nayak, K. S. (2019). 3d dynamic mri of the vocal tract during natural speech. *Magnetic resonance in medicine*, 81(3):1511–1520.

[6] Mastrogiacomo, S., Dou, W., Jansen, J. A., and Walboomers, X. F. (2019). Magnetic resonance imaging of hard tissues and hard tissue engineered bio-substitutes. *Molecular Imaging and Biology*, pages 1–17.

[7] Nakai, S., Beavan, D., Lawson, E., Leplâtre, G., Scobbie, J. M., and Stuart-Smith, J. (2018). Viewing speech in action: speech articulation videos in the public domain that demonstrate the sounds of the International Phonetic Alphabet (IPA). *Innovation in Language Learning and Teaching*, 12(3):212–220.

[8] Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y.-C., Zhu, Y., Goldstein, L., Byrd, D., Bresch, E., Ghosh, P., Katsamanis, A., and Proctor, M. (2014). Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *The Journal of the Acoustical Society of America*, 136(3):1307–1311.

[9] Ng, I., Ono, T., Inoue-Arai, M., Honda, E., Kurabayashi, T., and Moriyama, K. (2011). Application of mri movie for observation of articulatory movement during a fricative/s/and a plosive/t/ tooth visualization in mri. *The Angle Orthodontist*, 81(2):237–244.

[10] Sampaio, R. D. A. and Jackowski, M. P. (2017). Vocal Tract Morphology Using Real-Time Magnetic Resonance Imaging. volume 2017, pages 359–366.

[11] Shadle, C., Proctor, M. I., and Iskarous, K. (2008). An MRI study of the effect of vowel context on English fricatives. *Journal of the Acoustical Society of America*, 123(5):3735.

[12] Silva, S. and Teixeira, A. (2016). Quantitative systematic analysis of vocal tract data. *Computer Speech and Language*, 36:307–329.

[13] Sorensen, T., Skordilis, Z., Toutios, A., Kim, Y. C., Zhu, Y., Kim, J., Lammert, A., Ramanarayanan, V., Goldstein, L., Byrd, D., Nayak, K., and Narayanan, S. (2017). Database of volumetric and real-time vocal tract MRI for speech science. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2017, pages 645–649.

[14] Takemoto, H., Kitamura, T., Nishimoto, H., and Honda, K. (2004). A method of tooth superimposition on mri data for accurate measurement of vocal tract shape and dimensions. *Acoustical science and technology*, 25(6):468–474.

[15] Toutios, A. and Narayanan, S. S. (2016). Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research. *APSIPA Transactions on Signal and Information Processing*, 5.

[16] Uecker, M., Zhang, S., Voit, D., Karaus, A., Merboldt, K. D., and Frahm, J. (2010). Real-time MRI at a resolution of 20 ms. *NMR in Biomedicine*, 23(8):986–994.

[17] Voskuilen, L., Italiaander, M., de Heer, P., Balm, A. J., van der Heijden, F., Strijkers, G. J., Smeele, L. E., and Nederveen, A. J. (2019). A 12-channel flexible receive coil for accelerated tongue imaging. In *Proceedings of the International Society for Magnetic Resonance in Medicine*.

5.4 Appendices

5.4.1 Proposed speech tasks, Dutch

Table 5.4.1: Proposed protocol for Dutch rtMRI database acquisition.

Task	Imaging	Assignment
CVC	2D	sɑs, sus, sis, zɑz, zuz, ziz, (ʃɑʃ, ʃuʃ, ʃiʃ), fɑf, fuf, fif, vɑv, vuv, viv, lɑl, lul, lil, χɑχ, χuχ, χiχ, ʁɑʁ, ʁiʁ, ʁoʁ
VCV	2D	ɑpɑ, upu, ipi, ɑtɑ, utu, iti, ɑkɑ, uku, iki, ɑbɑ, ubu, ibi, ɑdɑ, udu, idi, ɑgɑ, ugu, igi, ɑsɑ, usu, isi, (ɑʃɑ, uʃu, iʃi), ɑmɑ, umu, imi, ɑnɑ, unu, ini, ɑlɑ, ulu, ili, ɑfɑ, ufu, ifi, ɑmɑ, umu, imi, ɑhɑ, uhu, ihi, ɑwɑ, uwu, iwi, ɑjɑ, uju, iji
Sentences	2D	1. Op het gras mag men niet lopen. 2. De zon gaat in het westen onder. 3. De kat van de buren is weg. 4. Het was heel stil in de duinen 5. De rivier trad buiten haar oevers. 6. Moezaam klom de man naar boven. 7. De kat likt het schoteltje leeg. 8. Steile trappen zijn gevaarlijk. 9. De hond blafte de hele nacht. 10. De trein vertrekt over twee uur. 11. Hij rookte zijn sigaret op. 12. De jongen singen er gauw vandoor. 13. De biefstuk is vandaag erg mals.

Passage 1	2D	DE AUTO Er was eens een man uit Finland. Hij had veel geld gespaard. Dat was voor de auto van zijn dromen. Hij nam de trein om de auto te gaan kopen. Maar de man was bang voor dieven. Hij bewaarde het geld in zijn onderbroek. Hij droomde al van de eerste rit in de nieuwe wagen. Plots moest hij naar het toilet. De man dacht niet meer aan het geld. Het zakje met geld viel recht in de pot. En de man spoelde door. Daar ging zijn fraaie plan! Gelukkig was de politie in de buurt. Die vond het zakje terug op de sporen.
Passage 2	2D	DE NOORDENWIND EN DE ZON De noordenwind en de zon waren erover aan het redetwisten wie de sterkste was van hun beiden. Juist op dat moment kwam er een reiziger aan, die gehuld was in een warme mantel. Ze kwamen overeen dat degene die het eerst erin zou slagen de reiziger zijn mantel te doen uittrekken de sterkste zou worden geacht. De noordenwind begon toen uit alle macht te blazen, maar hoe harder ie blies, deste dichter trok de reiziger zijn mantel om zich heen; en ten lange leste gaf de noordenwind het op. Daarna begon de zon krachtig te stralen, en hierop trok de reiziger onmiddellijk zijn mantel uit. De noordenwind moest dus wel bekennen dat de zon van hun beiden de sterkste was.
Sustained vowels	3D	bɑd (bad), bɛd (bed), bɪt (bit), bɔt (bot), pʏt (put), bɑːt (baat), beːt (beet), bit (biet), boːt (boot), byt (buut), bɛit (bijt), høp (heup), buk (boek)
Sustained consonants	3D	ɑfɑ, ɑvɑ, ɑsɑ, ɑʃɑ, ɑzɑ, ɑmɑ, ɑnɑ, ɑʁɑ, ɑχɑ

6

Acknowledgments

Employees of the Antoni van Leeuwenhoek hospital,
Thank you all for being empathetic, open healthcare professionals that try to make the lives of people that have or have had cancer better. Thank you for being so approachable and willing to teach me what you know.

Paula,
Thank you for always being available for a chat or helping me connect to people in the hospital. You were maybe appointed for helping me out content-wise, but you noticed that what I needed more was someone that asked me how I was doing and if I recently performed some clinical activities. It was good to have you around to keep me on track both personally as professionally.

Rob,
Thank you for introducing me to the big, new world of linguistics. Never realized before how my tongue does the same thing when saying a /*w*/ and a /*u*/. Or that the fact that we can put intonation in our voiceless speech when whispering is still an unsolved problem. It was great to have your enthusiasm at my meetings. Thank you for being able to give me a motivation-boost when I lost sense of it all. I am happy you joined my exam committee halfway my internship. You belong in there!

Ludi,
Thank you for the trust you had in me, whether that was when assisting you amputating an ear, or regarding the research I did. Thank you for always making time in short term, when I asked for your input.

Ferdi,
Especially in the first half of the internship we have been in touch a lot, where you gave me great ideas on the technological approach. You helped me not to overflow my ambition and to stay aware of the different interests at stake. I will definitely take your advise regarding focus with me for the future.

Paul,
You say what you think and even though I did not always enjoy to hear that, I have

always appreciated it. I know it came from a genuine worry or believe. Usually, it was just a matter of time for me to realize you were absolutely right. I want to thank you for the always very quick response on my written thoughts. They made me feel understood and supported.

Mannes,
You have been my machine learning mentor since I knocked on your door to ask you to be my supervisor for the online machine learning course. It has been great to have you around in my learning journey the past three and a half years. Thank you for being there to fall back on in my first machine learning projects, and later to give me the opportunity to be a teacher assistant and even give a guest lecture in your course. I have not seen anyone happier than you with the message that I was going to stick around in academia for a bit longer. Thank you for being my biggest fan.

Luuk,
Thank you for the elaborate feedback that showed me how to write a proper paper. Every chapter in this thesis and that I will write in the future has become better because of you. I will never look the same at the word *'can'*. Also, thank you for the time you spent with me on trying to make the rtMRI and speech recordings work in the AUMC. All the best wrapping up your dissertation!

Maarten,
You have both been amazing in providing me with new ideas and great feedback to my thesis. Thank you for having taken the time to be there at my meetings, showing such interest, and always being there to help. The heatmap manometry idea from you has really made Chapter 4.

Kilian,
Thank you for all the peer-pressure in making me drink more coffee than I wanted. Just kidding, was great to have you as an office-mate. It was nice to have each other as a moral support and to share whatever with. I enjoyed listening to your rabbit and air-conditioning stories and am grateful for you always looking at my newly

created figures with interest when I really wanted to show it to someone.

Bence,
It was fun to have you as an office-mate and great to have another speech researcher around. I am pretty sure the winner of the nerd-points will be you and the price stroopwafels. Good luck enduring the companion of Kilian for another year or so and hope you can resist his coffee-peer-pressure after quitting.

Stefano,
The hero computer scientist in a clinical center. You can stop worrying now about whether I liked the internship: I liked it (and yes, I am even staying in research, who would have thought). Thanks for being my unofficial deep learning supervisor. Even though the whole thing ended up not to be much deep learning, having had you around in the beginning was definitely helpful and enjoyable! Other than that, I really liked the deep learning journal clubs you organized.

Rita,
Thank you for your interest and attendance at the 6-weekly thesis meetings. I really appreciate your help and enthusiasm, especially since you are in a completely different department.

Klaske and Anne,
Many thanks for having taken the time to give feedback on the results. Your experiences as speech therapists provided me with valuable context for the clinical problem being faced.