Deep-Learning-Based Methods for Automatic Articulator and Levator Veli Palatini Segmentation and Motion Quantification in Magnetic Resonance Images of the Vocal Tract

Ruthven, Matthieu

Awarding institution:
King's College London

KING'S *College* LONDON

School of Biomedical Engineering & Imaging Sciences

Deep-Learning-Based Methods for Automatic Articulator and *Levator Veli Palatini* Segmentation and Motion Quantification in Magnetic Resonance Images of the Vocal Tract

Matthieu Ruthven

Supervised by Dr Andrew King and Dr Marc Miquel

PhD Thesis

July 2023

Abstract

Articulators such as the soft palate play an essential role in the production of speech. In combination with the *levator veli palatini* (LVP), the soft palate causes velopharyngeal closure, a key requirement for the production of most speech sounds.

Velopharyngeal insufficiency (VPI) is an anatomical or structural defect that prevents velopharyngeal closure and consequently impairs speech. While several well-established surgical techniques to treat VPI exist, there is currently no consensus on which is most effective and consequently a variety of techniques are used. In addition, treatment is not always successful and further surgery is required.

Typically in clinical assessments of speech, imaging is used to enable identification of the defects preventing velopharyngeal closure and inform the choice of treatment. While currently the most commonly used imaging techniques are videofluoroscopy and nasendoscopy, use of magnetic resonance imaging (MRI) is increasing due to its unique ability to dynamically image the articulators during speech and acquire detailed three-dimensional (3D) images of the LVP. In addition, there is increasing interest in extracting quantitative information about the vocal tract, articulators and LVP from the images. The work presented in this thesis makes several contributions towards addressing the unmet need for this quantitative information.

Segmentation of medical images is a common first step to enable automatic measurement of anatomical features. In the work presented in this thesis, two segmentation methods, both of them deep learning based, were developed and evaluated. One method segments the vocal tract, soft palate and four other relevant anatomical features in two-dimensional (2D) magnetic resonance (MR) images of speech. At the time it was published, the method overcame the limitations of existing segmentation methods that either only segmented air-tissue boundaries between the vocal tract and adjacent tissues or only fully segmented the vocal tract. The other method segments the LVP and pharynx in 3D MR images of the vocal tract.

In addition, a framework for quantification of articulator motion in 2D MR images of speech was developed and evaluated. This deep learning framework for nonlinear registration builds on the 2D image segmentation method by employing knowledge of

region boundaries as well as images to estimate displacement fields between 2D MR images of speech. The framework was compared with several state-of-the-art traditional registration methods and deep learning frameworks for nonlinear registration and found to estimate displacement fields that more accurately captured velopharyngeal closures.

To enable the development and evaluation of the segmentation methods and motion quantification framework, a new dataset of 15 3D MR images of the vocal tract was acquired and ground-truth (GT) segmentations were created for it and an existing dataset of 392 2D MR images of speech. Prior to acquiring the new dataset, an investigation was performed to identify the parameters that resulted in the optimal image contrast for LVP visualisation.

To be suitable for use in clinical speech assessment, a key requirement of segmentation and motion quantification methods is that they capture any velopharyngeal closures that occur. Since standard evaluation metrics do not provide such information, a novel metric based on velopharyngeal closure was developed to enable more clinically relevant evaluation. Particularly in the comparison of motion quantification frameworks, the metric revealed differences between the frameworks that standard metrics did not.

To conclude, while future work is required to fully address the unmet need for quantitative information about the vocal tract, soft palate and LVP in MR images, the work presented in this thesis has nevertheless contributed towards addressing this need and created several new opportunities to contribute to the ultimate goal of improving the treatment outcomes of patients with VPI.

Acknowledgements

Matthieu Ruthven (Clinical Doctoral Research Fellow, ICA-CDRF-2018-04-ST2-032) is funded by Health Education England (HEE) / National Institute for Health and Care Research (NIHR) for this research project. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR, NHS or the UK Department of Health and Social Care.

I wish to express my deep gratitude to several people for enabling the work presented in this thesis. First, I wish to thank HEE and the NIHR for funding this work. Second, I wish to thank my Thesis Progression Committee at King's College London, Prof Pablo Lamata, Dr Marc Modat and Mr Loshan Kangesu, for their useful feedback during my thesis progress review meetings and for helping me to better understand the clinical motivation for this work. Third, I wish to thank my two excellent supervisors, Dr Andrew King and Dr Marc Miquel, for generously and effectively sharing their expertise with me, and for their support, patience and kindness during a PhD journey that was more complicated than anticipated due to factors such as the COVID-19 pandemic. Finally, I wish to thank my family, especially my partner, Sandra, for their extensive support, patience and love, in particular during the thesis writing-up period which took me more time to complete than anticipated. This work would not have been possible without you.

Matthieu Ruthven, July 2023

Contents

List of Figures

List of Tables

Symbols and Abbreviations

2D	Two-dimensional
3D	Three-dimensional
ASD	Average symmetric surface distance
BD	Boundary distance weighting
CE	Cross entropy
CF	Class frequency weighting
CNN	Convolutional neural network
D	Displacement field
DL	Deep-learning
DSC	Dice coefficient
FCN	Fully convolutional network
FFD	Free-form deformation
GE	Gradient echo
GT	Ground-truth
HD	General Hausdorff distance
JRS	Joint registration and segmentation
LVP	*Levator veli palatini*
MI	Mutual information
mm	Millimetres
MR	Magnetic resonance
MRI	Magnetic resonance imaging
ms	Milliseconds
MSE	Mean squared error
PD	Proton-density
PDw	Proton-density-weighted
ReLU	Rectified linear unit
RF	Radiofrequency
rtMR	Real-time magnetic resonance
rtMRI	Real-time magnetic resonance imaging

s	Second
SDM	Statistical deformation model
SE	Spin echo
SIFFD	Segmentation-informed free-form deformation
SIVXM	Segmentation-informed VoxelMorph
t	Time
T	Tesla
T_1	Spin-lattice relaxation time
T_1w	T_1-weighted
T_2	Spin-spin relaxation time
T_2w	T_2-weighted
TE	Echo time
TR	Repetition time
TSE	Turbo spin echo
VPI	Velopharyngeal insufficiency
VXM	VoxelMorph

Chapter 1: Introduction

1.1 Motivation

Speech is one of the principal forms of human communication. Its production is a complex process involving several body parts including articulators such as the tongue and soft palate. A phenomenon called velopharyngeal closure regularly occurs during normal speech. Velopharyngeal closure prevents airflow into the nasal cavity and is required to produce most speech sounds. For velopharyngeal closure to occur, the soft palate must elevate and come into contact with the pharyngeal walls. The muscle primarily responsible for elevating the soft palate is called the *levator veli palatini* (LVP).

Velopharyngeal insufficiency (VPI) is an anatomical or structural defect that prevents velopharyngeal closure and consequently impairs speech [1]. Speech impairments negatively affect social and educational development as well as quality of life [2]. Mitigating such impairments is therefore crucial in order to avoid these negative effects.

Since VPI is an anatomical or structural defect, it can only be treated via surgery [1]. Several well-established surgical techniques to treat VPI exist, however, there is no consensus on which is most effective and consequently a variety of techniques are used [3–5]. The technique most likely to improve the speech of a patient depends on the defect(s) preventing velopharyngeal closure. If the defect is a poorly functioning LVP, a treatment that aims to improve LVP function is performed, while if the defect is an abnormally deep pharynx or an insufficiently long soft palate, a treatment that aims to address these defects is performed. However, treatment is not always successful and consequently further surgery can be required: studies have reported persistence of VPI requiring further surgery in 0-50% of cases [5,6]. Key drawbacks of further surgery are additional distress and disruption for patients and their carers, increased workloads for clinicians and additional costs for healthcare services.

Typically in clinical assessments of speech, imaging is used to enable identification of the defects preventing velopharyngeal closure. Since the treatment most likely to improve the speech of a patient depends on the defect(s) preventing velopharyngeal closure, imaging has an important role in the management of patients with VPI by providing clinicians with information that aids treatment decision making. Imaging enables visualisation of the vocal tract and soft palate during speech and therefore provides information about the size, shape and motion of these anatomical features. Visualisation of vocal tract and soft palate size and shape enables clinicians to identify if the defect is an abnormally deep pharynx or an insufficiently long soft palate, while visualisation of soft palate motion enables clinical teams to infer if the LVP is functioning adequately. Currently, the imaging techniques most commonly used in clinical speech assessment are videofluoroscopy and nasendoscopy [3,4]. These imaging techniques enable two-dimensional (2D) visualisation of the vocal tract and soft palate during speech. However, neither technique enables visualisation of the LVP.

There is increasing interest in using magnetic resonance imaging (MRI) in clinical speech assessment, due to the unique ability of MRI to noninvasively and dynamically image the vocal tract and articulators during speech and acquire detailed three-dimensional (3D) images of the LVP without using ionising radiation [7,8]. Three-dimensional visualisation of the LVP would provide clinicians with additional clinically relevant information that could aid VPI treatment decision making.

A range of MRI techniques to dynamically image the vocal tract and articulators during speech have been developed [9–11]. While techniques that enable imaging at very high spatio-temporal resolutions have been developed, these require specialised MRI equipment and software [9,10] and are therefore very challenging to implement in other centres. This barrier to adoption has motivated the development of techniques that only require standard MRI equipment and software [11]. While these techniques image at lower spatio-temporal resolutions, the resolutions are nevertheless sufficient to capture the general motion of articulators such as the soft palate [12].

Due to the small size of the LVP and its 3D structure, 3D MRI at a high spatial resolution is required to fully visualise the muscle. The LVP and the soft tissue that surrounds it have very similar tissue properties. Consequently, a key challenge when imaging the LVP is ensuring that the image contrast is sufficient to discriminate between the two. In previous work, T_2-weighted images of the LVP have predominantly been acquired [8]. However, the results of recent work suggest that the image contrast in T_1-weighted or proton-density-weighted (PD-weighted) images may result in improved LVP visualisation [13]. There is therefore currently no consensus on the optimal image contrast for LVP visualisation.

Currently in clinical speech assessment, analysis of vocal tract and soft palate size, shape and motion is qualitative and no analysis of the size, shape and configuration of the LVP is performed. However, in combination with the increasing interest in using MRI in clinical speech assessment, there is increasing interest in quantitative analysis of vocal tract and soft palate size, shape and motion, and also LVP size, shape and configuration in magnetic resonance (MR) images [7,8]. Such quantitative analysis would provide objective information to aid treatment decision-making. In previous work, measurement of the size, shape and configuration of the soft palate and LVP was manually performed [7,8]. Such measurements are time consuming, require input from specialists, are prone to intra- and inter-observer variability and are consequently not feasible on a large scale. There is therefore a growing unmet need for methods to automatically perform these measurements. This unmet need is not limited to the clinical speech assessment community; the speech science community is increasingly using MRI to visualise the vocal tract and articulators during speech and is also increasingly interested in methods to automatically measure the vocal tract and articulators in MR images.

Segmentation of medical images is a common first step to enable automatic measurement of anatomical structures. Several methods have been developed to segment air-tissue boundaries between the vocal tract and adjacent articulators in MR images of speech [14–16], however, these methods do not fully segment articulators such as the soft palate and therefore do not enable analysis of articulator size, shape and motion. Instead, methods that fully segment articulators in MR images of speech are required. Regarding methods to automatically segment the LVP in 3D MR images, there is only one report in the literature of such methods: in very recent work (postdating the work described in this thesis) deep-learning-based (DL-based) methods to perform this task were developed and compared [17].

Development and evaluation of segmentation methods requires datasets with corresponding ground-truth (GT) segmentations. While there are publicly available speech MRI datasets [18,19], none of these include GT segmentations of articulators. In addition, there are no publicly available MRI datasets in which the LVP can be adequately visualised. Due to the lack of suitable publicly available MRI datasets, acquisition of new datasets and creation of GT segmentations is required to enable the development and evaluation of methods to segment the vocal tract, soft palate and LVP in MR images.

An established way to automatically quantify complex motion in an image series is by using a nonlinear registration method to estimate displacement fields between the images. While traditional registration methods have been used to register MR images of speech in several previous works [20,21], there is only a single report in the literature of these methods being used to quantify articulator motion: in [20], such methods were used to quantify tongue motion. However, there are no reports in the literature of such methods being used to quantify soft palate motion.

To be suitable for use in clinical speech assessment, a key requirement of image analysis methods is that they capture any velopharyngeal closures that occur. However, standard metrics for evaluating segmentation and motion quantification method accuracy do not provide such information. To enable clinically relevant evaluation of the accuracy of image analysis methods, there is an unmet need for the development of such metrics.

The main aim of the work presented in this thesis is to begin to address the unmet need for methods to perform automatic quantitative analysis of the vocal tract, soft palate and LVP in MR images, by developing methods to segment such images and developing a framework to quantify motion in such images.

1.2 Contributions

The work presented in this thesis makes several contributions towards addressing the unmet need for methods to perform automatic quantitative analysis of the vocal tract, soft palate and LVP in MR images. More specifically, as part of the work two segmentation methods and a motion quantification framework were developed, GT segmentations were created for an existing speech MRI dataset, a new MRI dataset including GT segmentations was created and a novel metric based on velopharyngeal closure was developed to enable a more clinically relevant evaluation of segmentation method and motion quantification framework accuracy. More information about these contributions is provided in the following sections.

1.2.1 Articulator Segmentation in MR Images of Speech

As a first step towards enabling automatic measurement of vocal tract and soft palate size, shape and motion in 2D MR images of speech, a method to automatically segment the vocal tract, soft palate and four other anatomical features in this type of image was developed.

The DL-based method includes an extension to automatically calculate the minimum distance between the soft palate and the posterior pharyngeal wall, a measurement of particular interest to clinicians who perform clinical speech assessments. Although primarily designed to enable automatic measurement of vocal tract and soft palate size, shape and motion in 2D MR images of speech, the 2D segmentation method was designed to also enable measurement of tongue size, shape and motion in order to broaden its potential applications and utility.

1.2.2 Quantification of Articulator Motion in MR Images of Speech

As an additional step towards enabling automatic measurement of soft palate motion in 2D MR images of speech, a framework to automatically estimate the motion of the soft palate and five other anatomical features in this type of image was developed. This deep learning framework for nonlinear registration of 2D MR images of speech builds on the 2D image segmentation method by incorporating knowledge of region boundaries into the registration and automatically estimates displacement fields between this type of image.

1.2.3 LVP Segmentation in 3D MR Images

As a first step towards enabling automatic measurement of LVP size, shape and configuration in 3D MR images of the vocal tract, a method to automatically segment the LVP and pharynx in this type of image was developed. Similarly to the 2D image segmentation method and the motion quantification framework, the 3D image segmentation method is deep learning based.

1.2.4 Speech MRI Dataset GT Segmentation Creation

There are currently no publicly available speech MRI datasets that include GT segmentations of the entire vocal tract or soft palate. To enable the development and evaluation of a method to segment such images, GT segmentations for an existing speech MRI dataset were created. A dataset acquired using a speech MRI technique that does not require specialised MRI equipment and software was deliberately chosen in order to facilitate acquisition of similar images in other centres and consequently application of the 2D image segmentation method presented in this thesis. GT segmentations of the vocal tract, soft palate and four other anatomical structures were manually created for each image in the dataset.

1.2.5 New MRI Dataset and GT Segmentation Creation

There is currently no consensus on the optimal image contrast for visualising the LVP in 3D MR images. In addition, there are currently no publicly available MRI datasets that include GT segmentations of the LVP. To enable the development and evaluation of a method to segment the LVP in 3D MR images, a new dataset of 3D MR images of the vocal tract was acquired after performing an investigation to identify the parameters that result in the optimal image contrast for visualising the LVP in this type of image. GT segmentations of the LVP and pharynx were manually created for each image in the dataset.

1.2.6 Novel Metric for Clinically Relevant Method Accuracy Evaluation

To be suitable for use in clinical speech assessment, a key requirement of segmentation and motion quantification methods is that they capture any velopharyngeal closures that occur. However, standard metrics for evaluating segmentation and motion quantification method accuracy do not provide such information. To enable more clinically relevant evaluation of the accuracy of segmentation methods and motion quantification frameworks, a novel metric based on velopharyngeal closure was developed.

1.3 Outline

This thesis is divided into eight chapters, including this Introduction. The other seven chapters of this thesis are outlined below:

Chapter 2: Clinical Background

Chapter 2 provides the clinical background to the work presented in this thesis. As this work involves developing methods to analyse images of speech, the chapter provides an overview of speech production. The chapter also provides an overview of VPI, the health problem that this work ultimately aims to address, and the management of patients with VPI.

Chapter 3: Technical Introduction

Chapter 3 introduces the technical background to this work. First, as MRI data was acquired and used in this work, the chapter provides an overview of MRI and then reviews speech MRI techniques and MRI techniques for LVP visualisation. Second, as DL-based methods were developed in this work, the chapter provides an overview of deep learning and its application to medical image analysis. Third, as DL-based segmentation methods were developed in this work, the chapter provides an overview of medical image segmentation focusing on DL-based methods and then reviews the literature on the segmentation of speech MR images and the segmentation of the LVP in MR images. Fourth, as a motion quantification framework based on image registration was developed in this work, the chapter provides an overview of medical image registration and then reviews the literature on the registration of speech MR images.

Chapter 4: Materials

Chapter 4 describes the datasets acquired and used in this work. First, the chapter describes the previously acquired speech MRI dataset that was used in this work and how GT segmentations were created for this dataset. Second, the chapter describes the new MRI dataset that was acquired in this work, including the image contrast optimisation investigation that was performed prior to acquiring the dataset. The chapter then describes how GT segmentations were created for this new dataset. The speech MRI dataset and corresponding GT segmentations were used to develop the segmentation method presented in chapter 5 and the registration framework presented in chapter 6, while the new MRI dataset and corresponding GT segmentations were used to develop the segmentation method presented in chapter 7.

Chapter 5: DL-Based Segmentation of Speech MRI Data

Chapter 5 presents a DL-based method to segment the vocal tract and articulators in 2D MR images of speech. The chapter also presents an extension to the method to calculate the minimum distance between the soft palate and the posterior pharyngeal wall. Finally, the chapter presents a novel clinically relevant metric based on velopharyngeal closure to evaluate the accuracy of segmentations estimated by the method. The speech MRI dataset described in chapter 4 was used to develop the segmentation method.

Chapter 6: DL-Based Nonlinear Registration of Speech MRI Data

Chapter 6 presents a deep learning framework for nonlinear registration of 2D MR images of speech. The framework builds on the segmentation method presented in chapter 5. Chapter 6 also presents the results of experiments comparing the performance of the proposed framework to state-of-the-art traditional nonlinear registration methods and deep learning frameworks for nonlinear registration. One of the metrics used in this comparison was the novel clinically relevant metric based on velopharyngeal closure presented in chapter 5. The speech MRI dataset and corresponding GT segmentations described in chapter 4 were used to develop the proposed framework and in the performance comparison experiments.

Chapter 7: DL-Based LVP Segmentation in 3D MR Images

Chapter 7 presents a DL-based method to segment the LVP and pharynx in 3D MR images of the vocal tract. It also presents the results of experiments investigating the effect of different data augmentation methods on the accuracy of the segmentation method. The new MRI dataset and corresponding GT segmentations described in chapter 4 were used to develop the segmentation method and in the data augmentation experiments.

Chapter 8: Conclusions

Chapter 8 first summarises the contributions of the work presented in this thesis. The chapter then discusses limitations of this work and makes suggestions on future work.

Chapter 2: Clinical Background

This chapter introduces the clinical background to the work presented in this thesis. It consists of an overview of speech and its production, followed by an overview of VPI, the health problem that this work ultimately aims to tackle.

2.1 Speech

Speech is one of the principal forms of human communication. Its production is a complex process involving several body parts (see Figure 1), notably the lungs, vocal folds (also known as the vocal cords) and articulators including the lips, tongue and soft palate (also known as the velum) [1,22].

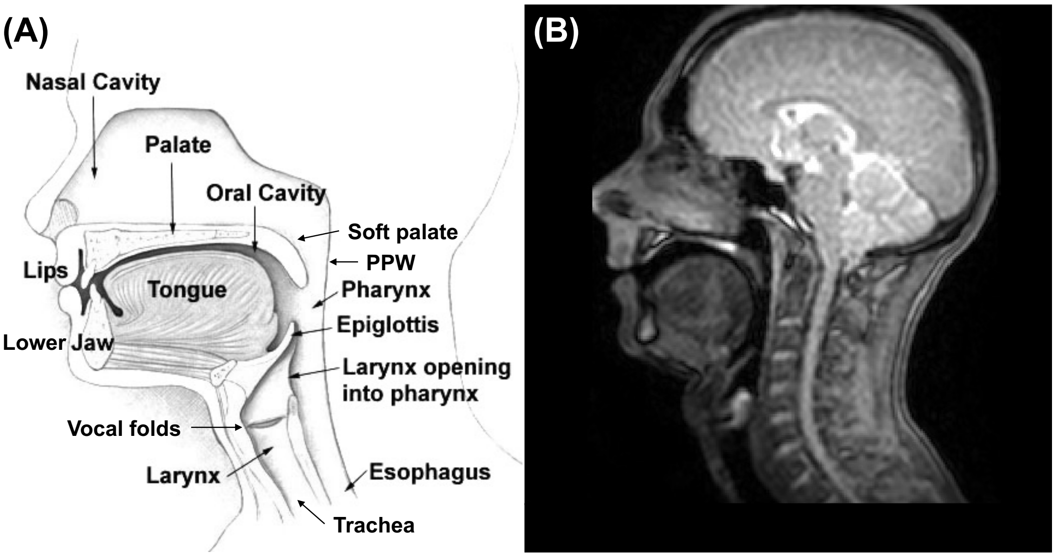

Figure 1: (A) A diagram of a midsagittal slice of the head, showing anatomical features with key roles in speech production (modified from [23]*). (B) A real-time magnetic resonance image of a midsagittal slice of the head. PPW: posterior pharyngeal wall.*

Speech production requires a flow of air from the lungs. For the majority of speech sounds, the airflow passes through the trachea, larynx (which contains the vocal folds), pharynx and oral cavity, and leaves the body via the mouth. For a few speech sounds, such as [m], [n], [ng] in English, the airflow passes through the trachea, larynx, pharynx, oral cavity and nasal cavity, and leaves the body via the nose.

Airflow past the vocal folds causes them to vibrate. This vibration modulates the airflow and generates sound. The tension of the vocal folds and their separation, both of which can be controlled by the speaker, determine the frequencies of the sound that is generated.

The oral cavity acts as a resonator that modifies the sound generated by the vocal folds to the desired speech sound. The sound modification depends on the size and shape of the oral cavity. The speaker can control these properties of the cavity by moving articulators including the lips, tongue and soft palate to different positions.

As well as modifying the shape of the oral cavity, the soft palate is responsible for preventing airflow into the nasal cavity. Prevention of such airflow is required to produce all speech sounds in English apart from [m], [n] and [ŋ]. The soft palate prevents such airflow by blocking the opening between the pharynx and the nasal cavity. It achieves this by elevating and coming into contact with the pharyngeal walls (see Figure 2). Blockage of the opening in this way is known as velopharyngeal closure.

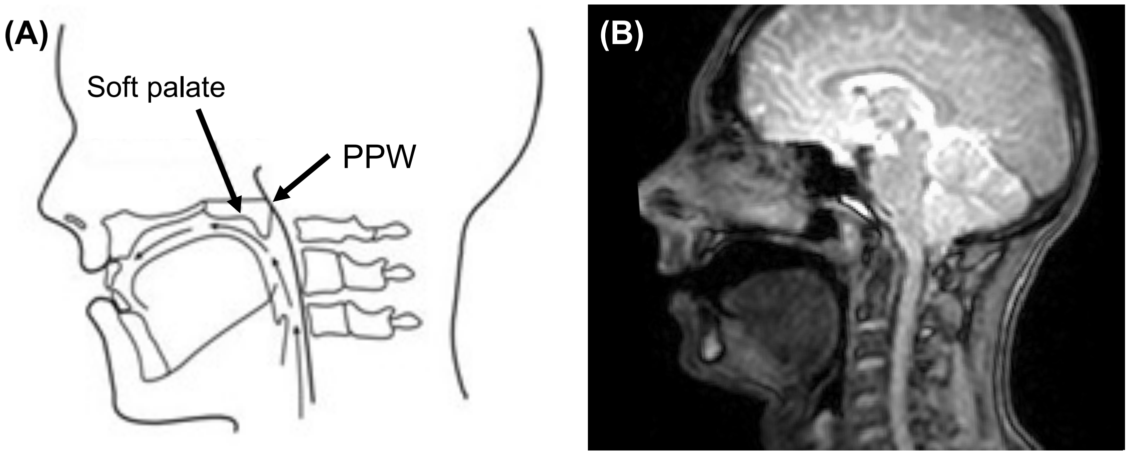

Figure 2: (A) A diagram of a midsagittal slice of the head showing velopharyngeal closure: the soft palate is elevated and in contact with the posterior pharyngeal wall (PPW) (modified from [24]*). (B) A real-time magnetic resonance image of a midsagittal slice of the head showing velopharyngeal closure.*

Elevation of the soft palate is primarily caused by a muscle called the LVP. The LVP forms a U-shaped sling that lifts the soft palate (see Figure 3). The muscle originates from the base of the skull, close to the petrous part of the temporal bone, and connects to the midsection of the soft palate at approximately 40% of the length of the soft palate [25,26] (see Figure 4).

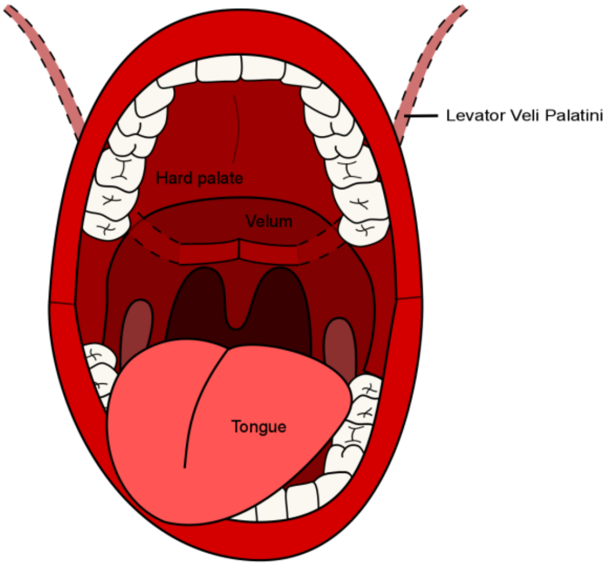

Figure 3: A diagram of an open mouth, showing the tongue, velum (soft palate), hard palate and levator veli palatini (modified from [27]*).*

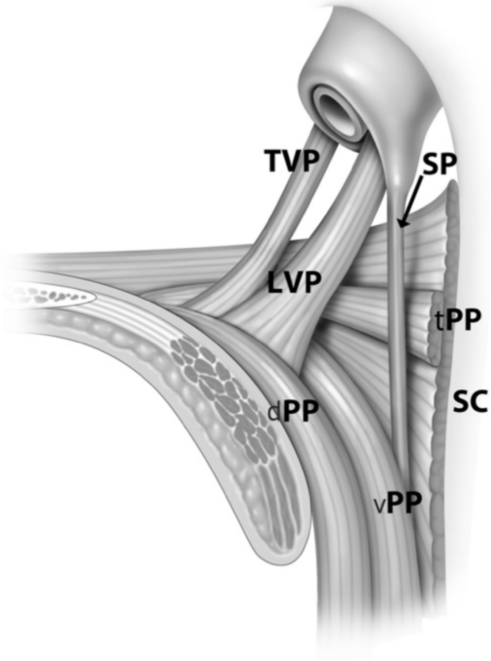

Figure 4: A diagram of a midsagittal view of the soft palate and its muscles [28]*: the levator veli palatini (LVP), tensor veli palatini (TVP), salphingopharyngeus (SP), superior pharyngeal constrictor (SC), and the transverse fascicle (tPP), dorsal fascicle (dPP) and ventral fascicle (vPP) of the palatopharyngeus muscle (PP).*

2.2 Velopharyngeal Insufficiency

2.2.1 Causes and Effects

VPI is an anatomical or structural defect that prevents velopharyngeal closure [1]. Examples of defects include the pharynx being abnormally deep, the LVP not elevating the soft palate sufficiently to block the opening between the pharynx and the nasal cavity, and the soft palate not being sufficiently large to block the opening between the pharynx and the nasal cavity [29]. As a result of the defect, airflow during speech is disrupted as air flows into the nasal cavity when it should not. This disruption can make it challenging or impossible to produce certain speech sounds and therefore impairs speech. The extent to which speech is impaired by VPI is variable. In cases where the defect is minor, individuals can produce most speech sounds correctly. However, in more severe cases, individuals can only produce a few speech sounds correctly. Speech and language impairments have been found to negatively affect social and educational development [2,30–33]. VPI is the health problem which the work described in this thesis ultimately aims to address.

2.2.2 Prevalence

Two populations of individuals are particularly prone to VPI: individuals with a repaired cleft palate and individuals with velocardiofacial syndrome [34–38]. The incidence of VPI has been found to be 16-37% in individuals with a repaired cleft palate [34,35] and 27-92% in individuals with velocardiofacial syndrome [38].

Orofacial clefts are abnormal fissures in the lip and/or palate that are present from birth (see Figure 5). They occur when different sections of the lip and/or palate do not fuse together correctly during prenatal development. A cleft can be in the lip only (cleft lip), the palate only (cleft palate) or both (cleft lip and palate). Cleft lips can be further categorised as being either unilateral or bilateral depending on whether they are on both sides of the face (see Figure 5). A cleft in the palate results in an abnormal opening between the oral and nasal cavities that negatively affects feeding and speech.

In the United Kingdom (UK), approximately 800 babies per year are born with a orofacial cleft that involves the palate [39,40]. Cleft palates are surgically repaired, usually six to 12 months after birth. Since children with a repaired cleft palate are known to be prone to speech impairments, their speech is assessed by Speech and Language Therapists (SLTs) every two years until the age of 18.

Velocardiofacial syndrome is a genetic condition caused by a hemizygous deletion of chromosome 22q11.2 [38]. It is characterised by heart anomalies and mild-to-moderate immune deficiencies. Additional common characteristics of individuals with the syndrome include facial dysmorphia, developmental delay and VPI. The prevalence of velocardiofacial syndrome in the UK has been found to be approximately 1 per 4000 births [46].

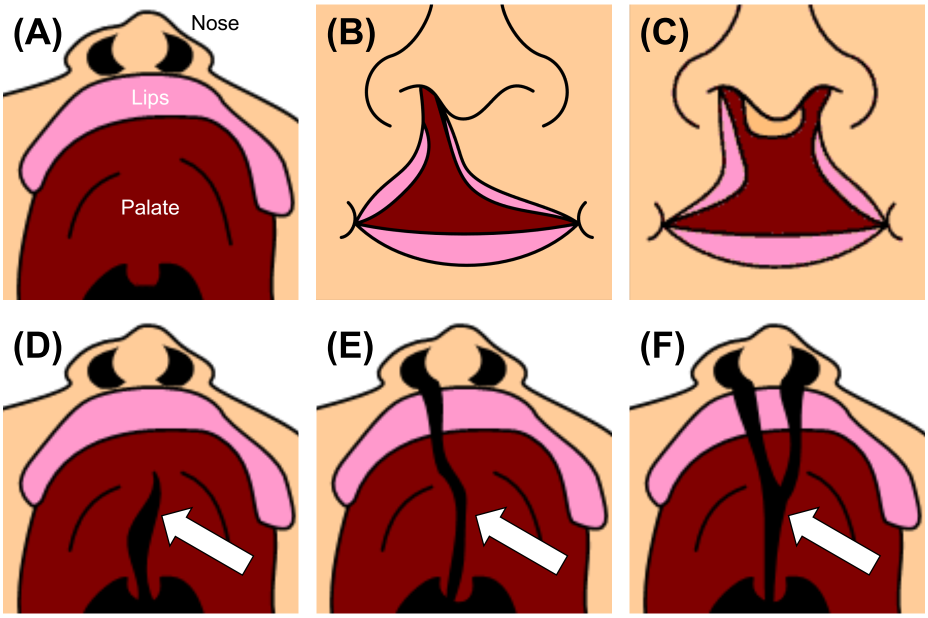

Figure 5: Different types of orofacial clefts: (A) normal palate (modified from [41]*), (B) unilateral cleft lip* [42]*, (C) bilateral cleft lip* [43]*, (D) cleft palate* [41]*, (E) unilateral cleft lip and palate* [44]*, (F) bilateral cleft lip and palate* [45]*. White arrows indicate clefts.*

2.2.3 Treatment

Since VPI is an anatomical or structural defect, it can only be treated via surgery [1]. As part of their rehabilitation following surgery, patients receive speech therapy to help them learn to use the modified anatomy effectively and to eliminate compensatory placements (i.e. abnormal articulator positioning during speech production in order to compensate for the anatomical or structural defect). Several well-established surgical techniques to treat VPI exist, including intravelar veloplasty, palate re-repair, pharyngeal flap, sphincter pharyngoplasty and Furlow Z-palatoplasty [47]. Intravelar veloplasty and palate re-repair both aim to change the position and orientation of the LVP to improve the function of the muscle. Pharyngeal flap and sphincter pharyngoplasty both aim to reduce the size of the opening between the pharynx and nasal cavity. Furlow Z-palatoplasty aims to both change the position and orientation of the LVP and reduce the size of the opening between the pharynx and nasal cavity. However, there is no consensus on which technique is most effective and consequently a variety of techniques are used [4–6,47]. The technique most likely to improve the speech of a patient depends on the defect(s) preventing velopharyngeal closure, however, the choice of technique can also be influenced by surgeon experience [47]. If the defect is a poorly functioning LVP, a technique that aims to improve LVP function such as a palate re-repair is performed, while if the defect is an abnormally deep pharynx or an insufficiently large soft palate, a technique that aims to reduce the size of the opening between the pharynx and nasal cavity is performed. Surgical treatment is most often performed when the patient is approximately six years old [48–52]. However, treatment is not always successful: studies have reported persistence of VPI following surgery in 16-100% of cases [48–52]. VPI persistence can necessitate further surgery [48,49,52]. For patients and their carers, further surgery results in additional hospital visits. These visits can be distressing and inconvenient, and usually cause patients to miss school. In addition, accompanying carers must usually take time off work for the visits and post-surgery care at home. For health and care services, further surgery results in additional workloads for clinical teams and additional costs. In the UK, the cost of a surgery and its planning and follow-up is approximately £8500. Avoiding further surgery would therefore avoid large additional costs. The ultimate goal of the work presented in this thesis is to develop methods to help clinical teams improve the treatment outcomes of patients with VPI and therefore reduce the rates of further surgery.

Clinical assessments of speech are performed to identify the defect preventing velopharyngeal closure and thus inform treatment decisions. These assessments are performed by SLTs and Plastic Surgeons. Clinical speech assessments usually involve imaging to enable clinical teams to visualise the pharynx and soft palate of a patient while (s)he is speaking [3,4,53]. Visualisation of the shape of the pharynx and soft palate enables clinical teams to identify if the defect preventing velopharyngeal closure is an abnormally deep pharynx or an insufficiently long soft palate. Visualisation of the motion of the soft palate during speech enables clinical teams to infer how well the LVP is functioning and whether it is connected to the soft palate at an abnormal location. Identification of the defect preventing velopharyngeal closure in turn aids clinical teams to decide on the treatment required to correct the defect. Imaging therefore has an important role in the management of patients with VPI, by providing key information for treatment decision-making. In the UK, the imaging techniques most commonly used in clinical speech assessments are videofluoroscopy and nasendoscopy [4,53]. Videofluoroscopy is a technique that uses X-rays to visualise the inside of the body. Nasendoscopy is when a small camera is threaded into the nasal cavity via the nose, enabling visualisation of the top of the soft palate. Both these imaging techniques enable 2D visualisation of the pharynx and soft palate, however, neither enables visualisation of the LVP.

2.3 Conclusions

Articulators such as the soft palate play an essential role in the production of speech. In combination with the LVP, the soft palate causes velopharyngeal closure, a key requirement for the production of most speech sounds. VPI is an anatomical or structural defect that prevents velopharyngeal closure and consequently impairs speech. While several well-established surgical techniques to treat VPI exist, there is currently no consensus on which is most effective and consequently a variety of techniques are used. The technique most likely to improve the speech of a patient depends on the defect(s) preventing velopharyngeal closure. Imaging is used in clinical speech assessments to aid identification of the defect(s) and therefore inform treatment decisions. However, treatment is not always successful and further surgery can be required, causing additional distress and disruption for patients and their carers, additional workloads for clinical teams and additional costs for health and care services. The ultimate goal of the work presented in this thesis is to develop methods to help to improve the treatment outcomes of patients with VPI.

Chapter 3: Technical Introduction

This chapter introduces the technical background to the work presented in this thesis. The first section provides an overview of MRI and its use for visualising the articulators and LVP. The second section provides an overview of deep learning and its use for medical image analysis tasks. The third section provides an overview of DL-based image segmentation methods, followed by a review of existing methods for segmenting articulators and the LVP in MR images. The final section provides an overview of medical image registration methods, followed by a summary of previous work in which these methods were applied to MR images of speech.

3.1 Magnetic Resonance Imaging

3.1.1 Introduction

MRI is a non-invasive imaging technique primarily used to acquire images of the inside of the body. It is widely used in clinical practice and has an important role in the diagnosis and monitoring of a wide range of diseases including cancer and dementia. In addition, MRI is widely used in multiple research areas and is itself a topic of much research and development. MRI is primarily known for its ability to acquire detailed 2D or 3D images of static parts of the body. However, due to advances in MRI technology and data acquisition acceleration strategies, MRI can now be used to acquire images of dynamic processes such as speech production.

Providing a detailed coverage of all relevant aspects of MRI is beyond the scope of this section. Instead, brief introductions to the fundamentals of key aspects are provided in the following sections. For further details, readers are referred to [54] for an introduction to the components of an MRI scanner, [55] for an introduction to MR signal creation and relaxation, [56] for an introduction to MR image acquisition and k-space, and [57] and [58] for introductions to pulse sequences.

3.1.2 Pulse Sequences

In MRI, a sequence of radiofrequency (RF) pulses and magnetic field gradients are applied to produce the signals required for image formation. This sequence is known as a pulse sequence. There are many different types of pulse sequence, the most basic of which are spin echo (SE) and gradient echo (GE) [59]. In practice, variants of SE and GE pulse sequences that enable faster image acquisition are primarily used. Particularly widely used sequences include turbo SE (TSE) sequences and fast GE sequences [60].

While an SE sequence produces a single MR signal per RF excitation pulse, a TSE sequence produces multiple signals, thus accelerating image acquisition. TSE sequences produce multiple signals by applying additional RF pulses and magnetic field gradients between the RF excitation pulses [57]. The number of signals that are produced per RF excitation pulse is known as the echo train length or the turbo factor. TSE sequences are primarily used to acquire detailed 2D or 3D images of static parts of the body. A key advantage of TSE sequences is that they can acquire images with a wide range of different contrasts. However, while TSE sequences accelerate image acquisition, the acceleration is not usually sufficient to enable dynamic imaging.

Fast GE sequences sufficiently accelerate image acquisition to enable dynamic imaging. The most commonly used types of fast GE sequences are spoiled GE sequences and refocused GE sequences [58]. Spoiled GE sequences are almost identical to GE sequences, except that an additional magnetic field gradient, known as a spoiler gradient, is applied after signal acquisition in order to remove any remaining transverse magnetisation and therefore prevent it from affecting the production of subsequent signals. In refocused GE sequences, additional magnetic field gradients are applied to manipulate the residual transverse magnetisation so that it contributes to the production of subsequent signals. While fast GE sequences enable more rapid image acquisition than other types of sequences such as TSE sequences, the range of image contrasts in images acquired using such sequences is more limited.

Pulse sequences have a range of parameters that can be modified to affect the image acquisition speed, the spatial resolution of imaging and the contrast of the images that are acquired. The key parameter that affects image acquisition speed is the repetition time (TR).

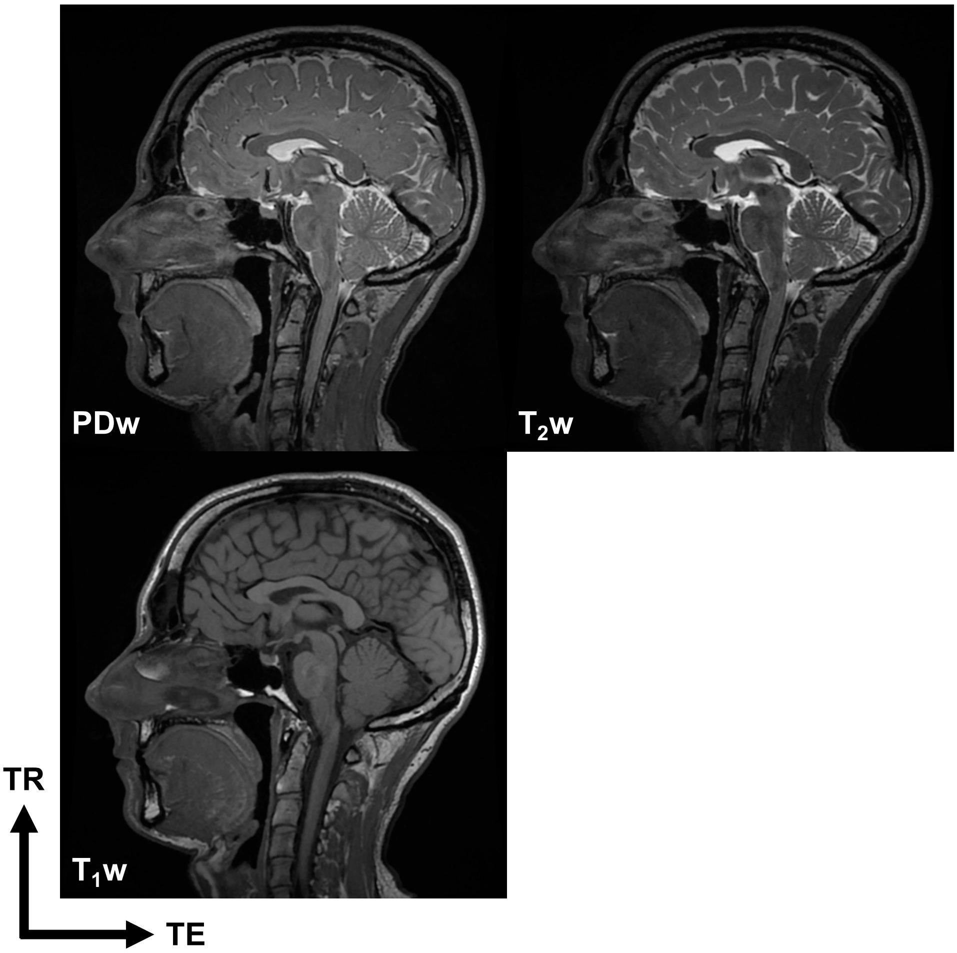

Figure 6 Magnetic resonance images with different contrasts: T_1-weighted (T_1w), proton-density-weighted (PDw) and T_2-weighted (T_2w). The x- and y-axes are repetition time (TR) and echo time (TE) respectively.

3.1.3 Relaxation and Image Contrast

In MRI, after the application of an RF excitation pulse, the recovery of the longitudinal magnetisation (M_z) is characterised by the spin-lattice relaxation time (T_1) and is commonly modelled using the following equation:

$$M_z \propto 1 - e^{-t/T_1} \tag{1}$$

where t is time. The decay of the transverse magnetisation (M_{xy}) is characterised by the spin-spin relaxation time (T_2) and is commonly modelled using the following equation:

$$M_{xy} \propto e^{-t/T_2} \tag{2}$$

T_1 and T_2 are substance dependent. For example, in the body at a magnetic field of 3.0 T, the T_1 of fat and muscle is approximately 400 ms and 900 ms respectively, while the T_2 of fat and muscle is approximately 70 ms and 30 ms respectively [55]. In MRI, these differences in relaxation times are exploited in order to acquire images with different contrasts.

A key advantage of MRI over other imaging techniques is its ability to acquire images with a range of different contrasts. Several factors affect MR image contrast including the proton density and the relaxation times of the volume being imaged, the strength of the main magnetic field, the type of pulse sequence used in image acquisition and the parameters of the pulse sequence. It is common to describe an MR image as T_1-, T_2- or proton-density-weighted (PD-weighted), depending on the factor that most influenced the image contrast. Examples of images with different contrasts are shown in Figure 6. The contrast in a T_1-weighted image depends primarily on the differences in the amplitudes of the longitudinal magnetisations in different regions of the volume being imaged, while the contrast in a T_2-weighted image depends primarily on the differences in the amplitudes of the transverse magnetisations. The contrast in a PD-weighted image depends primarily on the proton density of the volume being imaged.

The parameters of pulse sequences can be modified in order to acquire images with different contrasts. The key parameters that affect the image contrast are the TR, echo time (TE) and, for GE-based sequences, the flip angle. Generally, to acquire a T_1-weighted image, a pulse sequence with a relatively short TR and TE is required. Conversely, to acquire a T_2-weighted image, a pulse sequence with a relatively long TR and TE is required. To acquire a PD-weighted image, a pulse sequence with a relatively long TR and a relatively short TE is required.

A suitable image contrast is required to be able to distinguish between different regions in an image and ultimately visualise anatomical features in medical images. In MRI, the process to identify the pulse sequence parameters that result in an optimal contrast for anatomical feature visualisation is known as pulse sequence optimisation.

3.1.4 Tradeoffs in MRI

MRI involves an unavoidable tradeoff between the image acquisition speed, image quality and spatial resolution. The optimal tradeoff for a given application depends on the relative importance of these three factors. MRI of dynamic processes such as speech production requires fast image acquisition to ensure that the temporal resolution of imaging is sufficiently high to capture the processes as they occur. Nevertheless, visualisation of the processes also requires adequate image quality and a sufficiently high spatial resolution.

Commonly used strategies for accelerating MR image acquisition include using faster pulse sequences, parallel imaging, non-Cartesian k-space sampling, novel image reconstruction methods and custom receive coils. While some strategies such as faster pulse sequences and parallel imaging are widely available on standard MRI scanners, others such as non-Cartesian k-space sampling, novel reconstruction methods and bespoke receive coils are only available on specialised MRI scanners. Generally, the former type of scanner is much more common in clinical practice than the latter.

3.1.5 Dynamic MRI Techniques

MRI is primarily known for its ability to acquire detailed 2D or 3D images of static parts of the body. However, due to advances in MRI technology and data acquisition acceleration strategies, MRI can now be used to acquire images of dynamic processes such as speech production. Dynamic MRI techniques use a variety of data acquisition acceleration strategies, usually in combination, to enable imaging at high temporal resolutions while maintaining adequate image quality and spatial resolution [12,27,61–63]. Applications of dynamic MRI techniques include in cardiac MRI [61–63], MRI-guided invasive procedures [61] and speech MRI [12,27,61].

Dynamic processes that regularly repeat in a similar manner, such as the beating of the heart, can be dynamically imaged at high spatio-temporal resolutions using triggered and gated MRI techniques [62,63]. However, these types of technique require monitoring of the dynamic process. For example, the beating of the heart is monitored using electrocardiography [62,63]. Using this monitoring, triggered MRI techniques synchronise data acquisition so that it only occurs at specific stages of the process, while gated MRI techniques continuously acquire data and then retrospectively use the recorded monitoring signal to determine at which stage of the process data were acquired. To acquire all the data required to create an image, triggered and gated MRI techniques require several repetitions of the dynamic process. Consequently, these techniques can require up to several minutes to acquire all the data required to create an image.

Real-time MRI (rtMRI) techniques enable imaging of dynamic processes as they occur, without requiring any repetition of the processes. This type of technique is therefore, unlike triggered and gated MRI techniques, not restricted to imaging dynamic processes that regularly repeat in a similar manner. However, achieving the desired spatio-temporal resolutions is more challenging due to the lack of repetition.

3.1.6 Vocal Tract and Articulator Visualisation during Speech

Visualisation of the vocal tract and articulators during speech provides information about the size, shape, motion and position of these anatomical features during speech production. In a research context, primarily in speech science research, this information is desirable as it provides insights into speech production, while, as described in section 2.2.3, in clinical practice this information is desirable as it enables identification of the causes of speech problems and consequently informs decisions on how to treat the problems [1,3,4].

Due to their location in the body, imaging is required to visualise the vocal tract and articulators during speech. Several different imaging techniques enable visualisation of these anatomical features. The most commonly used techniques are nasendoscopy [1,3,4], videofluoroscopy [4], ultrasound (US) [64–67] and MRI [12,27,61]. Each of these techniques has its advantages and disadvantages. Nasendoscopy is free from ionising radiation and requires relatively inexpensive technology but is minimally invasive, potentially affecting speech, and visualisation is limited to external surfaces of articulators. Videofluoroscopy is non-invasive and quick to perform. However, it involves exposure to ionising radiation, specialist staff and facilities are required to perform it, and visualisation is limited to projections of the anatomy. US imaging is non-invasive, free from ionising radiation and requires relatively inexpensive technology, but visualisation is limited to the tongue. MRI is non-invasive, free from ionising radiation and enables visualisation of any view of the vocal tract and articulators. However, it requires expensive equipment and specialist staff and facilities to perform.

3.1.7 Dynamic MRI of Speech

Use of MRI to visualise the vocal tract and articulators during speech is increasing due to the growing availability of MRI scanners, the development of dynamic MRI techniques for such visualisation, and the unique ability of MRI to non-invasively acquire images of any orientation without using ionising radiation [12,27,61]. Currently, the main application of dynamic MRI of speech is in speech science research [68–76]. However, there is increasing interest in using dynamic MRI in the clinical assessment of speech of patients with VPI [7,77–82], apraxia [83], stutter [84] or sleep apnea [85,86], or patients following glossectomy [87,88]. Dynamic MRI has also been used to visualise the vocal tract and articulators during singing [89,90], swallowing [91–93], laughter [94], beatboxing [95,96] and the playing of musical instruments [97,98,107,108,99–106].

Accurate vocal tract and articulator visualisation during speech requires imaging at spatio-temporal resolutions sufficient to capture the motion of these anatomical features. Recommendations on dynamic speech MRI spatio-temporal resolutions have been published by a group of dynamic speech MRI experts [12]. For example, the group recommended an in-plane spatial resolution of <5 mm^2 and a temporal resolution of <150 ms for capturing the general motion of the soft palate during speech. The spatio-temporal resolutions recommended by the experts are shown in Figure 7.

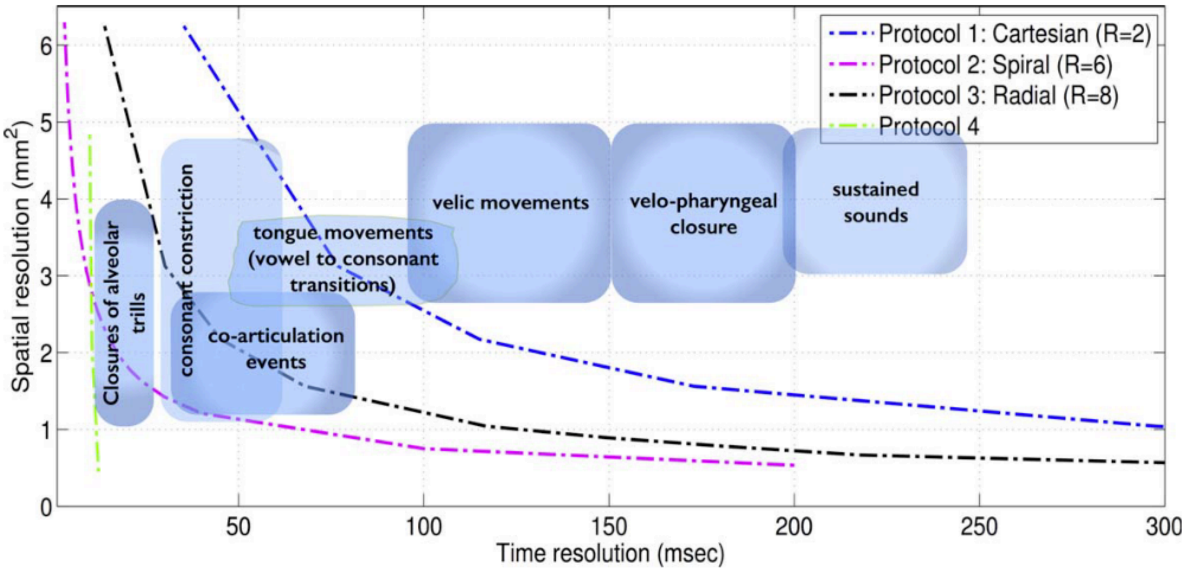

Figure 7: Spatio-temporal resolutions recommended by dynamic speech MRI experts for accurate capture of vocal tract and articulator motion during speech [12].

A wide variety of triggered and rtMRI techniques have been developed for 2D (both single- and multi-slice) [9,10,12,27,61,109], pseudo-3D (i.e. stacks of contiguous slices) [110] and 3D imaging [111–113] of the vocal tract and articulators during speech. Overviews of most of these techniques are given in the review articles of Scott et al. [27] and Nayak et al. [61]. While techniques have been developed for multi-slice 2D, pseudo-3D and 3D imaging of the vocal tract and articulators during speech, typically a series of 2D images of a midsagittal slice of the head are acquired in dynamic speech MRI studies. Examples of such images are shown in Figure 8. Acquisition of 2D midsagittal image series is desirable in clinical speech assessment as the images show a view of the vocal tract and articulators similar to videofluoroscopy, one of the imaging techniques most commonly used in clinical speech assessment, and therefore a view that clinicians are familiar with and can more easily interpret.

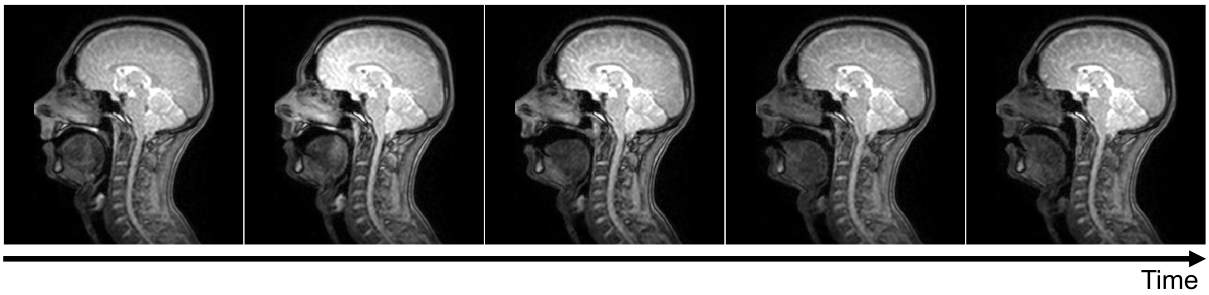

Figure 8: A series of magnetic resonance images of a midsagittal slice of the head during speech, acquired at a temporal resolution of 100 ms.

State-of-the-art triggered techniques enable imaging of speech at the highest spatio-temporal resolutions. More specifically, these techniques enable 2D imaging of a single slice at a spatial resolution of 2.2×2.2 mm^2 and a temporal resolution of 9.8 ms [109], pseudo 3D imaging at a spatial resolution of 1.875×1.875×2.000 mm^3 and a temporal resolution of 28 ms [110], and 3D imaging at a spatial resolution of 2.2×2.2×5.0 mm^3 and a temporal resolution of 6 ms [112]. However, triggered techniques require continuous repetition of a speech task during an extended period of time. For example, the state-of-the-art 2D, pseudo 3D and 3D imaging techniques require continuous repetition of a speech task for 1.7, 19.5 and 7.2 minutes respectively [109,110,112]. While continuous repetition of a speech task for these durations may be feasible for healthy subjects, it is not for patients with speech problems. Triggered techniques are therefore not the most suitable for use in clinical speech assessment.

Real-time techniques allow imaging of speech as it occurs, without requiring any repetitions, and are therefore more suitable for use in clinical speech assessment than triggered techniques. State-of-the-art real-time techniques enable 2D imaging of a single slice at a spatial resolution of <2.4×2.4 mm^2 and a temporal resolution of <20 ms [9,10], and 3D imaging at a spatial resolution of 2.2×2.2×5.8 mm^3 and a temporal resolution of 61 ms [113]. However, these techniques require highly specialised MRI equipment and software, namely custom receive coils [10] and/or specialised pulse sequences and reconstruction methods [9,10], that are not widely available especially in clinical practice. These requirements therefore prevent the widespread adoption of the techniques, a limitation that has motivated the development of techniques that only require widely available standard MRI equipment and software [11,27,114,115]. Techniques that only require standard MRI equipment and software enable 2D imaging at spatial resolutions of <2.4×2.4 mm^2 and temporal resolutions <100ms. While these spatio-temporal resolutions are lower than those of state-of-the-art techniques, they are nevertheless sufficient to capture the general motion of articulators such as the soft palate [12].

To widen access to real-time speech MRI data and therefore stimulate research in the field, several datasets have been made publicly available [18,19,116–121]. Most of these datasets include 2D midsagittal image series of English [18,116,117] or French [19,118] speakers performing phonologically comprehensive speech tasks (i.e. speech tasks designed to include most phonemes in a wide range of contexts). The other datasets include 2D midsagittal image series of English speakers producing emotional speech [119], repeating several speech tasks consisting of vowel-consonant-vowel sequences [120], and imitating unfamiliar speech sounds [121].

3.1.8 *Levator Veli Palatini* Visualisation

Visualisation of the LVP provides information about the shape and configuration of the muscle. There is increasing interest in LVP visualisation, to better understand variations in the shape and configuration of the muscle [25,122,131–140,123,141–143,124–130], to aid planning of surgical treatment of VPI [144,145], and for medical education purposes [146]. MRI is predominantly used for LVP visualisation [13,25,130–139,122,140,142,143,123–129], due to its unique ability to acquire images of any orientation with excellent soft tissue contrast without using ionising radiation.

3.1.9 MRI of the *Levator Veli Palatini*

Due to the small size of the LVP and its 3D structure, 3D imaging at a high spatial resolution is required to fully visualise the muscle. Previous work has predominantly used 3.0 T MRI at a spatial resolution of 0.8×0.8×0.8 mm^3 for 3D LVP visualisation [25,126,138–140,127–129,131–133,136,137]. The motivation for imaging at 3.0 T rather than at lower magnetic field strengths is the acquisition of images with greater signal-to-noise ratios, enabling improved visualisation of anatomical features [147]. Nevertheless, a few previous works used 1.5 T MRI for 3D LVP visualisation [13,132,133].

The LVP and the soft tissue that surrounds it have very similar tissue properties. Consequently, a challenge when imaging the LVP is ensuring that the image contrast between the LVP and the surrounding soft tissue is sufficient to discriminate between the two. Previous work has predominantly acquired T_2-weighted 3D images of the LVP at 3.0 T using TSE pulse sequences [25,126,139,140,127–129,131,134,136–138]. In addition, a recommendation to acquire T_2-weighted images for assessing the LVP in clinical practice was recently made [8]. However, the results of recent work which investigated the optimal image contrast for identification of LVP landmarks in 3D images acquired at 1.5 T suggest that T_1- or PD-weighted images may enable more accurate identification [13]. However, the literature contains no reports of equivalent investigations into the optimal image contrast for 3D LVP visualisation at 3.0 T.

3.2 Deep Learning

3.2.1 Machine Learning

A machine learning algorithm is an algorithm that is able to learn from data [148]. In this context, an algorithm is considered to learn if its ability to perform a task improves with experience [149]. Machine learning algorithms can be broadly categorised as supervised or unsupervised, depending on the data they learn from. Supervised learning algorithms learn from data that includes ground-truth (GT) labels, while unsupervised learning algorithms learn from data that do not include such labels. Other categories of machine learning algorithms exist. Two notable examples of these are semi-supervised learning algorithms and reinforcement learning algorithms. Semi-supervised learning algorithms learn from relatively small amounts of data that include GT labels and relatively large amounts of data that do not. This type of algorithm therefore lies between supervised and unsupervised learning algorithms. Reinforcement learning algorithms interact with a dynamic environment and learn from these interactions via feedback loops.

Machine learning algorithms create models using data. In recent years, models based on artificial neural networks (ANNs) have attracted much attention, in particular those based on ANNs with many layers. Since the number of layers of an ANN is referred to as its depth, ANNs with many layers are considered to be deep and are therefore referred to as deep learning models. In recent years, the field of deep learning has advanced and expanded rapidly. Deep learning models have been developed to perform various tasks in a wide variety of fields including medical image analysis. The predominant type of ANN that deep learning models for medical image analysis are based on is the convolutional neural network (CNN), although recently deep learning models based on vision transformers (ViTs), another type of ANN, have also begun to gain popularity. The next sections will introduce ANNs, CNNs and ViTs, and provide an overview of how these models are developed.

3.2.2 Artificial Neural Networks

ANNs are a type of machine learning model [150]. They are networks that consist of interconnected layers of units (also known as artificial neurons since they aim to mimic to some degree the operation of biological neurons), as illustrated in Figure 9A. The first and last layers of an ANN are known as the input and output layers respectively, while layers between these are known as hidden layers. ANNs with multiple hidden layers are considered to be deep neural networks. Consequently, machine learning using deep neural networks is known as deep learning.

Each unit of an ANN has one or more inputs, $\boldsymbol{x}$, and transforms these into a scalar output, a, in a non-linear manner according to the following equation:

$$a = \sigma(\boldsymbol{w}^T \boldsymbol{x} + b) \tag{3}$$

where $\boldsymbol{w}$ is a vector of weights, b is a scalar bias and σ is a non-linear function such as the sigmoid function or the hyperbolic tangent function. The combination of the weights and biases of the units of an ANN are the model parameters, denoted by θ, that are updated as the ANN learns from data. Due to the multiple layers in an ANN, data are transformed in a non-linear manner multiple times as they pass through the ANN. This series of transformations enables ANNs to learn complex non-linear patterns in data. Feedforward ANNs (also known as multi-layer perceptrons) contain no feedback connections. In other words, the outputs of the units in a layer are only used as inputs to units in deeper layers. ANNs that include feedback connections are known as recurrent neural networks (RNNs). An example of an RNN is shown in Figure 9B.

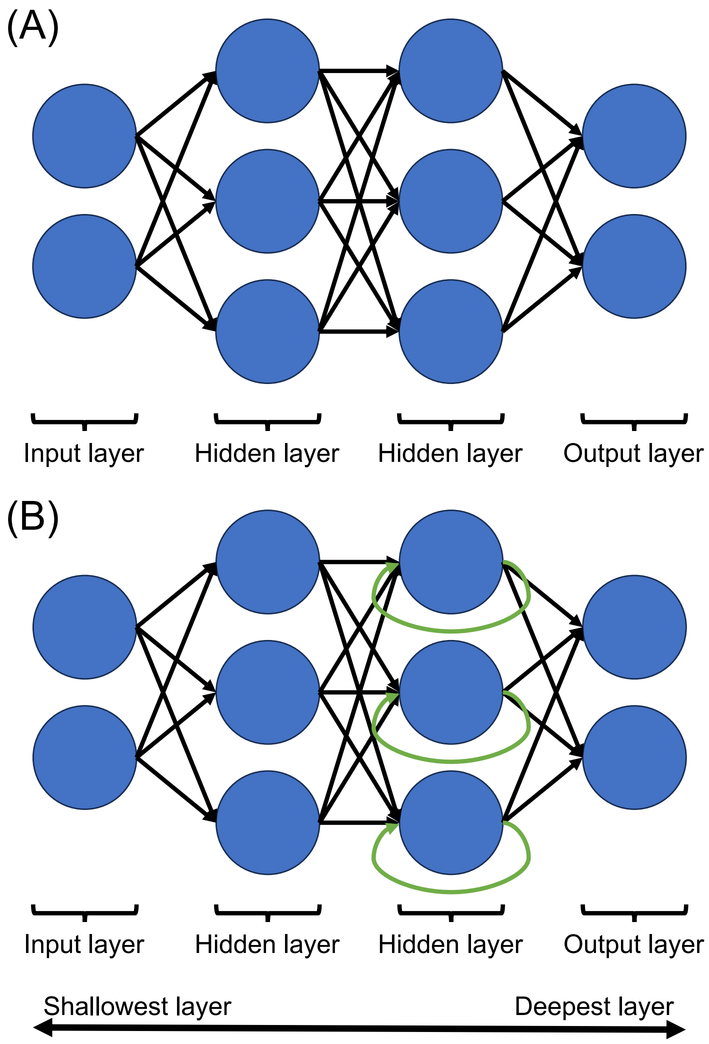

Figure 9: Examples of artificial neural networks. (A) A multi-layer perceptron with four layers. (B) A recurrent neural network with four layers. Blue circles indicate units, black arrows indicate connections between units and green arrows indicate feedback connections.

3.2.3 Supervised Training of ANNs

An ANN is effectively a function, f, that maps an input, $\boldsymbol{x}$, to an output, y:

$$y = f(\boldsymbol{x};\theta) \tag{4}$$

An iterative process is carried out to enable the ANN to learn parameters that result in f approximating the function, f^*, that maps the input to the corresponding GT label, y^*:

$$y^* = f^*(\boldsymbol{x}) \tag{5}$$

This process is known as supervised training of the network and consists of four main steps that are repeated multiple times. First, the network is inputted with data vectors, $X \in \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_i}\}$, and estimates labels, $Y \in \{\boldsymbol{y_1}, \boldsymbol{y_2}, \dots, \boldsymbol{y_i}\}$, for these vectors:

$$Y = f(X;\theta) \tag{6}$$

This step is known as forward propagation and the number of data vectors that the network is inputted with, i, is known as the mini-batch size. Second, the labels estimated by the network are compared with the GT labels of the data vectors, $Y^* \in \{\boldsymbol{y_1^*}, \boldsymbol{y_2^*}, \dots, \boldsymbol{y_i^*}\}$. The errors between the estimated and GT labels (known as the loss) are quantified using a function, $L(Y^*, Y)$. This function is known as the loss function and can consist of one or more terms. Third, the derivatives of the loss with respect to the parameters of each unit of the network are calculated using the chain rule:

$$\frac{\partial L}{\partial w_{jk}^l} \tag{7}$$

where w_{jk}^l is the weight between the kth unit in layer $l-1$ and the jth unit in layer l.

Finally, an optimizer is used to update θ according to the derivatives calculated in the third step:

$$w_{jk}^{l} \leftarrow w_{jk}^{l} - \lambda \frac{\partial L}{\partial w_{jk}^{l}} \tag{8}$$

where λ is a hyperparameter called the learning rate. Commonly used optimizers include stochastic gradient descent and Adam [151]. The final two steps are known as backpropagation and are a key requirement to enable θ to be updated in a way that reduces the loss. The goal of training is to find the network parameters, θ^{*}, that minimise the loss function. Training can therefore be formulated as the following optimisation problem:

$$\theta^{*} = \underset{\theta}{\operatorname{argmin}}\, L(Y^{*}, Y) \tag{9}$$

Supervised training requires a dataset consisting of input data vectors and corresponding GT labels, known as a training dataset.

Once a network is fully trained, its performance is quantitatively evaluated using one or more metrics. To enable evaluation of the performance of a fully trained network, a dataset that does not include any of the data in the training dataset is required. This dataset is known as the test dataset.

3.2.4 Generalisation of ANNs

A key challenge in machine learning and deep learning is training networks that perform well on data other than those used in network training. In other words, creating networks that generalise to new data. Techniques commonly used during network training to improve the generalisation of a network include weight decay, dropout, data augmentation and the use of a validation dataset.

Weight decay, also known as L_2 regularisation, aims to prevent individual units from having an excessive influence on the output of the network. It is implemented by including the following term to the loss function:

$$L_{WD}(Y^{*}, Y, \theta) = L(Y^{*}, Y) + \epsilon \|\theta\|_{2}^{2} \tag{10}$$

where ϵ is a scalar constant. The purpose of this term is to prevent the values of the network weights from becoming too large, thus preventing individual units from becoming overly influential. The term achieves this by increasing the loss when the weight values increase, thus encouraging smaller values.

Dropout [152] is when the outputs of a random group of units in a network are set to zero during network training, temporarily preventing the units from contributing to the output of the network. At the end of training involving dropout, the resulting network is in effect an average of several slightly less complex networks, causing an improvement in its generalisation.

Data augmentation aims to increase the generalisation of a network by synthetically increasing the diversity of the training dataset. This increase is achieved by creating modified versions of the training data. When training networks for image analysis tasks, commonly used augmentations include rotation, translation and cropping, in addition to augmentations that modify image pixel or voxel values such as addition of random Gaussian noise [153]. An overview of data augmentation techniques commonly used in the training of networks for medical image analysis tasks is provided in [154].

The duration of training can greatly influence the performance of a network. Training for an insufficient duration prevents the network from maximising its performance, while training for an excessive duration results in the network overfitting the training dataset, thus compromising network generalisation. Typically a validation dataset is used to identify the optimal training duration. Evaluation of a network performance using this dataset, which does not include any data in either the training or test datasets, gives an indication of the network performance on the test dataset. During network training, regular evaluation of network performance using the validation dataset enables identification of overfitting and therefore informs when training should be stopped. A validation dataset is also often used to identify the values of hyperparameters such as the learning rate that result in a network with the greatest generalisation, a process known as hyperparameter optimisation.

3.2.5 Limitations of ANNs

Typically in ANNs, units in adjacent layers are fully connected. In other words, in two adjacent layers, each unit in the shallower layer is connected to every unit in the deeper layer. Increasing the number of layers and units of an ANN results in a large increase in the number of connections and consequently the number of parameters to be learned. This relationship causes an unavoidable tradeoff between the complexity of a model and the computational and data requirement to train and deploy it. To mitigate this tradeoff, other types of ANN that are more sparsely connected have been developed, such as CNNs.

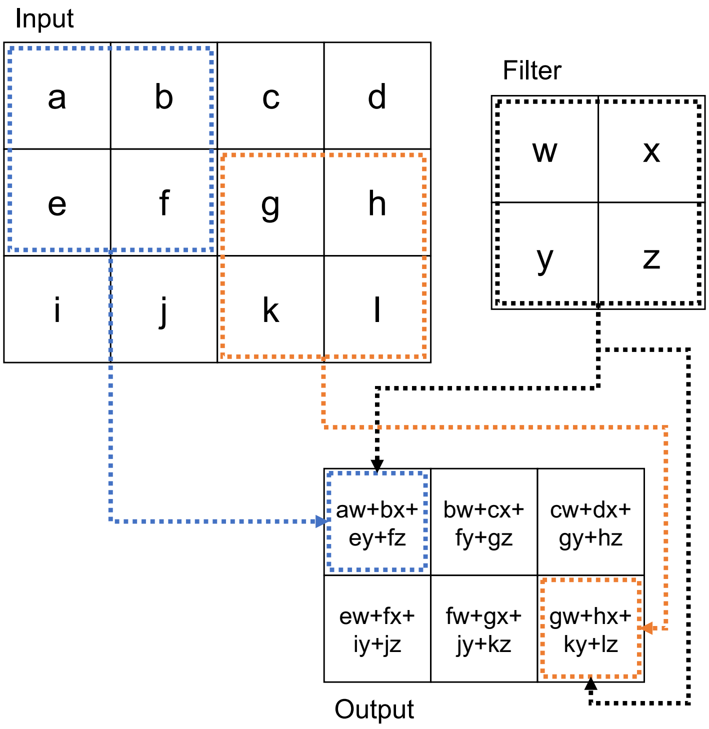

Figure 10: An example of two-dimensional convolution. The 3×4 input is convolved with a 2×2 filter to create a 2×3 feature map.

3.2.6 Convolutional Neural Networks

CNNs are a specific type of ANN that are designed to learn from data with a grid-like arrangement such as images [155]. These networks contain one or more layers in which a mathematical operation called convolution is used to transform the input to the layers. These layers are known as convolutional layers. Typically, a convolutional layer consists of three consecutive mathematical operations: convolution, normalisation and then a non-linear transformation.

The convolution operation consists of convolving the input to the layer with a set of filters (also known as kernels) with learnable weights and biases. The output of this operation is a set of feature maps that correspond to the response of the filters at different spatial locations of the input. Typically, filters for 2D and 3D convolutions have a size of 3×3 pixels and 3×3×3 voxels respectively, although other filter sizes are possible. An example of 2D convolution is shown in Figure 10.

Following the convolution operation, the feature maps are normalised. The method most commonly used to normalise feature maps is batch normalisation [156] where feature maps are normalised across a mini-batch rather than on an individual basis. The motivation for normalising feature maps is to stabilise the distribution of outputs from the convolutional layer, as such stabilisation has been shown to accelerate the training of CNNs [156]. Other methods to normalise feature maps have also been developed, such as instance normalisation [157] where feature maps are normalised on an individual basis rather than across a mini-batch.

Finally, normalised feature maps are transformed in an element-wise manner using a non-linear function. The most commonly used non-linear function in CNNs is the rectified linear unit (ReLU) which transforms a scalar value, z, in the following way:

$$ReLU(z) = \max{(0, z)} \tag{11}$$

Another commonly used non-linear function is the leaky ReLU [158] which transforms z in the following way:

$$LeakyReLU(z) = \max(0, z) + k \cdot min(0, z) \tag{12}$$

where k is a scalar constant.

In addition to convolutional layers, CNNs contain pooling layers. Pooling layers typically occur after convolutional layers and reduce the spatial dimensions of the outputs of convolutional layers, usually by a factor of two. Pooling layers achieve this by first partitioning the outputs into non-overlapping regions, then calculating a summary statistic such as the maximum value in each region and finally creating a new output consisting of the summary statistics. The aim of pooling layers is to make CNNs approximately invariant to small translations of the input image. The operator most commonly used in pooling layers is max pooling. Max pooling layers identify the maximum value in each region and then create a new output consisting of these values, as illustrated in Figure 11.

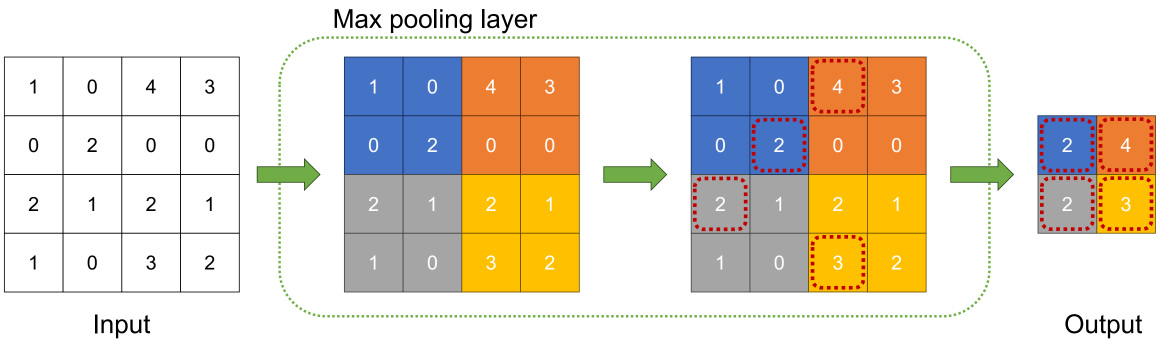

Figure 11: An example of a max pooling layer. The input to the layer is a feature map of size 4×4. First, the feature map is partitioned into non-overlapping regions of size 2×2 pixels (different colours indicate different regions). The maximum value in each region is then identified (dotted red line indicates pixel with maximum value in each region). The output of the layer is a feature map of the maximum values.

3.2.7 CNN-Based Image Analysis

CNN-based methods have been developed to perform a range of image analysis tasks. The most common of these tasks include image classification, object detection and image segmentation. Image classification is the process of assigning a label to an image according to its content. For example, given a set of images of handwritten digits, an image classification method would label the images according to the digit they show. Object detection is the process of detecting and locating instances of objects in images. Typically, CNN-based object detection methods estimate the coordinates of the bounding box that contains the object. Image segmentation is the process of partitioning images into regions of pixels (or voxels) [159]. It consists of assigning a label to each pixel in an image, in such a way that pixels with shared characteristics (such as pixels of the same object) are assigned the same label. There are two main types of image segmentation: semantic segmentation and instance segmentation. Semantic segmentation assigns a class label to every pixel in an image. If there are multiple instances of an object in an image, a semantic segmentation method would assign the same label to each pixel showing an instance of the object. Instance segmentation assigns a different label to pixels of different instances of an object in an image. If there are multiple instances of an object in an image, an instance segmentation method would assign a different label to pixels showing a different instance of the object.

Several CNN-based image analysis methods have particularly influenced the deep learning and image analysis communities in recent years. These methods include AlexNet [160], VGG [161], the method developed by [162] and ResNet [163]. AlexNet [160], a method for image classification, is credited for triggering renewed interest in CNN-based methods for image analysis after winning the ImageNet Large Scale Visual Recognition Challenge 2012 by a large margin. Inspired by AlexNet, Simonyan and Zisserman [161] developed several CNN-based methods for image classification with different depths, and showed that increased CNN depth resulted in improved CNN performance. This work is also credited for triggering the trend of using 3×3 filters in convolutional layers. Long et al. [162], were the first to develop a fully convolutional neural network (FCN) for semantic segmentation. This network consisted of convolutional layers and pooling layers only, hence its description as an FCN. Inspired by the work of Simonyan and Zisserman, He et al. [163] developed residual blocks for CNNs that enabled the training of even deeper CNNs with improved performance.

3.2.8 Vision Transformers

Transformers are a type of deep learning model initially developed for natural language processing (NLP) that are particularly effective at capturing long range correlations in data [164]. Unlike CNNs, transformers do not involve convolutions and instead use self-attention mechanisms. Following their great success in NLP, transformers were extended to be suitable for image analysis tasks [165]. Such transformers, known as vision transformers (ViTs), have attracted much interest from the medical image analysis community, resulting in their application to a range of medical image analysis tasks including image classification, segmentation and registration. Two recent review articles provide overviews of ViT-based medical image analysis methods [166,167]. The emergence of ViTs has prompted the community to reconsider the supremacy of CNNs for medical image analysis tasks.

3.3 Medical Image Segmentation

Image segmentation has numerous clinical applications ranging from radiotherapy [168,169] to neuroimaging [170] and cardiac imaging [171,172]. In many of these applications, image segmentation is an important step to enable measurement of clinical biomarkers that inform diagnosis or treatment decisions. For example, segmentation of the heart chambers in cardiac MR images enables measurement of biomarkers such as ejection fraction [171,172]. A wide variety of medical image segmentation methods have been developed, to segment anatomy/pathology ranging from the heart to the brain in medical images acquired using imaging techniques ranging from US to MRI. Consistent with trends in other image analysis fields, recently most of the medical image segmentation methods that have been developed are DL-based. Several recent review articles provide overviews of these methods [154,166,168–174]. However, a key requirement for the development of such methods is the availability of GT segmentations. Obtaining such segmentations is typically a time-consuming manual process which, particularly for medical images, requires input by specialists and is prone to intra-and inter-observer variability.

3.3.1 Deep-Learning-Based Medical Image Segmentation

The majority of medical image segmentation methods that have been developed in recent years have been based on FCNs. To begin with, vanilla FCN-based methods were developed and then widely applied. Notable examples of such methods include U-Net [175] and SegNet [176] for 2D segmentation and 3D U-Net [177] for 3D segmentation.

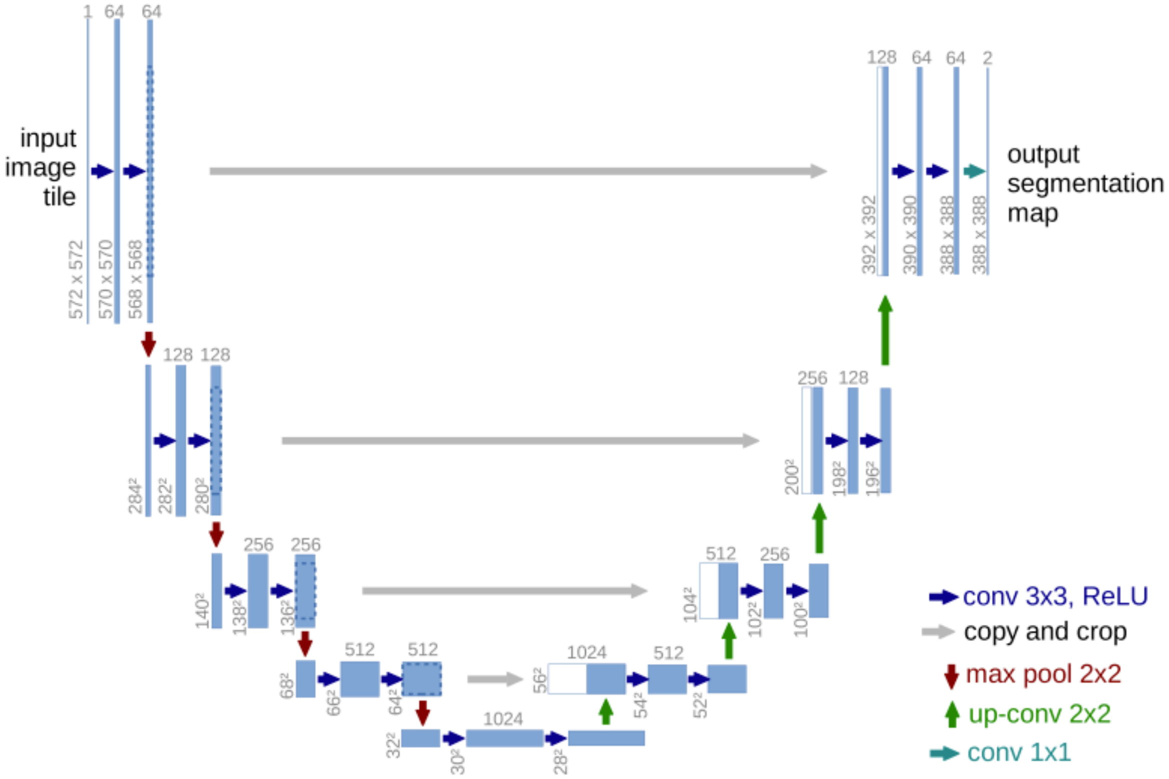

Figure 12: An overview of the architecture of the U-Net [175] *fully convolutional network (image from* [175]*), consisting of two-dimensional convolutions (conv), transposed convolutions (up-conv), rectified linear units (ReLUs), skip connections (grey arrows) and max pooling.*

Inspired by the work of [162], U-Net used an FCN with the architecture shown in Figure 12 to segment 2D images [175]. The FCN consisted of a multi-layer encoder to downsample feature maps, followed by a multi-layer decoder to upsample feature maps. The network included skip connections that enabled the combination of feature maps from corresponding layers in the encoder and decoder. A weighted cross entropy loss was used to train the FCN. This loss was weighted according to the number of pixels in each class and the distance of a pixel from a boundary. However, typically the latter weighting is not included in the training of widely-applied U-Net-based methods in medical image analysis.

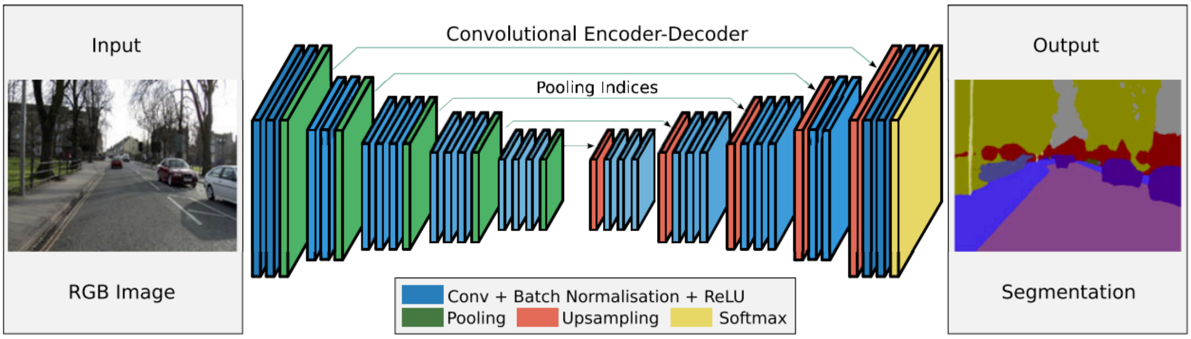

Figure 13: An overview of the architecture of the SegNet [176] *fully convolutional network (image from* [176]*), consisting of two-dimensional convolutions (Conv), batch normalisation, rectified linear units (ReLUs), max pooling (Pooling), upsampling using max pooling indices (Upsampling) and a softmax activation function (Softmax).*

SegNet used an FCN with a similar architecture to U-Net to segment 2D images. However, the FCN did not include skip connections. In addition, feature map upsampling in the multi-layer decoder was not achieved using transposed convolutions. Instead, upsampling was achieved using max pooling indices from corresponding layers in the multi-layer encoder (see Figure 13).

U-Net was extended to enable 3D segmentation by replacing the 2D operations by 3D operations [177]. Similarly to the original U-Net, a weighted cross entropy loss was used to train the FCN. However, this loss was only weighted according to the number of pixels in each class. In addition, unlike the original U-Net, the 3D U-Net included batch normalisation.

Since the development of vanilla FCN-based methods, there have been several general trends in the way these methods have been extended by the medical image analysis community. Unsurprisingly, these trends mirror those in other image analysis fields. These trends include adding residual blocks to FCNs, adding RNNs to FCNs, using series of consecutive FCNs and, most recently, developing hybrid FCN-ViT-based segmentation methods.

The success of CNNs with residual blocks for image classification [163] motivated the inclusion of such blocks in FCN-based medical image segmentation methods. One such method that has been particularly widely applied and extended by the medical image analysis community is V-Net [178]. V-Net used an FCN with residual blocks to segment 3D images. Similarly to the 3D U-Net FCN, the V-Net FCN consisted of a multi-layer encoder followed by a multi-layer decoder with skip connections. However, feature map downsampling in the V-Net FCN was achieved using 2×2×2 convolutions with a stride of 2 rather than using max pooling. In addition, the V-Net FCN included parametric ReLUs rather than ReLUs and was trained using a loss function based on the Dice coefficient rather than cross entropy.

Adjacent or consecutive medical images provide contextual information, however, vanilla FCN-based segmentation methods were not designed to exploit this information. This limitation motivated the inclusion of RNNs, which are designed to exploit such information, in FCN-based medical image segmentation methods. Examples of such methods include the one developed by [179] for multi-slice cardiac MR image segmentation and the one developed by [180] for 3D electron microscopy image segmentation.

Typically, the anatomical features of interest in a medical image occupy a relatively small proportion of the image. This fact has motivated the development of methods that use multi-stage approaches to segment an image [181,182]. Generally, the first stage aims to identify the area or volume that contains the anatomical features of interest. Only this area or volume is segmented in subsequent stages. For example, in the first stage of the method developed by [181], a U-Net-based FCN is used to estimate an initial segmentation of an entire image. The region of the image containing the anatomical features of interest is determined from this segmentation and then another U-Net-based FCN is used to estimate a more detailed segmentation of this region of the image only.

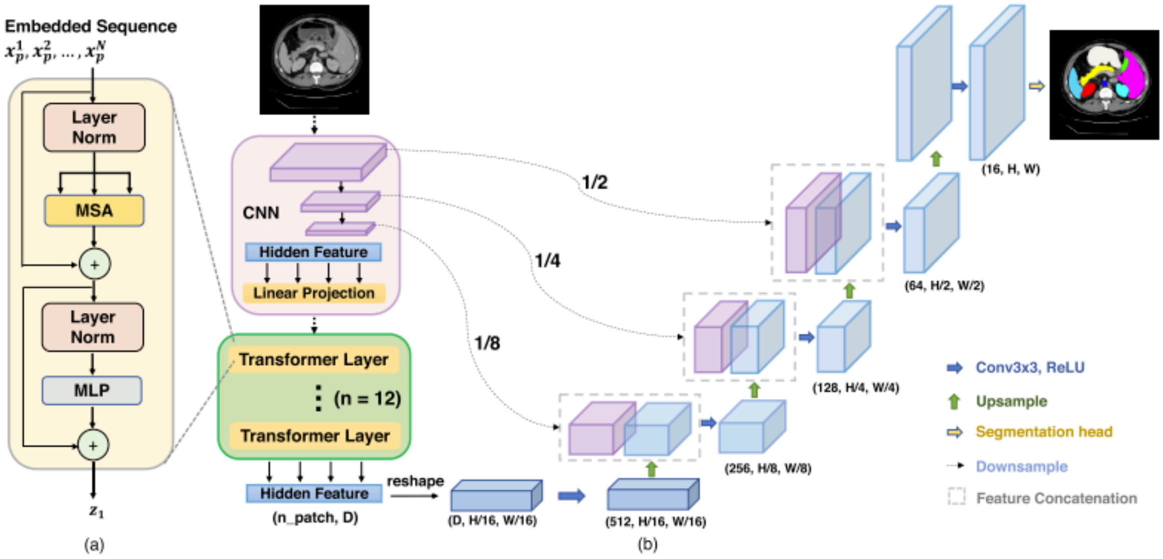

Figure 14: An overview of the TransUNet [183] *hybrid network, a fully convolutional network with a vision transformer between the encoder and decoder (image from* [183]*). (a) An overview of a transformer layer consisting of layer normalisations (Layer Norms), a multi-head self-attention block (MSA) and a multi-layer perceptron (MLP) block. (b) An overview of the hybrid network consisting of two-dimensional convolutions (Conv), rectified linear units (ReLUs), cascaded upsamplers (Upsample). The height (H) and width (W) of the feature maps are shown.*

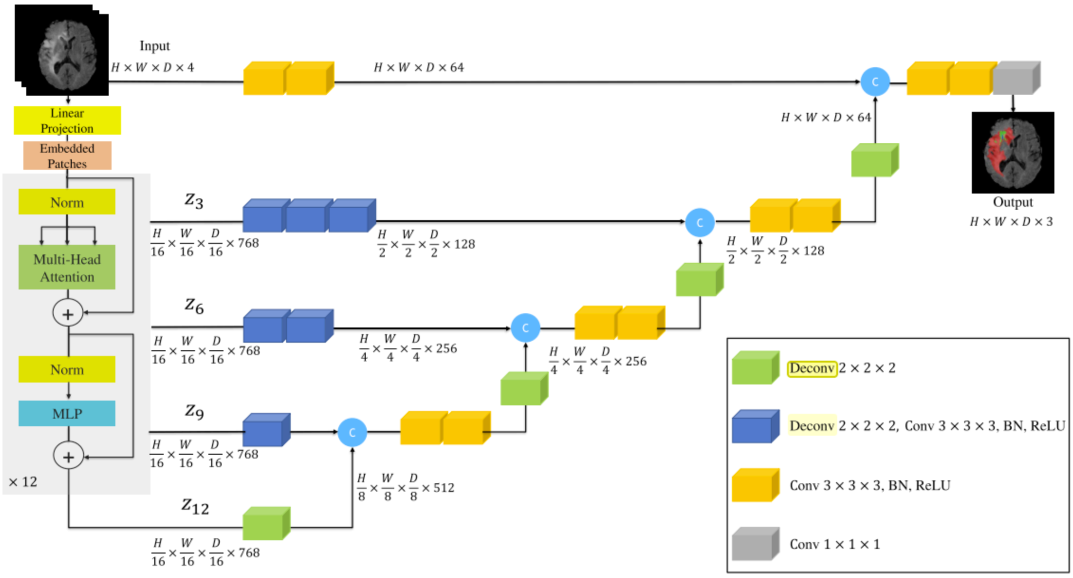
Input
$H \times W \times D \times 4$
Linear Projection
Embedded Patches
Norm
Multi-Head Attention
Norm
MLP
× 12
z_3
z_6
z_9
z_{12}
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$H \times W \times D \times 64$
$H \times W \times D \times 64$
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 128$
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 128$
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 256$
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 256$
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 512$
c
Output
$H \times W \times D \times 3$
Deconv $2 \times 2 \times 2$
Deconv $2 \times 2 \times 2$, Conv $3 \times 3 \times 3$, BN, ReLU
Conv $3 \times 3 \times 3$, BN, ReLU
Conv $1 \times 1 \times 1$

Figure 15: An overview of the UNETR [184] *hybrid network consisting of a vision transformer (ViT) as the encoder and a fully convolutional network (FCN) as the decoder (image from* [184]*). Each of the 12 layers in the ViT include layer normalisations (Norms), a multi-head self-attention block (Multi-Head Attention) and a multi-layer perceptron (MLP) block. The FCN includes three-dimensional convolutions (Conv), rectified linear units (ReLUs), batch normalisation (BN), transposed convolutions (Deconv). The height (H), width (W) and depth (D) of the feature maps are shown.*

Most recently, the success of ViTs in other fields of deep learning has motivated the development of hybrid FCN-ViT-based medical image segmentation methods. A recent review article provides an overview of such methods [166]. The rationale for combining FCNs and ViTs is to exploit both the accurate localisation abilities of the former and the global context identification abilities of the latter. Methods such as TransUNet [183] used a U-Net-based FCN with a ViT between the encoder and decoder (see Figure 14) to segment abdominal CT images and cardiac MR images, while methods such as UNETR [185] and Swin UNETR [185] used a U-Net-based FCN with the encoder replaced by a ViT and a swin transformer respectively to segment MR and CT images of various body organs (see Figure 15).

As noted earlier, the U-Net architecture [175,177] has proved to be particularly popular in the medical image analysis field, influencing the development of a range of other methods [186]. While a wide variety of complex FCN-based medical image segmentation methods have been developed, recent work has shown that vanilla U-Net-based methods can outperform these more complex ones if configured and trained effectively [187]. Based on this observation, nnU-Net, an automatic method for configuring and training U-Net effectively has been developed [187]. Given a training dataset, the method determines a suitable U-Net FCN architecture, whether image pre-processing steps are required (for example, cropping and normalisation) and an effective FCN training strategy including selection of key hyperparameters such as the image patch size and the mini-batch size. Recently, methods developed using nnU-Net have segmented a wide variety of medical images including 2D cardiac MR images [188], 3D CT images of the kidney [189], pseudo-3D MR images of the brain [190] and 3D CT images of the lungs [191,192] with state-of-the-art accuracy.

3.3.2 Loss Functions and Evaluation Metrics

Supervised training of deep learning models for medical image segmentation consists of the four main steps described in section 3.2.3. Loss functions commonly used in this training include the mean cross entropy loss, L_{CE}, and Dice loss, L_{DSC}:

$$L_{CE} = -\frac{1}{K}\sum_{c=1}^{C}\sum_{k=1}^{K} g_k^c \log s_k^c \qquad (13)$$

where K is the number of pixels or voxels in the image, C is the number of segmentation classes, g_k^c is the binary GT label indicating if pixel k belongs to segmentation class c and s_k^c is the probability estimated by the CNN that pixel k belongs to segmentation class c.

$$L_{DSC} = 1 - 2\frac{\sum_{c=1}^{C}\sum_{k=1}^{K} g_k^c s_k^c}{\sum_{c=1}^{C}\sum_{k=1}^{K}(g_k^c + s_k^c)} \tag{14}$$

Once a model has been trained, its performance is quantitatively evaluated. A recent review article provides an overview of the metrics commonly used by the medical image analysis community to quantify the accuracy of segmentation models [193]. Particularly commonly used metrics include the Dice coefficient (DSC) [194], the intersection over union (IoU), general Hausdorff distance (HD) and the average symmetric surface distance (ASD). The DSC and IoU quantify the overlap between two segmentations, while the HD and ASD quantify the discrepancies between the boundaries of two segmentations. The way in which the DSC and IoU are calculated is illustrated in Figure 16, while the way in which the HD and ASD are calculated are illustrated in Figure 17 and Figure 18 respectively.

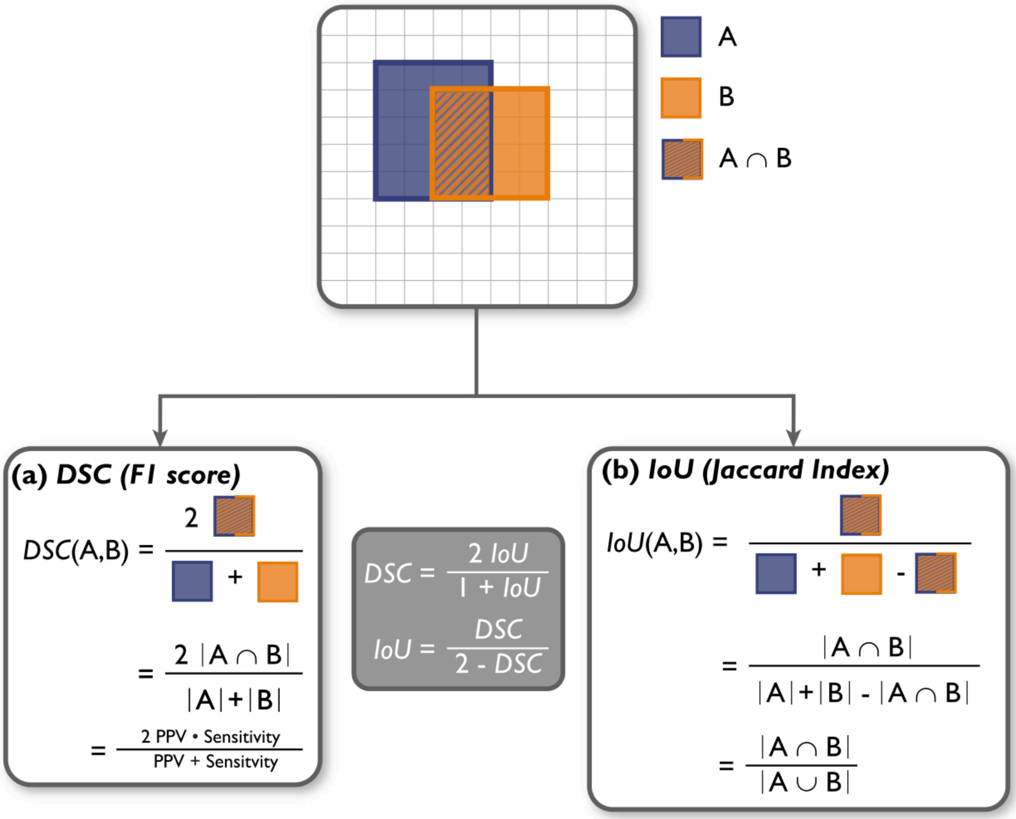

Figure 16: Metrics to quantify the overlap between two segmentations A and B (image from [193]*). (a) the Dice coefficient (DSC). (b) the intersection over union (IoU).*

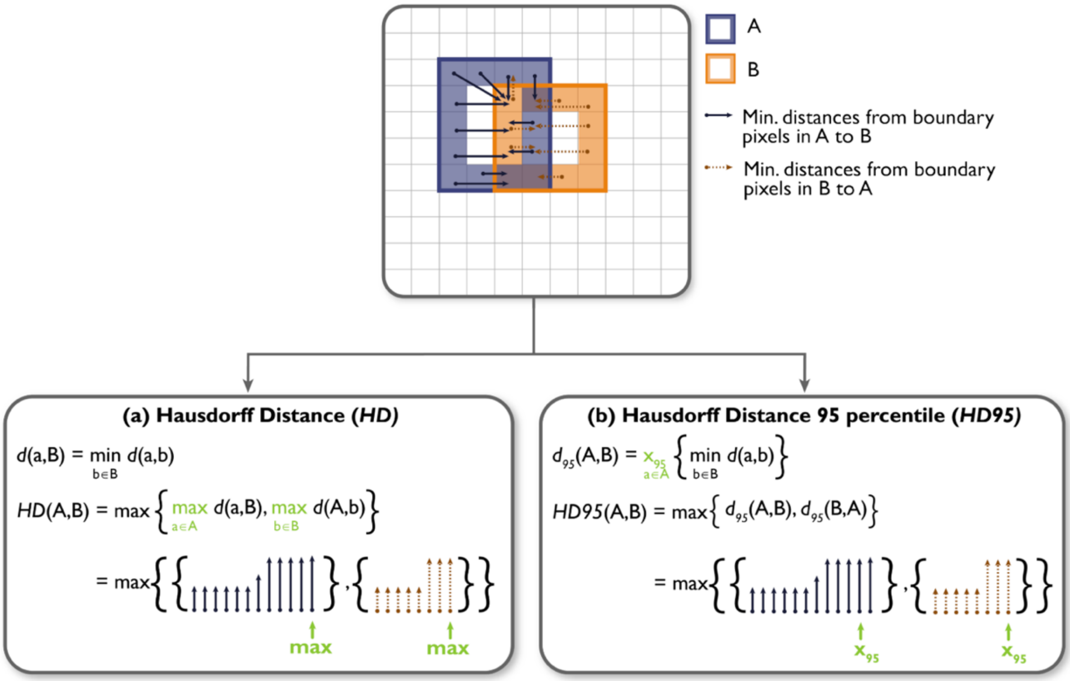
A
B
Min. distances from boundary pixels in A to B
Min. distances from boundary pixels in B to A
(a) Hausdorff Distance (HD)
d(a,B) = min d(a,b) b∈B
HD(A,B) = max { max a∈A d(a,B), max b∈B d(A,b) }
= max
max
max
(b) Hausdorff Distance 95 percentile (HD95)
d95(A,B) = x95 a∈A { min b∈B d(a,b) }
HD95(A,B) = max { d95(A,B), d95(B,A) }
= max
x95
x95

Figure 17: Calculation of the general Hausdorff distance (HD), a metric to quantify the maximum discrepancy between the boundaries of two segmentations A and B (image from [193]*).* $d(a, b)$ *indicates the Euclidean distance between pixel a (a boundary pixel of segmentation A) and pixel b (a boundary pixel of segmentation B).*

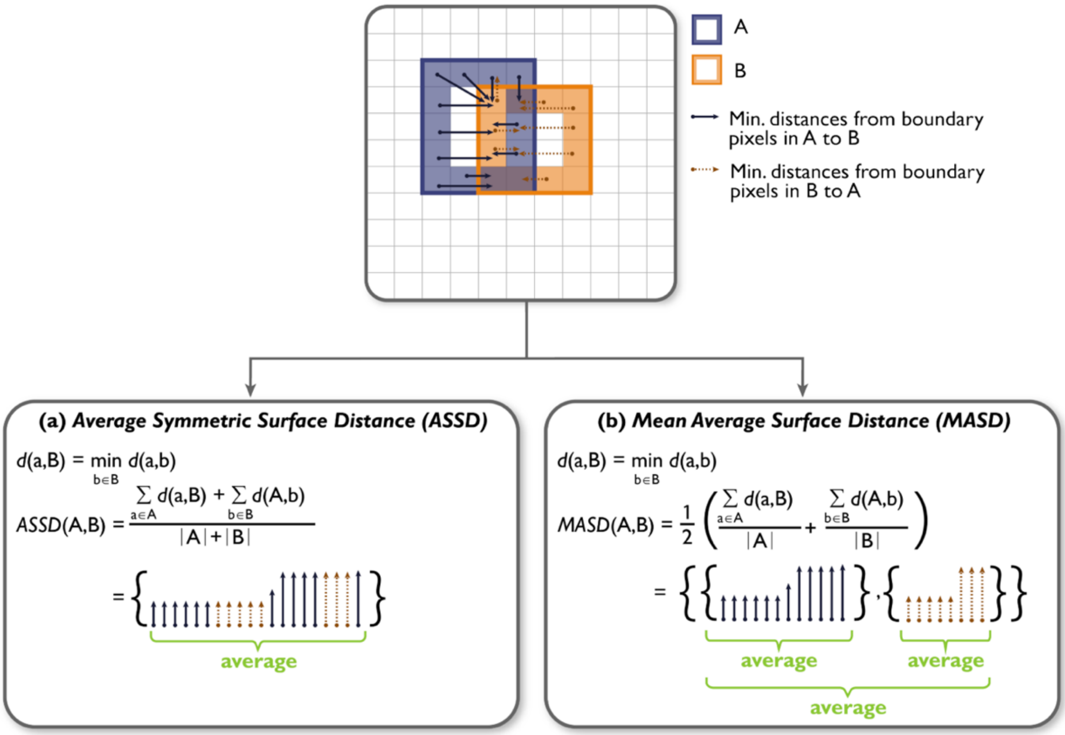
A
B
Min. distances from boundary pixels in A to B
Min. distances from boundary pixels in B to A
(a) Average Symmetric Surface Distance (ASSD)
d(a,B) = min d(a,b) b∈B
ASSD(A,B) = (Σ a∈A d(a,B) + Σ b∈B d(A,b)) / (|A|+|B|)
=
average
(b) Mean Average Surface Distance (MASD)
d(a,B) = min d(a,b) b∈B
MASD(A,B) = 1/2 (Σ a∈A d(a,B) / |A| + Σ b∈B d(A,b) / |B|)
=
average
average
average

Figure 18: Metrics to quantify the average discrepancy between the boundaries of two segmentations A and B (image from [193]*). (a) the average symmetric surface distance. (a) the mean average surface distance.* $d(a, b)$ *indicates the Euclidean distance between pixel a (a boundary pixel of segmentation A) and pixel b (a boundary pixel of segmentation B).*

3.3.3 Real-Time Speech MR Image Analysis

As described in section 3.1.7, use of rtMRI to visualise the vocal tract and articulators during speech is increasing in both research and clinical settings, and typically a series of 2D images of a midsagittal slice of the vocal tract is acquired using real-time speech MRI techniques. In addition, there is increasing interest in extracting quantitative information about the vocal tract and articulators from such images [7,14,75,76,84,88,195–198,20,68–74]. More specifically, there is interest in measuring the size and shape of the vocal tract [14,68,197,198,70,74–76,84,88,195,196], the size, shape and motion of the soft palate [72,73,75,84,198], lip motion [69,75,84], tongue motion [20,84] and the distance between the soft palate and the posterior pharyngeal wall [7,72,198]. Manual measurement to obtain this information is time-consuming, requires input by specialists and is prone to intra- and inter-observer variability. The increasing interest in extracting quantitative information, in combination with the need to avoid manual measurement, have motivated the development of a range of methods to (semi-)automatically extract this information [15,16,206,207,20,199–205]. Almost all these methods are segmentation based [15,16,207,199–206].

3.3.4 Real-Time Speech MR Image Segmentation

Numerous methods based on a variety of approaches have been developed to segment real-time MR (rtMR) images of speech [15,16,207,199–206]. More specifically, these methods segmented 2D images of a midsagittal slice of the head, the type of image most commonly acquired using real-time speech MRI techniques. The majority of these methods [15,199–204,208] were designed to enable (semi-)automatic analysis of the size and shape of the vocal tract in the images, an analysis of particular interest to the speech science community. To enable this analysis, the methods created contours of air-tissue boundaries between the vocal tract and adjacent articulators. Some of these methods created contours without articulator labels [199–202] while others created contours with such labels [15,203,204]. Examples of each type of contour are shown in Figure 19.

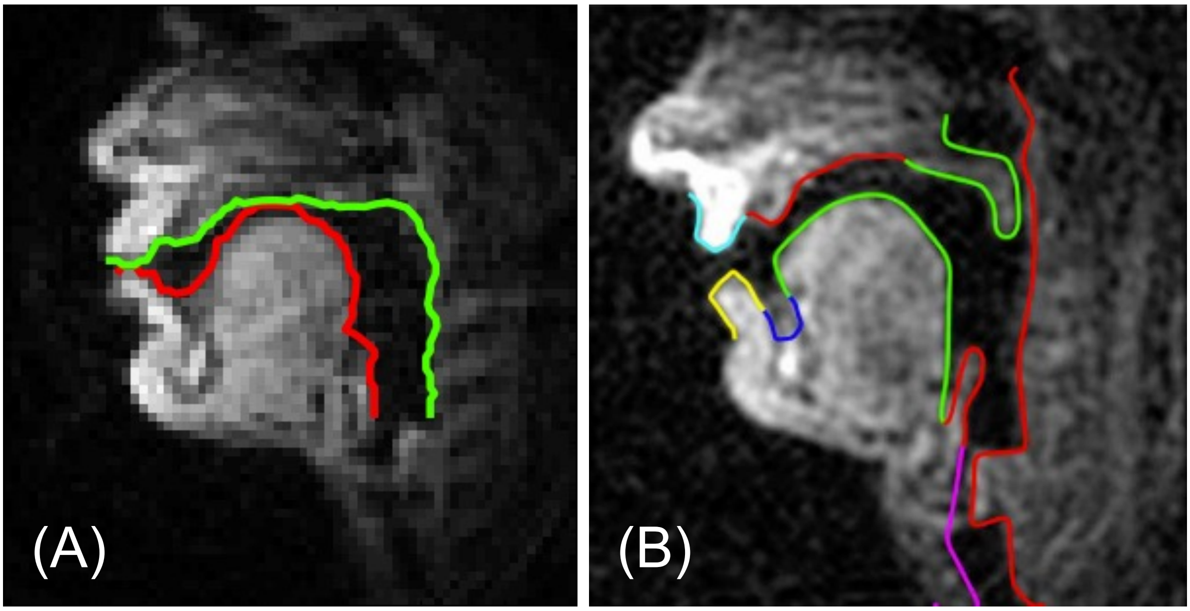

Figure 19: Examples of contours of air-tissue boundaries between the vocal tract and adjacent articulators in real-time speech magnetic resonance images. (A) Contours of the upper (green) and lower (red) air-tissue boundaries without articulator labels (image from [199]*). (B) Contours of air-tissue boundaries with articulator labels indicated by colour coding (image from* [203]*).*

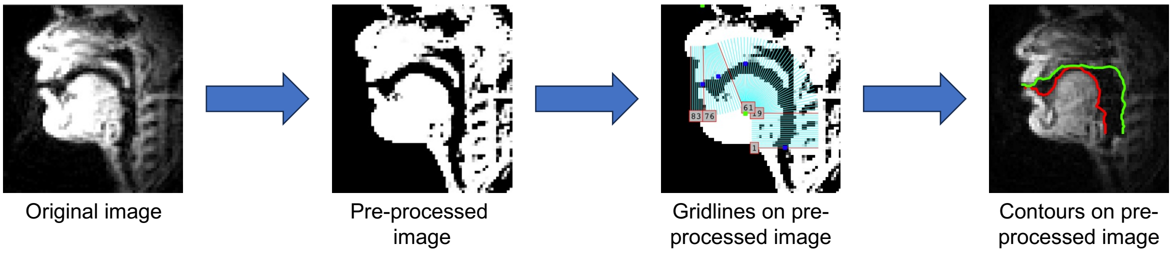
Original image
Pre-processed image
Gridlines on pre-processed image
Contours on pre-processed image

Figure 20: Overview of method developed by [199] *to create contours of air-tissue boundaries between the vocal tract and adjacent articulators in real-time speech magnetic resonance images (modified from* [199]*).*

Instead of creating contours with articulator labels, several methods [199–202] have been developed to create two contours per image: one of the upper air-tissue boundaries, the other of the lower boundaries (see Figure 19A). One method created such contours by first pre-processing the images to increase the image contrast between air and tissue, then superposing gridlines on the pre-processed images and analysing pixel values along these gridlines, and finally using the Viterbi algorithm to identify contours [199]. An overview of this method is shown in Figure 20. Three DL-based methods to create such contours have also been developed [200–202]. These methods all created contours using the same three-stage approach, an overview of which is shown in Figure 21. First, separate FCNs were used to segment three groups of anatomical features in the image as a preliminary step. Second, the contours of the segmentations were identified. Third, the contours were pruned to only include sections corresponding to air-tissue boundaries between the vocal tract and adjacent articulators. One method used FCNs based on SegNet [209] to create contours [201], while the methods developed by [200] and [202] used FCNs based on those developed by [162] and [210] respectively.

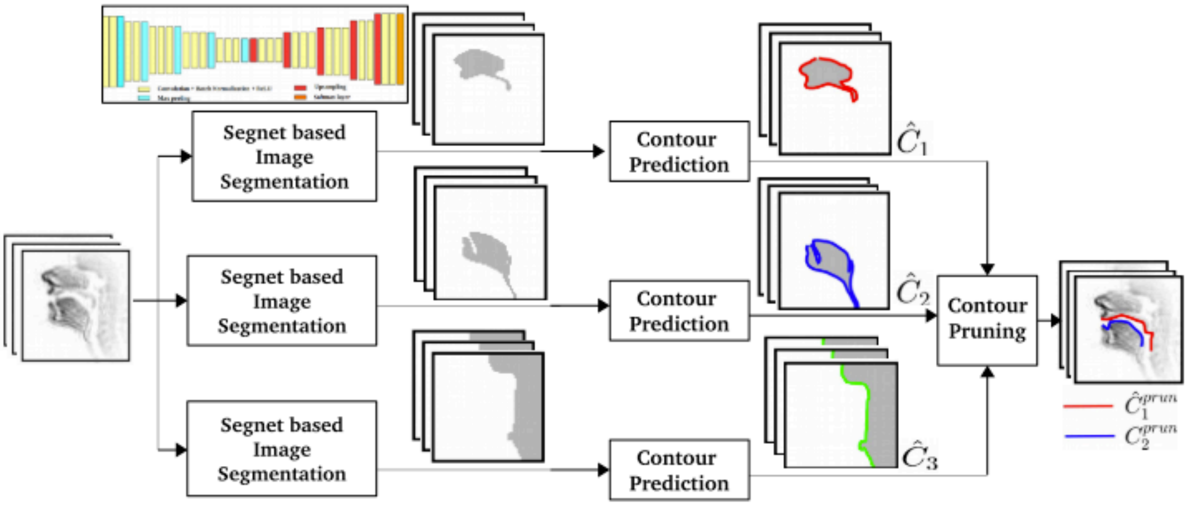

Figure 21: Overview of approach taken by deep-learning-based methods [200–202] *to create contours of air-tissue boundaries between the vocal tract and adjacent articulators in magnetic resonance image of speech (image from* [201]*).*

Several methods have been developed to create contours with articulator labels [15,203,204]. One method created such contours using an optimisation algorithm to iteratively adjust an anatomically informed synthetic image of the vocal tract until the *k*-space of the synthetic image was as similar as possible to the *k*-space of the MR image [203]. Another method based on active appearance models (AAMs) [211] has also been developed to create such contours [204]. This method included two AAMs to create contours: one for images in which the soft palate was in contact with the posterior pharyngeal wall and another for images in which there was no contact. In other work, methods based on multiple linear regression (MLR), active shape models (ASM) [212] and shape particle filtering (SPF) [213] were developed and compared [15]. These methods created separate contours for 10 articulators including the tongue, soft palate and pharyngeal wall. The ASM- and SPF- based methods were initialised using the contours created by the MLR-based method. Evaluated using the mean sum of distances between the closest points on each contour, the ASM-based method was found to be the most accurate. A DL-based method has also been developed to create contours with articulator labels [16]. This method used a single SegNet-based FCN to estimate contours and then refined these contours using an algorithm inspired by the connected component labelling one developed by [214].

Rather than creating contours of air-tissue boundaries between the vocal tract and adjacent articulators, a method to fully segment the vocal tract (see Figure 22A) in real-time speech MR images has also been developed [208]. This method used a FCN based on the original U-Net [175] to segment the vocal tract.

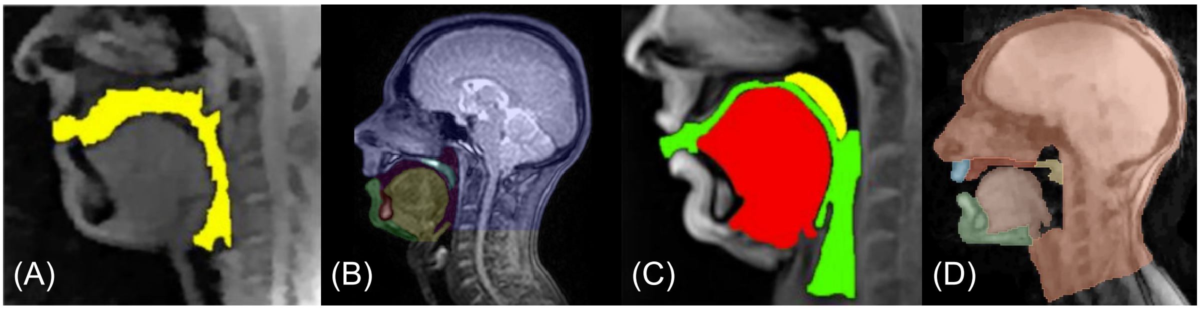

Figure 22: Magnetic resonance images of speech with segmentations estimated by different methods overlaid. (A) The method developed by [208] *segmented the vocal tract only (image from* [208]*). (B) The method developed by* [205] *segmented the head (dark blue), soft palate (light blue), jaw (green), tongue (yellow), vocal tract (pink), tooth space (red). (C) The method developed by* [207] *segmented the soft palate (yellow), tongue (red) and vocal tract (green) (image from* [207]*). (D) The methods developed by* [206] *segmented the head (orange), upper lip (blue), hard palate (red), soft palate (yellow), jaw (green) and tongue (brown) (image from* [206]*).*

While contours of air-tissue boundaries between the vocal tract and adjacent articulators enable analysis of the size and shape of the vocal tract, they only partially segment articulators and consequently do not enable analysis of the size, shape, motion or position of the articulators during speech. Increasing interest in such analysis, by clinicians as well as speech science researchers, has recently motivated the development of methods to fully segment articulators in rtMR images of speech [205–207]. These methods all used U-Net-based FCNs to fully segment the soft palate and tongue in addition to other anatomical features. One method, presented in chapter 5 of this thesis, used a single U-Net-based FCN to estimate segmentations and then refined these segmentations using a post-processing step that removed anatomically impossible regions [205]. This method segmented the following six anatomical features: the head (including the upper lip and hard palate), soft palate, lower lip and jaw, tongue (including the epiglottis), vocal tract and lower incisor space (see Figure 22B). Another method used several U-Net-based FCNs to segment the following three anatomical features: the soft palate, tongue (not including the epiglottis) and vocal tract (see Figure 22C). This method used a separate FCN to segment each anatomical feature. In other work, several methods each using a different U-Net-based FCN were developed and then compared [206]. Each of these methods used a single U-Net-based FCN to segment the following seven anatomical features: the head (not including the upper lip and hard palate), soft palate, lower lip and jaw (including the lower incisor space), tongue (not including the epiglottis), upper lip and hard palate (see Figure 22D). More specifically, the methods used FCNs based on the original U-Net and QuickTumourNet [215], the FCN developed by [216] and CEL-Unet [217].

In addition to the methods to segment rtMR images of speech described above, a method to segment 2D static MR images of the vocal tract has recently been developed [218]. In contrast to the real-time speech MR image segmentation methods described above, this method segmented sagittal images of the vocal tract as well as midsagittal images. However, similarly to several of the DL-based real-time speech MR image segmentation methods, the method used FCNs based on the original U-Net. The method segmented three anatomical features (the pharynx, tongue and soft palate) in the images using a three-stage approach. In each stage, a different U-Net-based FCN was used to segment one anatomical feature in the images. The pharynx, tongue and soft palate were segmented in the first, second and third stages respectively.

A key requirement for the development of DL-based segmentation methods is the availability of GT segmentations. While GT segmentations of articulators in rtMR images of speech have been created and used to develop methods to segment such images [205–207], these segmentations have not been made publicly available. As described in section 3.1.7, several real-time speech MRI datasets have been made publicly available and all these datasets include 2D midsagittal image series [18,19,116–121]. However, none of these datasets include articulator GT segmentations. The current lack of publicly available articulator GT segmentations is a barrier to the development of DL-based methods to segment (and ultimately analyse) real-time speech MR images, in addition to preventing rigorous comparison of segmentation methods.

3.3.5 LVP MR Image Segmentation and Analysis

As described in section 3.1.9, use of MRI to visualise the LVP is increasing. In addition, there is increasing interest in measuring aspects of the LVP in MR images [13,25,130–139,122,140–143,123–129]. In all previous work [13,25,130–139,122,140–143,123–129], measurements such as the length and thickness of the LVP were manually obtained from MR images. However, obtaining measurements in this way is time-consuming, requires input by specialists and is prone to intra- and inter-observer variability. To avoid the burden of manual measurements and to facilitate LVP measurement on a larger scale, there is currently an unmet need for automatic LVP measurement methods. A common approach for automating the measurement of anatomical features in biomedical images is to first segment the features and then perform measurements using the segmentations. As a first step towards developing an automatic LVP measurement method, in very recent work [17], four state-of-the-art DL-based methods were used to segment the LVP and five other anatomical features (adenoids, lateral pharyngeal wall, posterior pharyngeal wall, pterygoid raphe and soft palate) in 3D T_1-weighted MR images. More specifically, two methods based on 3D U-Net [177] (one of which was developed using nnU-Net [187]), the Swin UNETR method [185] and the 3D UX-Net method [219] were used. Evaluated using the DSC, the 3D UX-Net method was found to most accurately segment the LVP and three of the other anatomical features.

GT segmentations of the LVP have been created in previous work [17,144,146], however, these segmentations have not been made publicly available. While there are publicly available MRI datasets that include 3D images of the vocal tract [18,19,117,118,220,221], these datasets either do not include GT segmentations of anatomical features [18,19,117,118] or only include GT segmentations of the vocal tract [220,221]. The current lack of publicly available LVP GT segmentations is a barrier to the development of DL-based methods to segment (and ultimately quantify aspects of) the LVP in MR images, in addition to preventing rigorous comparison of segmentation methods.

3.4 Medical Image Registration

3.4.1 Introduction

Usually, corresponding regions in different images are not in spatial alignment. In other words, if the images were superposed, corresponding anatomical regions within them would not overlap. There are several reasons for this lack of alignment, ranging from differences in how the images were acquired to changes in the subject or object that was imaged. For example, the images may have been acquired using different modalities, from different views, or the subject or object being imaged may have moved or changed shape (e.g. images acquired during different scanning sessions or during organ motion such as that due to breathing or speech).

Image registration is the process of finding transformations that spatially align images. It is a key task in the field of medical image analysis and has a wide range of clinical applications, including radiotherapy [222] and neuroimaging [223]. Medical image registration has been an active area of research for over 30 years and a wide variety of methods have been developed for use in different scenarios. Several review articles give an overview of these methods [224–229].

Image registration is usually performed on a pair of images, to find a spatial transformation, φ, that describes how pixels or voxels in one of the images should move in space to align them with corresponding pixels or voxels in the other image. By convention, the image that will be transformed is referred to as the moving or source image, I_m, while the other image is referred to as the fixed or target image, I_f.

The transformation model (also known as a deformation model) of an Image registration method determines the range of possible transformations φ that can be applied to I_m. Rigid models allow translations and rotations, and affine models allow translations, rotations, shearing and scaling. These models can be compactly described using a single matrix. However, not all transformations can be described using translations, rotations, shearing and scaling. Instead, non-linear (also known as deformable) models are used to describe these more complex transformations. These models are usually defined by a dense field of vectors that describe either how each pixel or voxel in I_m should move in space to align it with the corresponding pixel or voxel in I_f or how each point in a grid of control points should move in space. Examples of the different types of transformations are shown in Figure 23.

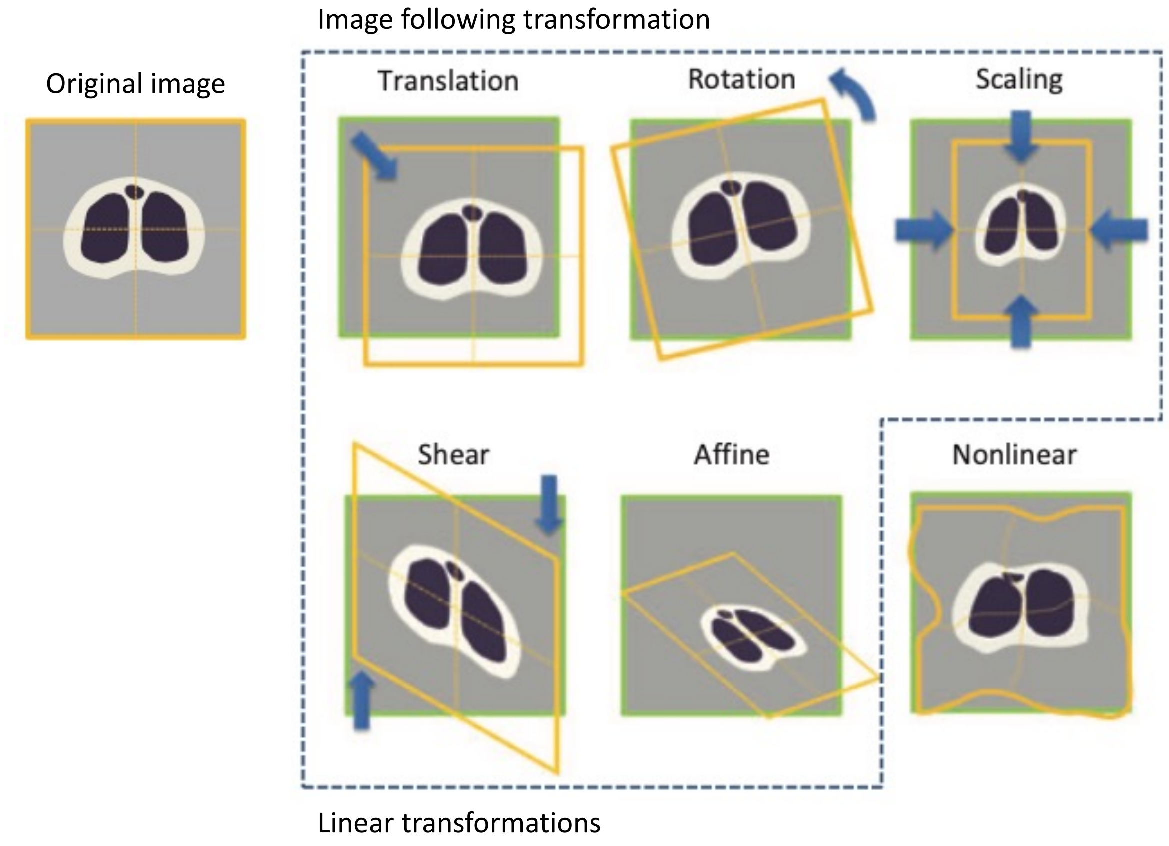

Figure 23: Examples of different types of transformations and their effect on an image (modified from [230]*).*

Finding transformations that align corresponding regions in medical images can be useful for three main reasons. First, they can enable the fusion of information contained in different images, which is useful for clinical applications such as image-guided interventions [231] and radiotherapy treatment planning [232]. Second, they can quantitatively describe differences or changes in the shapes of anatomical features in images, thus allowing quantitative analysis of shape variability within and between populations [233]. Third, they can quantitatively describe changes in the positions of anatomical features in images, thus allowing estimation of the motion of these features [233–235].

3.4.2 Traditional Registration Methods

Traditional registration methods (also known as classical registration methods) find an optimal spatial transformation, $\hat{\varphi}$, by solving an optimisation problem. In other words, they iteratively modify φ until it minimises the value of a cost function (also known as an objective function), $C(I_f, I_m, \varphi)$:

$$\hat{\varphi} = \underset{\varphi}{\operatorname{argmin}}\, C(I_f, I_m, \varphi) \tag{15}$$

Usually, $C(I_f, I_m, \varphi)$ consists of two terms: one to quantify the similarity between I_f and the moving image transformed according to φ, $I_m \circ \varphi$; and another to regularise φ:

$$C\big(I_f, I_m, \varphi\big) = M\big(I_f, I_m \circ \varphi\big) + R(\varphi) \tag{16}$$

$M\big(I_f, I_m \circ \varphi\big)$ is often referred to as the matching criterion or similarity metric. Commonly used matching criteria include cross-correlation (CC), mean squared error (MSE) and mutual information (MI) [224–227]. The choice of matching criterion usually depends on whether the images are mono- or multi-modal. MSE and CC are favoured in mono-modal image registration, as these criteria assume identity and linear mappings between the pixel or voxel intensities in the images. MI is favoured in multi-modal registration, as it is robust even when there are complex nonlinear mappings between the intensities.

The purpose of $R(\varphi)$ is to encourage φ to have certain desirable properties, usually being spatially smooth and continuous as these properties are often required for anatomically plausible deformation fields. A commonly used $R(\varphi)$ constrains the second derivatives of φ to encourage spatially smooth and continuous φ [236,237]:

$$R(\varphi) = \frac{1}{K}\sum_{k=1}^{K}\left[\left(\frac{\partial^2\varphi}{\partial x^2}\right)^2 + \left(\frac{\partial^2\varphi}{\partial y^2}\right)^2 + 2\left(\frac{\partial^2\varphi}{\partial xy}\right)^2\right] \tag{17}$$

where K is the number of pixels in the image.

Mathematically, image registration is challenging problem as there are many different spatial transformations that can align corresponding pixels or voxels in a pair of images. Regularising φ aims to make the problem easier to solve by penalising solutions that do not meet certain criteria, such as being smooth and continuous, as specified by the regularisation term.

Many different types of traditional registration methods have been developed and used to register a wide variety of medical images. Several review articles give an overview of these methods [224–227]. A popular rigid registration method is the block-matching method [238], while popular nonlinear registration methods include free-form deformations (FFDs) [236], demons [239] and their extensions such as diffeomorphic demons [240] and symmetric image normalisation (SyN) [223].

Implementations of several of these popular methods are publicly available. For example, NiftyReg [237,241] implements the block-matching method and FFDs, MATLAB (MathWorks, Natick, MA) implements demons and *ITK* [242] implements FFDs, SyN and demons. These implementations facilitate the adaptation and optimisation of the methods to new applications.

While a large number of mono- and multi-modal traditional registration methods have been developed and some of these translated into clinical practice [222], these methods register images in an iterative and therefore time-consuming way, preventing their use in clinical applications requiring near-real-time registration. This limitation has motivated the image registration community to explore alternative non-iterative ways to perform image registration.

3.4.3 Deep-Learning-Based Registration Methods

Recently, inspired by the successes of DL-based methods in other medical image analysis tasks, researchers have developed DL-based methods for medical image registration [228,229,251–254,243–250]. Two recent review articles give an overview of these methods [228,229]. The latest methods [243,244,253,254,245–252] are nonlinear registration methods that consist of CNNs (introduced in section 3.2.6) for estimating deformation fields between images and spatial transformers [255] for transforming images and/or segmentations according to the estimated deformation fields. These methods have achieved state-of-the-art accuracy in the registration of MR images of organs including the heart [243,244,247,251,253] and brain [245,246,248–250,252].

The latest DL-based nonlinear registration methods are unsupervised [243–246] or weakly-supervised [245–249,251–254] as, during their training, the deformation fields they estimate are not compared with GT deformation fields. The main motivation for avoiding the use of GT deformation fields in training is that these are rarely available and, if not impossible, the process to obtain these fields is time-consuming and prone to inaccuracies. While images with GT deformation fields can be synthesised, the key challenges of this approach are synthesising images with a realistic appearance and synthesising realistic deformation fields.

Unsupervised methods are trained using images only. The loss functions of the latest unsupervised methods are inspired by the cost functions of traditional registration methods and typically consist of two terms: the matching criterion, $M(I_f, I_m \circ \varphi)$ that quantifies the similarity between I_f and $I_m \circ \varphi$, and the regularisation term, $R(\varphi)$, which regularises the deformation field to ensure it has desirable properties such as being spatially smooth and continuous. The equation for a typical loss function is therefore:

$$L_{unsup} = M(I_f, I_m \circ \varphi) + \epsilon R(\theta) \tag{18}$$

where ϵ is a scalar constant. Similarly to traditional registration methods, commonly used matching criteria include CC [243–245,250,251] and MSE [245,247–249,252]. An overview of how an unsupervised registration method is trained is shown in Figure 24.

Weakly-supervised methods are trained using images and additional information such as surfaces [246] or segmentations [245,247–250,252–254]. The loss functions of these methods therefore typically consist of the two terms in equation (18) and an additional term. Usually, this additional term quantifies the overlap between corresponding regions in the segmentations or surfaces. Commonly used terms to quantify this overlap include the DSC [245,249,250,252–254] and cross entropy [247,248] (both introduced in section 3.3.2). An overview of how an unsupervised registration method is trained is shown in Figure 24.

Implementations of several state-of-the-art DL-based nonlinear registration methods are publicly available [245–249,251].

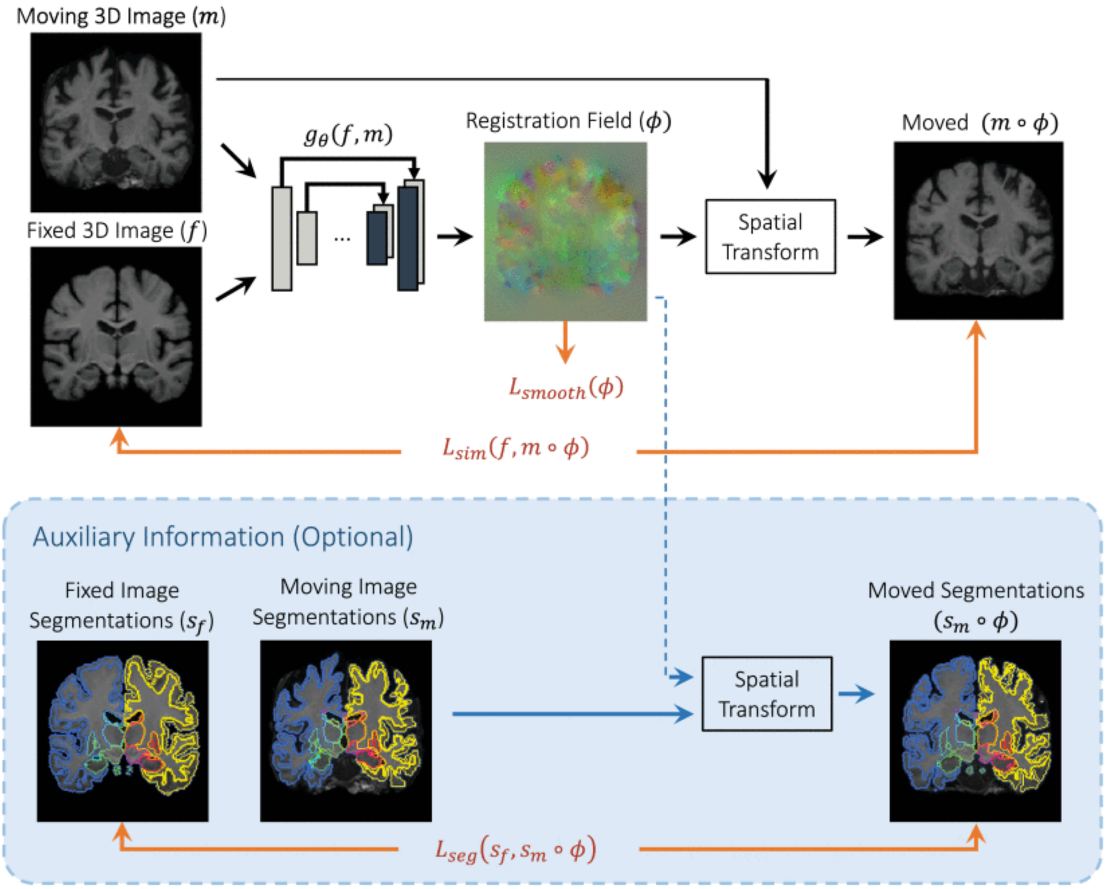
Moving 3D Image (m)
Fixed 3D Image (f)
$g_\theta(f, m)$
...
Registration Field (ϕ)
Spatial Transform
Moved ($m \circ \phi$)
$L_{smooth}(\phi)$
$L_{sim}(f, m \circ \phi)$
Auxiliary Information (Optional)
Fixed Image Segmentations (s_f)
Moving Image Segmentations (s_m)
Spatial Transform
Moved Segmentations ($s_m \circ \phi$)
$L_{seg}(s_f, s_m \circ \phi)$

Figure 24: An overview of VoxelMorph [245]*, a deep-learning-based nonlinear registration method that can be trained in either an unsupervised or weakly-supervised manner. The method consists of a convolutional neural network (g_θ) to estimate deformation fields that align images and a spatial transformer to warp images (and segmentations during weakly-supervised training) according to the deformation fields. Only images are used as inputs to the network. During unsupervised training of the network, the loss function consists of two terms: L_{sim} to quantify the similarity between the appearance of the moved and fixed images, and L_{smooth} to constrain the. During weakly-supervised training, the loss function includes an additional term, L_{seg}, that quantifies the overlap between the moved segmentations and the fixed image segmentations.*

3.4.4 Segmentation-Informed Registration Methods

Registration and segmentation can be related tasks, and there is increasing evidence that the performance of registration methods is improved if segmentation information is used in the registration process [245,247–254,256]. Such information is typically included in the training of DL-based nonlinear registration methods by adding region-overlap-based terms such as the DSC to their loss functions. Some methods, such as VoxelMorph [245] and joint registration and segmentation methods [247–250,252,257,258], only use segmentations during training, as shown in Figure 24, while others also use segmentations during deployment [251,253,254,256], as shown in Figure 26. VoxelMorph [245] has been used to register 3D MR images of the brain and performed this with an accuracy comparable to state-of-the-art traditional registration methods while reducing the computation time from hours to minutes on a central processing unit (CPU) and to under one second on a graphics processing unit (GPU) [245].

Compared with methods such as VoxelMorph, the methods that use segmentation during both training and deployment all include segmentation information in the registration process in one of two additional ways. The first approach is to use segmentations to modify the appearance of images in order to optimise them for the registration task [251,253,256]. In this approach, images are modified before being used as inputs to the registration CNN either by multiplying them by binary masks [251,253], as shown in Figure 25, or by using a fully convolutional image transformer network whose loss function includes a region-overlap-based term [256]. The second approach uses segmentations as well as images as inputs to the registration CNN [254], as shown in Figure 26. The rationale for using segmentations as inputs, even if these are estimates rather than ground truths, is that they provide information about the positions of anatomical features in the images and therefore help the registration CNN to estimate more accurate deformation fields. To enable their deployment when fixed and moving image segmentations are not available, the frameworks proposed in [251,254] all include automated segmentation methods.

Implementations of several segmentation-informed DL-based nonlinear registration methods are publicly available [245,247–249,251].

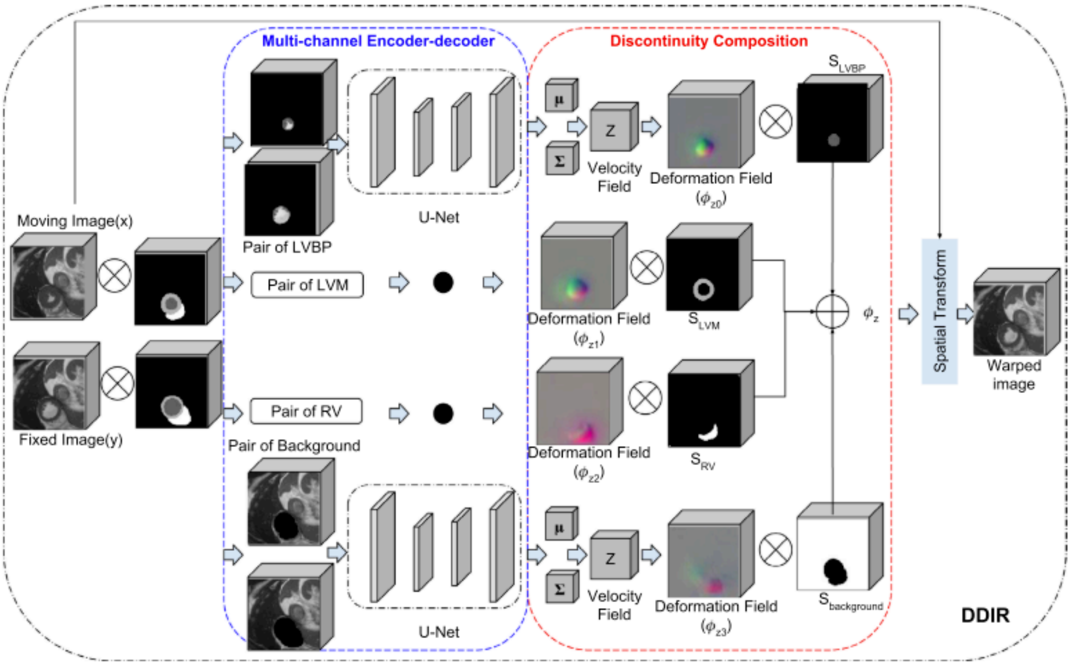

Figure 25: An overview of the deep-learning-based nonlinear registration method developed by [253]. *The method is both segmentation informed and discontinuity preserving. The method consists of the following steps. First, the input images are multiplied by binary masks to create multiple single-region versions of the images. These versions of the images are then used as inputs to U-Net-based FCNs to estimate region-specific velocity fields. Next, the velocity fields are converted into deformation fields, multiplied by binary masks to introduce discontinuities and then linearly combined to create an overall deformation field. Finally, a spatial transformer is used to warp the moving image according to the overall deformation field.*

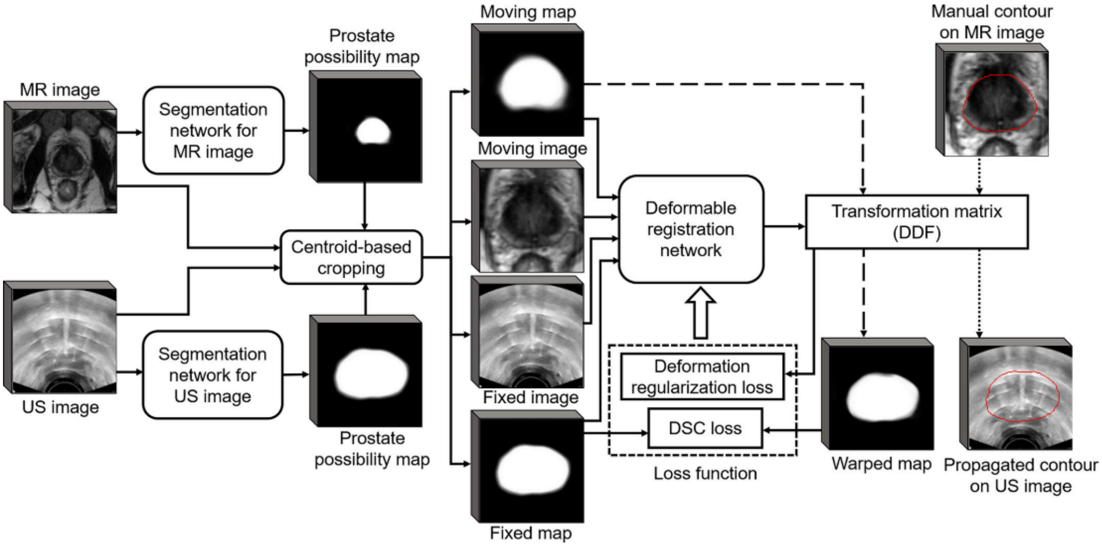
MR image
Segmentation network for MR image
Prostate possibility map
Centroid-based cropping
US image
Segmentation network for US image
Prostate possibility map
Moving map
Moving image
Fixed image
Fixed map
Deformable registration network
Deformation regularization loss
DSC loss
Loss function
Transformation matrix (DDF)
Manual contour on MR image
Warped map
Propagated contour on US image

Figure 26: An overview of the deep-learning-based nonlinear registration method developed by [254]. *The segmentation-informed method uses a convolutional neural network (CNN) to estimate deformation fields that align ultrasound (US) images to magnetic resonance (MR) images. Images and segmentations are used as inputs to the CNN during both training and deployment. Additional CNNs are required to estimate segmentations for the images. DSC: Dice coefficient, DDF: dense displacement field.*

3.4.5 Discontinuity-Preserving Registration Methods

Most nonlinear registration methods, both traditional and DL-based, feature regularisation terms that aim to ensure that the estimated deformation fields are smooth and continuous. However, such fields cannot accurately capture certain types of motion such as organs sliding past each other or organs coming into contact and then separating from each other. Instead, deformation fields with discontinuities are required to capture these types of motion. While several traditional methods [235,259–265] have been developed to capture the former type of motion (i.e. sliding motion), only one of these [261] can capture the latter type (i.e. changes in organ contact). Since during speech the articulators routinely come into contact and then separate from each other, this type of method would be particularly suitable for capturing their motion. The method, inspired by the extended finite element method [266], extended the FFD method by introducing an additional term to the FFD B-spline basis functions to enable them to estimate more realistic deformations at discontinuities. The method was segmentation-informed: it required information about the location of discontinuities in f to be provided in the form of binary masks. The method was used to register 4D computed tomography (CT) images of the lungs and liver achieved this more accurately than other state-of-the-art methods including those developed by [235,259,260,263,264]. However, unfortunately there is no publicly available implementation of the method developed by [261].

Two DL-based methods have been developed to estimate deformation fields with discontinuities [253,267]. The first method [267], consisting of a U-Net-based FCN to estimate deformation fields to align pairs of images, was trained in an unsupervised manner using a loss function that included a regularisation term to preserve discontinuities. However, this regularisation term was designed to capture sliding motion only. The other method [253], an overview of which is shown in Figure 25, used separate U-Net-based FCNs to estimate deformation fields for different regions of the input images, and then used segmentations to create discontinuities in these fields before combining them into an overall displacement field. It was designed to capture cardiac motion and its suitability for capturing motion where organs come into contact and then separate from each other has not yet been investigated. However, there is currently no publicly available implementation of the method.

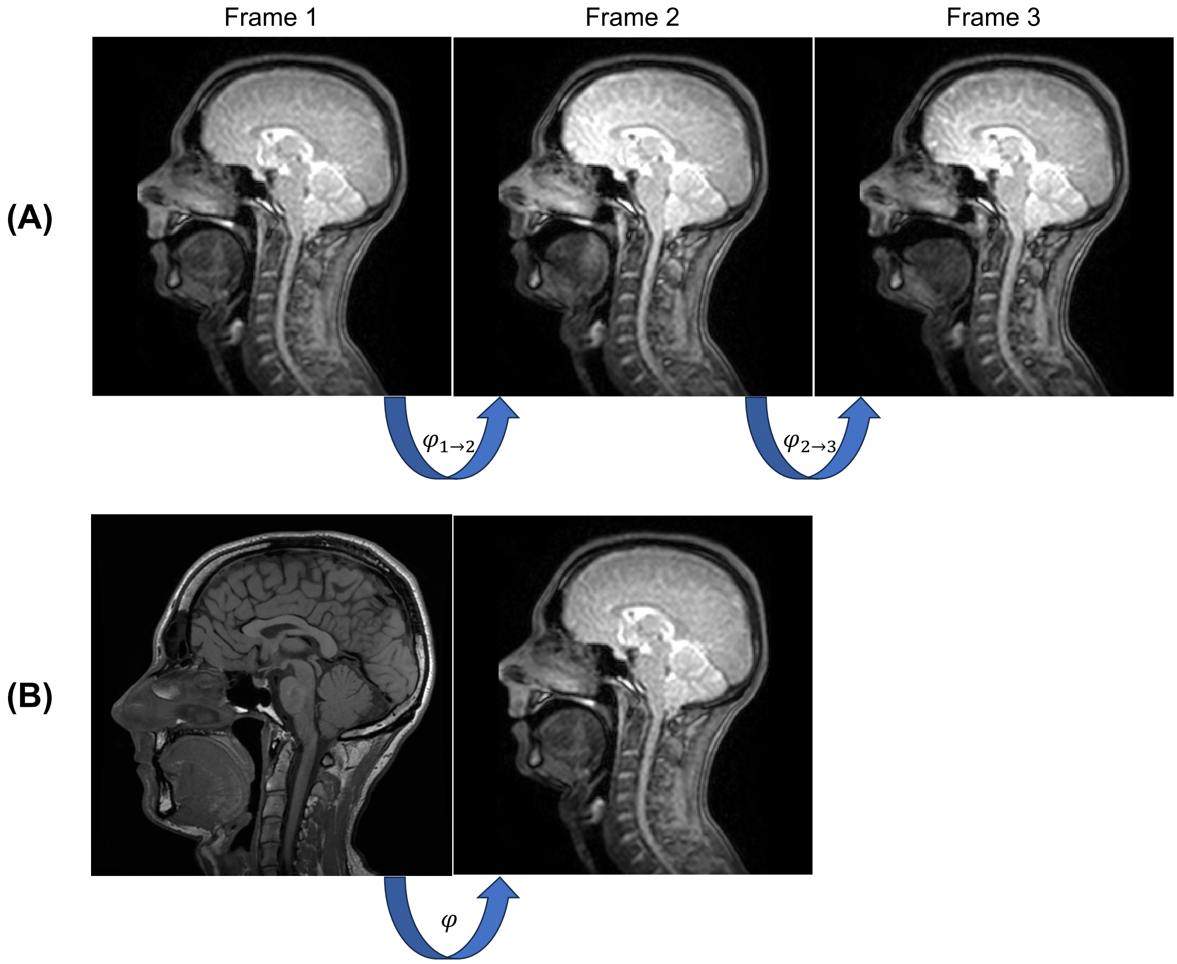

Figure 27: (A) Registration of consecutive frames in a series of two-dimensional real-time magnetic resonance images of speech. (B) Registration of the midsagittal slice of a three-dimensional image of the vocal tract to a two-dimensional real-time magnetic resonance image of speech. φ indicates the deformation field required to align the left-hand image to the right-hand image.

3.4.6 Registration of Magnetic Resonance Images of the Vocal Tract

So far, only traditional registration methods have been applied to MR images of the vocal tract [20,21,69,72,75,196,268–270]. Rigid methods were used to correct for changes in head position in series of 2D rtMR images of speech [69,72,75,196], while nonlinear methods were used to synthesise rtMR image series of speech [21,268–270], create dynamic 3D atlases of the vocal tract during speech [21] and estimate the speed at which the tongue tip moves during speech [20]. More specifically, in [20,21,268–270] nonlinear methods were used to determine transformations to align:

1. Consecutive images in series of 2D rtMR images of speech [20,268,269], as shown in Figure 27A;

2. Adjacent sagittal slices in 3D images of the vocal tract acquired during sustained phonation [268];
3. Two-dimensional rtMR images of speech from different series [21,270];
4. Midsagittal slices in 3D images of the vocal tract acquired during sustained phonation to 2D rtMR images of speech [268], as shown in Figure 27B.

Using these transformations, 3D rtMR image series of speech were synthesised [268,270], 2D rtMR image series of speech were synthesised from single 2D rtMR images of speech [269], dynamic 3D atlases of the vocal tract during speech were created [21] and tongue tip speeds were estimated [20].

More specifically, the registration-based method for estimating tongue tip speeds, an overview of which is shown in Figure 28, consisted of the following steps. First, the nonlinear registration method described in [271] was used to estimate deformation fields between consecutive frames in series of 2D rtMR images of speech. Then, a point on the tip of the tongue was manually selected in the first image of the series. Next, the position of the point in all the other images was estimated using the deformation fields, thus enabling tracking of the trajectory of the tongue tip and the calculation of tongue tip speeds. Tongue tip speeds estimated using the method were found to be similar to those reported in the literature, suggesting that registration-based methods can accurately estimate the speed at which articulators move during speech.

In terms of the three previous works, the diffeomorphic demons method [240] was used to register adjacent sagittal slices in 3D images of the vocal tract [268], consecutive frames in 2D rtMR image series of speech [268–270] (see Figure 27A), and midsagittal slices in 3D images to 2D rtMR images of speech [268] (see Figure 27B). In another previous work, the FFD method [236] was used to register consecutive frames in series of 2D rtMR images of speech [21].

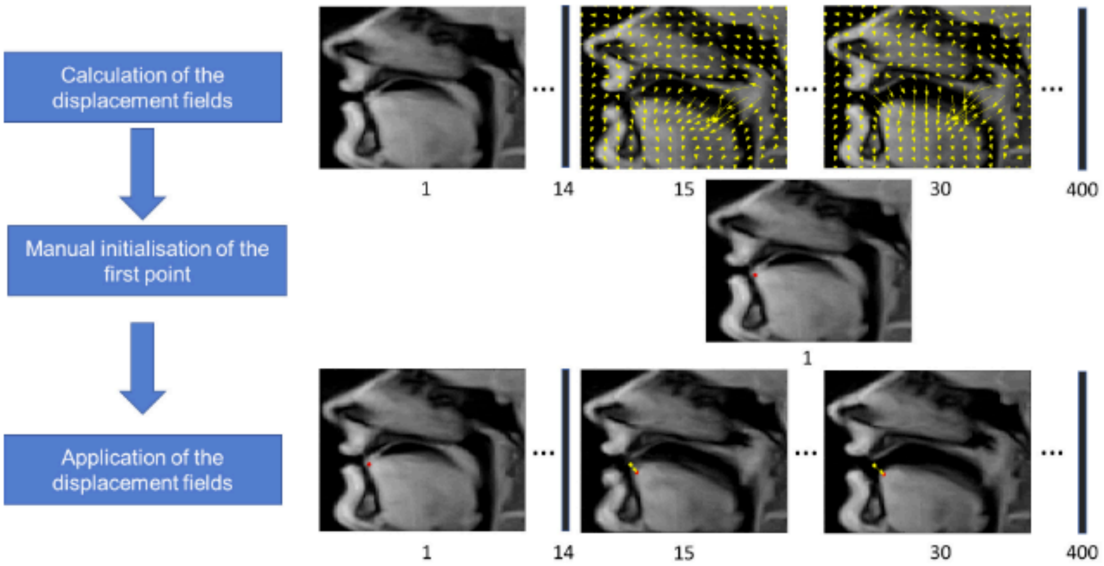
Calculation of the displacement fields
Manual initialisation of the first point
Application of the displacement fields
1
14
15
30
400
1
1
14
15
30
400

Figure 28: An overview of the registration-based method proposed by [20] *to estimate tongue tip speeds in series of two-dimensional magnetic resonance images of speech.*

In three previous works [20,268,269], images where articulators were in contact were nonlinearly registered to images where they were not and vice versa. However, the authors did not evaluate if their chosen registration methods captured these changes in contact. As explained in section 2.2.3, changes in contact such as those that occur because of velopharyngeal closure are clinically In [268], the authors reported that the diffeomorphic demons method did not capture articulators coming into contact (for example, the lips coming into contact). Nevertheless, the authors used the same method in similar subsequent work [269]. In [20,269], the authors did not discuss if their chosen registration methods captured changes in articulator contact. As described in section 2.2.3, in clinical speech assessment visualisation of soft palate motion provides information that aids VPI treatment decision making. Consequently, a key requirement of motion estimation methods intended for use in clinical speech assessment is that they accurately capture soft palate motion, including any velopharyngeal closures that occur.

3.5 Conclusions

Use of MRI to visualise the vocal tract and articulators during speech is increasing. Real-time MRI is the most suitable type of dynamic MRI technique for use in clinical speech assessment, as it allows imaging of speech as it occurs and does not require any repetitions of a speech task. However, a key requirement to facilitate the widespread adoption of rtMRI techniques in clinical speech assessment is that the techniques should only require standard MRI equipment and software. This requirement motivated the choice of the dataset (described in section 4.1) that was used in the work presented in chapter 5 and chapter 6.

Use of MRI to visualise the LVP is also increasing. In previous work, T_2-weighted 3D images of the LVP at 3.0 T using TSE pulse sequences were predominantly acquired. However, there is still no consensus on the optimal contrast for LVP visualisation in MR images. This lack of consensus motivated the optimal contrast investigation presented in section 4.2.2.

There is increasing interest in extracting quantitative information about articulators and the LVP from MR images. A common approach to automate the quantification of anatomical features in medical images is to first segment the features and then perform quantification using the segmentations. While many methods to segment air-tissue boundaries between the vocal tract and articulators in 2D rtMR images of speech have been developed, few methods have been developed to fully segment articulators, an important first step to enable their quantification. In addition, there is only a single report in the literature of methods to segment the LVP in 3D MR images. This lack of methods motivated the work presented in chapter 5 and chapter 7.

In addition, there is increasing interest in extracting quantitative information about the motion of the soft palate in series of 2D rtMR images of speech. An established way to quantify motion in image series is by using nonlinear registration methods to estimate displacement fields between the images. While a nonlinear-registration-based method has been developed to estimate tongue tip speed in series of 2D MR images of speech, there are no reports in the literature of nonlinear-registration-based methods for estimating the motion of other articulators such as the soft palate. This lack of methods motivated the work presented in chapter 6.

Chapter 4: Materials

This chapter describes the datasets that were used in the work presented in this thesis. The first section describes the real-time speech MRI datasets and corresponding GT segmentations that were used in the work described in chapter 5 and chapter 6, while the second section describes the 3D static MR images of the vocal tract and corresponding GT segmentations that were used in the work described in chapter 7.

4.1 Real-Time Speech MRI Datasets

4.1.1 Introduction

As explained in section 3.1.7, use of MRI to visualise the vocal tract and articulators during speech is increasing due to the growing availability of MRI scanners, the development of rtMRI techniques for such visualisation, and the unique ability of MRI to non-invasively acquire images of any orientation without using ionising radiation [12,27,61]. Real-time MRI is the most suitable type of MRI technique for use in clinical speech assessment, as it allows imaging of speech as it occurs and does not require any repetitions of a speech task.

Typically during rtMRI of speech, series of 2D images of a midsagittal slice of the vocal tract are acquired. To accurately capture articulator motion, imaging at relatively high spatio-temporal resolutions is required [12]. State-of-the-art real-time speech MRI techniques enable 2D imaging of a single slice at a spatial resolution of <2.4×2.4 mm^2 and a temporal resolution of <20 ms [9,10]. However, these techniques require highly specialised MRI equipment and software, namely custom receive coils [10] and/or specialised pulse sequences and reconstruction methods [9,10], that are not widely available especially in clinical practice. These requirements therefore prevent the widespread adoption of the techniques, a limitation that has motivated the development of techniques that only require widely available standard MRI equipment and software [11,27,114,115]. Techniques that only require standard MRI equipment and software enable 2D imaging at spatial resolutions of <2.4×2.4 mm^2 and temporal resolutions <100ms. While these spatio-temporal resolutions are lower than those of state-of-the-art techniques, they are nevertheless sufficient to capture the general motion of articulators such as the soft palate [12]. A key requirement to facilitate the widespread adoption of rtMRI techniques in clinical speech assessment is that the techniques should only require standard MRI equipment and software. This requirement motivated the choice of the dataset described in this section.

As explained in section 3.3.3, there is increasing interest in extracting quantitative information about articulators and the vocal tract from 2D rtMR images of speech [7,14,75,76,84,88,195–198,20,68–74]. More specifically, there is interest in measuring the size and shape of the vocal tract [14,68,197,198,70,74–76,84,88,195,196], the size, shape and motion of the soft palate [72,73,75,84,198], lip motion [69,75,84], tongue motion [20,84] and the distance between the soft palate and the posterior pharyngeal wall [7,72,198]. Manual measurement to obtain this information is time-consuming, requires input by specialists and is prone to intra- and inter-observer variability. The increasing interest in extracting quantitative information, in combination with the need to avoid manual measurement, have motivated the development of a range of methods to (semi-)automatically extract this information [15,16,206,207,20,199–205]. Almost all these methods are segmentation based [15,16,207,199–206].

Segmentation of medical images is a common first step to enable automatic measurement of anatomical structures. As explained in section 3.3.4, numerous methods based on a variety of approaches have been developed to segment rtMR images of speech [15,16,207,199–206]. The majority of these methods [15,199–204,208] were designed to enable (semi-)automatic analysis of the size and shape of the vocal tract in the images, an analysis of particular interest to the speech science community. To enable this analysis, the methods created contours of air-tissue boundaries between the vocal tract and adjacent articulators. While such contours enable analysis of the size and shape of the vocal tract, they only partially segment articulators and consequently do not enable analysis of the size, shape, motion or position of the articulators during speech. Increasing interest in such analysis, by clinicians as well as speech science researchers, has recently motivated the development of methods to fully segment articulators in rtMR images of speech [205–207], including the method presented in chapter 5.

Development and evaluation of segmentation methods requires datasets with corresponding GT segmentations. As explained in section 3.3.4, while there are publicly available speech MRI datasets [18,19,116–121], none of these include GT segmentations of articulators. This requirement, in combination with the lack of suitable publicly available MRI datasets, motivated the creation of the GT segmentations described in this section.

As described in section 2.2.3, visualisation of soft palate motion provides information that aids VPI treatment decision making in clinical speech assessment. Consequently, a key requirement of automatic image quantification methods intended for use in clinical speech assessment is that they accurately capture soft palate motion. In particular, the methods must capture any velopharyngeal closures that occur. To enable evaluation of the accuracy with which methods captured velopharyngeal closures, GT velopharyngeal closure labels were created for the datasets described in this section.

Table 1: Imaging parameters used to acquire the two-dimensional real-time magnetic resonance image series. The table lists repetition times (TRs), echo times (TEs), sensitivity encoding (SENSE) factors, number of signal averages (NSAs), water fat shifts (WFSs) and bandwidths (BWs).

Parameter	Value
TR (ms)	2.0
TE (ms)	0.9
Flip angle (°)	15
Field of view (mm^2)	300×230
SENSE factor	2
NSA	1
Actual WFS (pixel) / BW (Hz)	0.134 / 3240.4

4.1.2 Real-time MR Images of Speech

Five series of rtMR images of speech acquired in a previous study [272] were used in the work described in chapter 5 and 6 of this thesis. The series were of five healthy adult volunteers (two females, three males; age range 24-28 years). All volunteers were fluent English speakers with no history of speech and language disorders.

Each volunteer was imaged in a supine position using a 3.0 T TX Achieva MRI scanner and a 16-channel neurovascular coil (both Philips Healthcare, Best, Netherlands) while they performed the following speech task a single time: counting from 1 to 10 in English. Images of a 10 mm thick mid-sagittal slice of the head were acquired using a steady-state free precession (SSFP) pulse sequence based on the sequence identified by [11] as being optimal for vocal tract image quality. Example images are shown in Figure 29A. Imaging parameters are listed in Table 1. The acquired matrix size and in-plane pixel size were 120×93 and 2.5×2.45 mm^2 respectively. However, *k*-space data were zero padded to a matrix size of 256×256 by the scanner before being reconstructed, resulting in a reconstructed in-plane pixel size of 1.17×1.17 mm^2. To maximise the signal-to-noise ratio in the images, partial Fourier was not used. Images were acquired at a temporal resolution of 0.1 s and only one image series was acquired per volunteer. The volunteers were instructed to perform the speech task at a rate which they considered to be normal. Some performed the task faster than others and consequently not all series had the same number of images. The series had 105, 71, 71, 78 and 67 images each (392 images in total).

4.1.3 Velopharyngeal Closure Identification

The number of velopharyngeal closures shown in the rtMR image series had not been identified in any previous work. To identify this number, the following steps were taken:

1. Each image was visually inspected and labelled as either showing contact between the soft palate and posterior pharyngeal wall or not showing contact.
2. Line charts of the labels of each image series were created (an example chart is shown in Figure 29E) and visually inspected to determine the number of velopharyngeal closures shown in the series.

It can be challenging to determine if an image shows contact between the soft palate and posterior pharyngeal wall, especially if the soft palate is close to the posterior pharyngeal wall. To reduce the subjectivity of the labels, each image was independently labelled by four MRI Physicists. Raters one to four respectively had four, ten, two and one years of experience of rtMRI of speech. All the images were labelled again one month later by rater one (the author of this thesis). Intra- and inter-rater agreement was assessed by comparing the labels and velopharyngeal closures. In cases where one rater disagreed with the others, the majority label was considered to be the GT label. In cases where only two raters agreed, raters one and two (those with the most experience of speech MRI) jointly inspected the images and then reached a consensus on the labels for these images, similarly to how SLTs jointly inspect videofluoroscopy speech image series in clinical practice in the UK. Line charts of the GT labels of each image series were created (an example chart is shown in Figure 29E) and visually inspected to determine the number of velopharyngeal closures shown in the series.

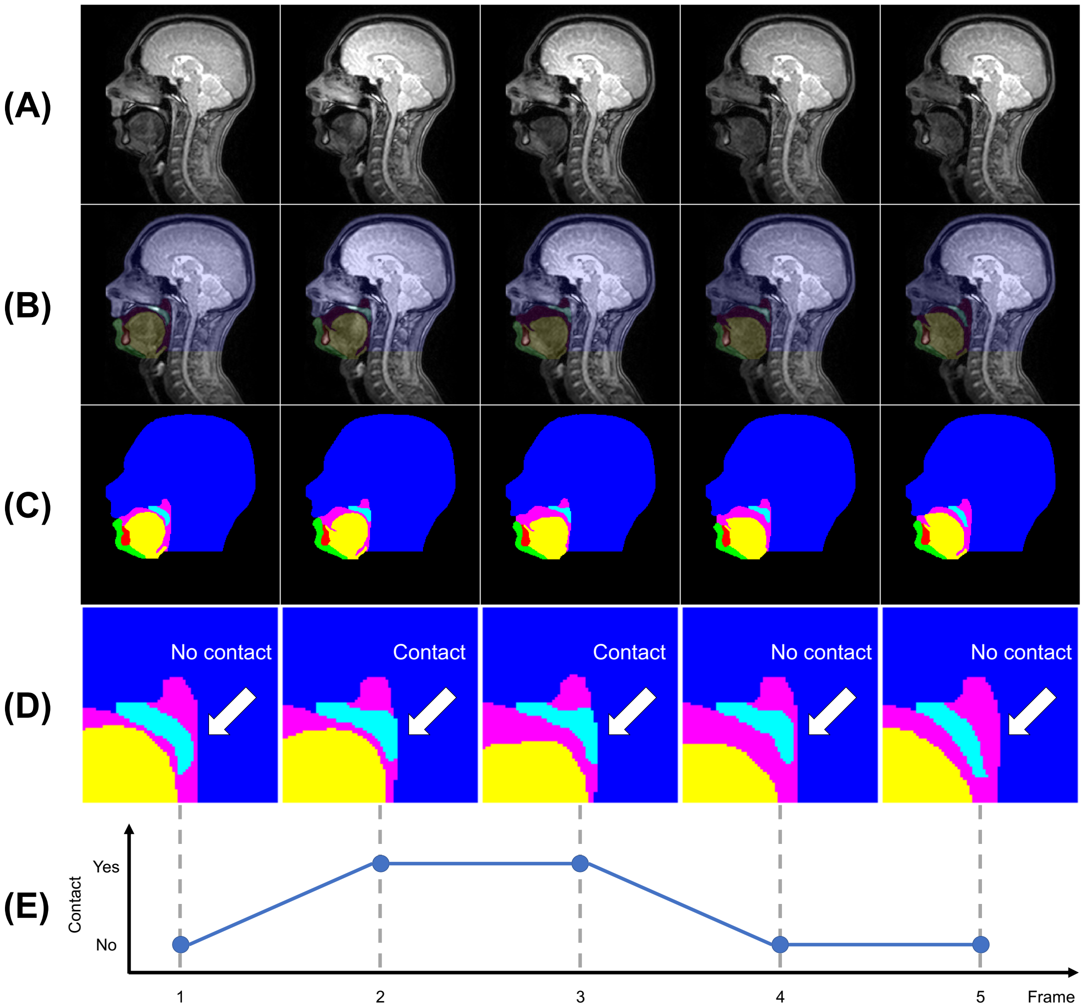

Figure 29: Five consecutive images from one of the real-time magnetic resonance image series (A) with ground-truth segmentations of anatomical features overlaid (B). The ground-truth segmentations are of the head (dark blue), soft palate (light blue), jaw (green), tongue (yellow), vocal tract (pink) and tooth space (red) classes. (C) shows ground-truth segmentations only. (D) shows cropped versions of the ground-truth segmentations in (C) with labels indicating if the soft palate is in contact with the posterior pharyngeal wall. (E) is a line chart of the contact labels.

4.1.4 Ground-Truth Segmentation Creation

GT segmentations of anatomical features in the rtMR image series had not been created in any previous work. GT segmentations were created by manually labelling pixels in each of the images. The segmentations consisted of six classes, each made up of one or more anatomical features. There was no overlap between classes: a pixel could not belong to more than one class. For conciseness, the classes were named as follows: head, soft palate, jaw, tongue, vocal tract and tooth space. However, the names of the head, jaw and tongue classes are simplifications. The head class consisted of all anatomical features superior to or posterior to the vocal tract. It therefore included the upper lip, hard palate, brain, skull, posterior pharyngeal wall and neck. The jaw class consisted of the lower lips, the soft tissue anterior to and inferior to the tooth space and the soft tissue inferior to the tongue. The tongue class included the epiglottis and the hyoid bone. Pixels not labelled as belonging to one of the classes were considered to belong to the background. Example GT segmentations are shown in Figure 29B. The reasons for including each class in the GT segmentations are given in Table 2.

Table 2: Reasons for including each class in the ground-truth segmentations of the real-time magnetic resonance images of the vocal tract during speech.

Class	**Reason(s) for inclusion**
Head	*Primary:* segmentation of the posterior pharyngeal wall would enable automatic measurement of the distance between the soft palate and the posterior pharyngeal wall *Secondary:* segmentation of the upper lip would enable automatic lip motion tracking
Soft palate	Segmentation would enable soft palate size, shape, motion and position analysis, and also automatic measurement of the distance between the soft palate and the posterior pharyngeal wall
Jaw	Segmentation of the lower lip would enable automatic lip motion tracking
Tongue	Segmentation would enable tongue size, shape, motion and position analysis
Vocal tract	Segmentation would enable vocal tract size and shape analysis
Tooth space	Included so that there were no holes in the ground-truth segmentations, thus facilitating the post-processing of estimated segmentations

Wherever possible, the boundaries of the classes were chosen to be clear anatomical boundaries in order to minimise the subjectivity of the GT segmentations. Apart from the tooth space class, the majority of the class boundaries were easily distinguishable air-tissue boundaries. However, there were no clear anatomical boundaries for some sections of the class boundaries. Instead, the following artificial boundaries were devised for these sections. The two main goals when devising these boundaries were firstly to include only relevant anatomical features and secondly to minimise the subjectivity of the boundaries.

The inferior boundary of the head class in the neck was defined as the horizontal line parallel to the inferior surface of the intervertebral disc between vertebrae C3 and C4 (see blue arrows in Figure 30). This choice was made to exclude the inferior section of the neck in the head class as this section was not required for the desired analyses and would have otherwise increased the imbalance between the number of pixels in the head class and the other classes.

The posterior boundary of the jaw class was defined as the anterior edge of the hyoid bone (see dotted green arrows in Figure 30), while the inferior boundary of the jaw class in the neck was defined as the horizontal line intersecting the point where the jaw meets the neck (see solid green arrows in Figure 30).

The inferior boundary of the vocal tract class was defined in the same way as that of the head class (see pink arrows in Figure 30), and the inferior boundary of the tongue class in the neck was defined in the same way as that of the jaw class in the neck (see yellow arrows in Figure 30).

GT segmentations were created by the MRI Physicist with four years of speech MRI experience (the author of this thesis), using bespoke software developed in house and implemented in MATLAB R2019b (MathWorks, Natick, MA). GT segmentations were consistent with the GT velopharyngeal closure label for the images: segmentations of the soft palate and posterior pharyngeal wall (part of the head class) were in contact for images labelled as showing contact and not in contact otherwise. To enable investigation of intra-rater agreement and therefore uncertainty in the segmentations, the Physicist created GT segmentations again for seven (approximately 10%) randomly chosen images in each series. The agreement was quantified using two metrics: the DSC and the HD.

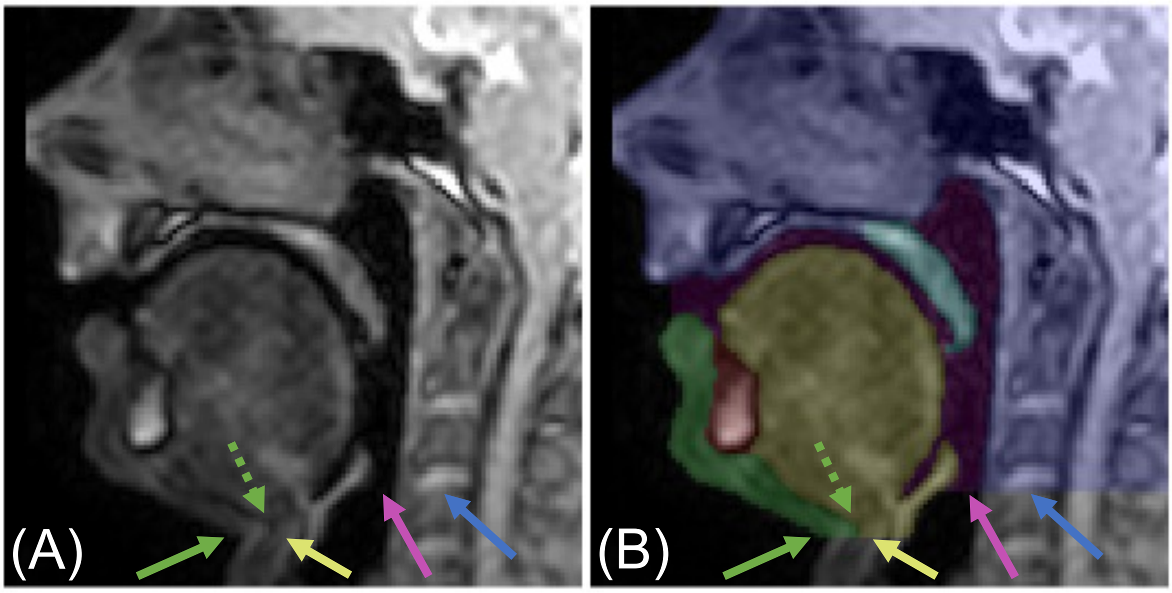

Figure 30: A real-time magnetic resonance image of speech cropped to only show the vocal tract (A) with ground-truth segmentations of anatomical features overlaid (B). The blue arrows point to the inferior surface of the intervertebral disc between vertebrae C3 and C4. The dotted green arrows point to the anterior edge of the hyoid bone, while the solid green arrows point to where the neck meets the jaw. The yellow arrows point to the inferior boundary of the tongue class in the neck, while the pink arrows point to the inferior boundary of the vocal tract class.

The process for creating the GT segmentations of an image series was as follows:

1. **Initial binary mask creation:** a series of binary masks of the entire head were created by applying a manually chosen threshold to the image series (see Figure 31). The chosen threshold was the minimum integer that resulted in as many of the binary masks as possible meeting the following criteria:
    a. Minimal noise in the vocal tract (see Figure 32A).
    b. Clear air-tissue boundaries.
    c. Jaw not divided into two or more regions (see Figure 32B).
    d. Tip of epiglottis not artificially in contact with tongue (see Figure 32C).

   No single threshold resulted in all the binary masks meeting all the criteria. The following iterative process was used to identify a suitable threshold:
    a. A threshold was applied to the image series to create a series of binary masks of the entire head.
    b. The series was visually inspected.
    c. If necessary, the threshold was modified and steps (a) and (b) above were repeated.

Once a suitable threshold has been identified, holes in the binary mask were manually removed (see Figure 31).

2. **Head class GT segmentation creation:** a series binary masks of the head class were created by:
    a. Manually defining an approximate outline of the head class in each image (see Figure 33A).
    b. Extracting the sections of the initial binary mask within the approximate outline (see Figure 33B).
    c. Manually refining the extracted binary masks (see Figure 33C).
3. **Soft palate class GT segmentation creation:** a series of binary masks of the soft palate class were created by following the same process as in step 2 above.
4. **Jaw class GT segmentation creation:** a series of binary masks of the jaw class were created by following the same process as in step 2 above.
5. **Tongue class GT segmentation creation:** a series of binary masks of the tongue class were created by following the same process as in step 2 above.
6. **Tooth space class GT segmentation creation:** a series of binary masks of the tooth space class were created using the binary masks of the jaw and tongue classes, as shown in Figure 34.
7. **Vocal tract class GT segmentation creation:** a series of binary masks of the vocal tract class were created using the binary masks of the head, soft palate, jaw, tongue and tooth space classes, as shown in Figure 35.

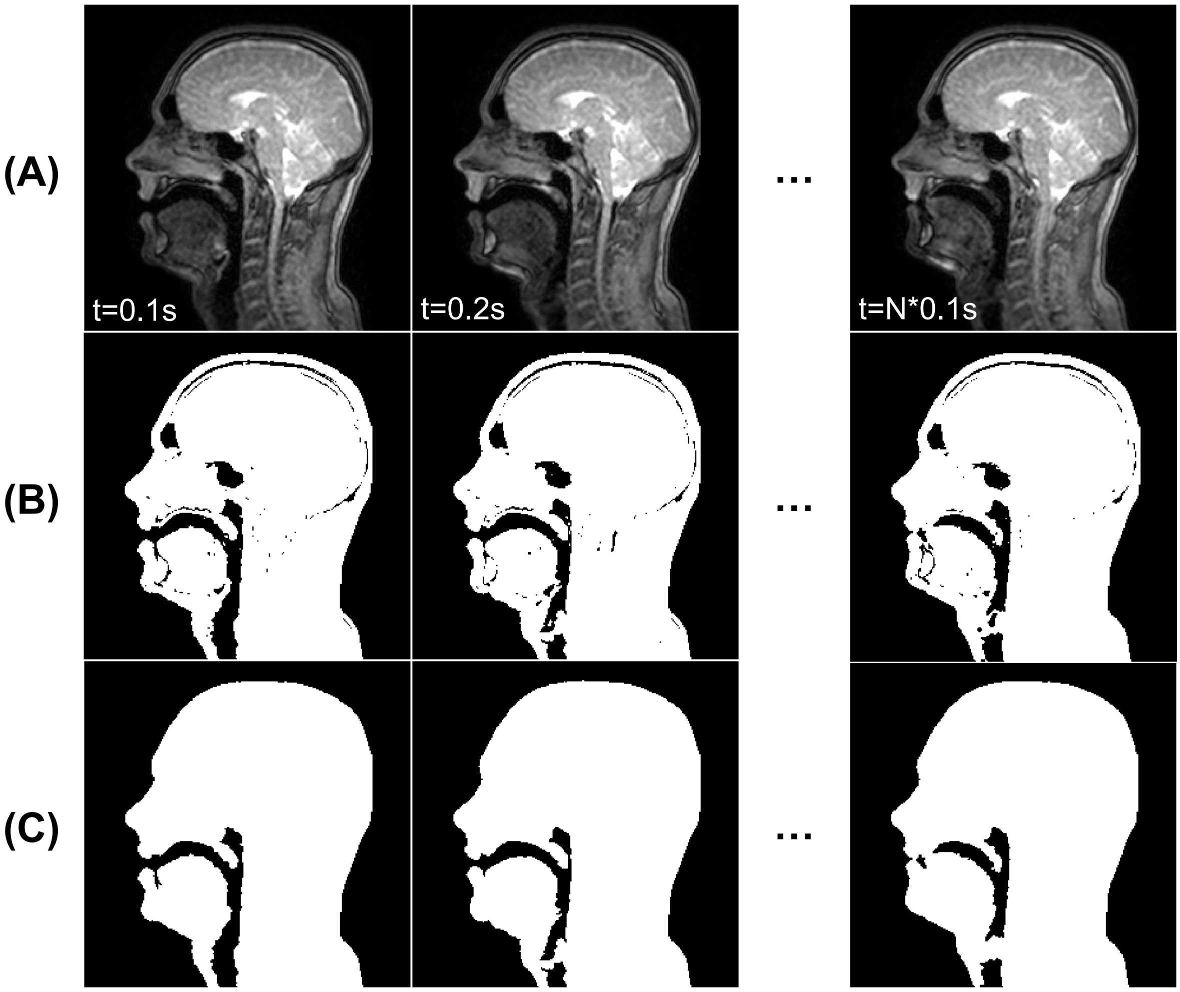

Figure 31: A series of real-time magnetic resonance images of speech (A), corresponding binary masks of the entire head created by applying a manually chosen threshold (B), and the binary masks after holes in them have been filled (C).

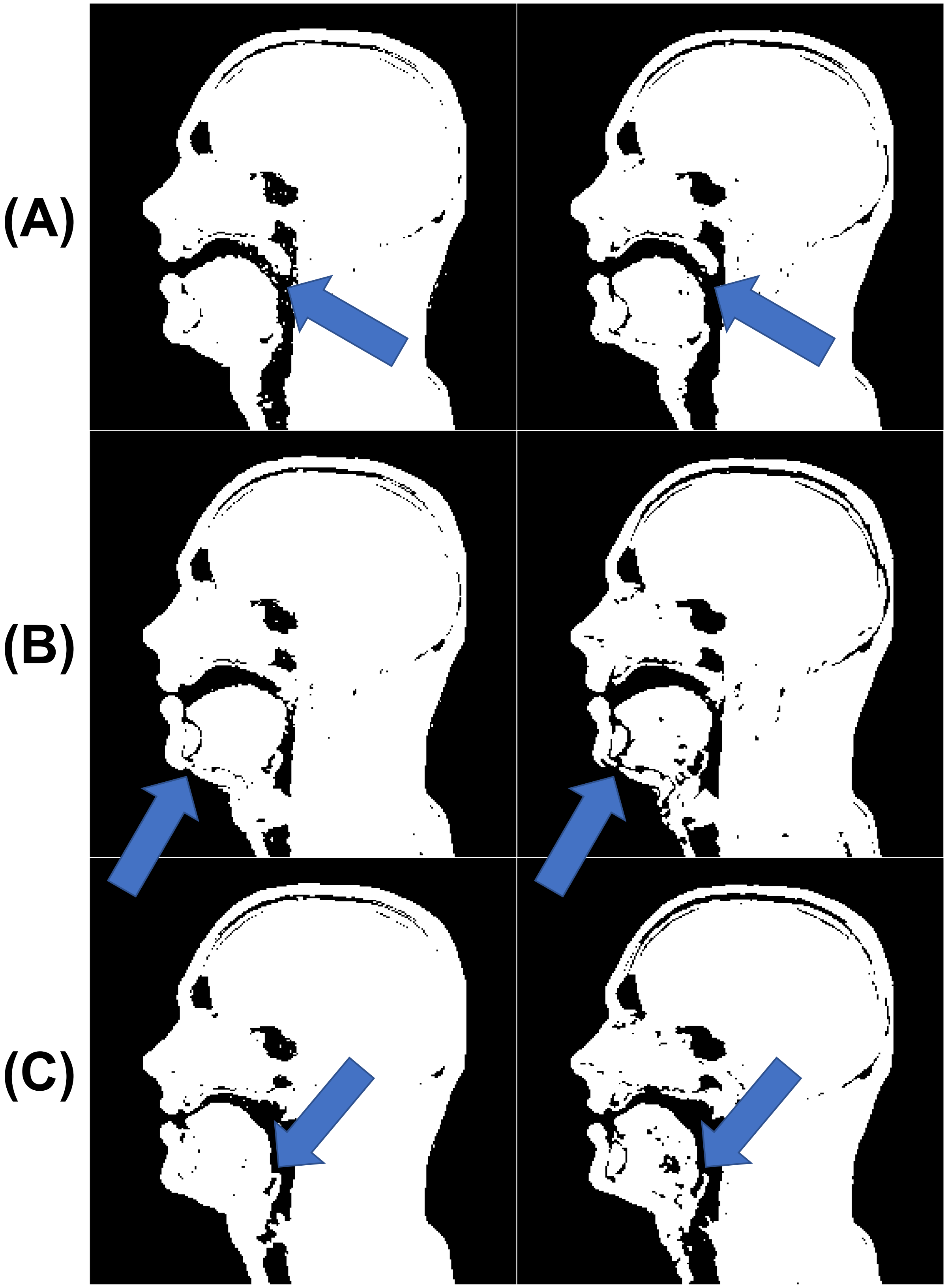

Figure 32: Pairs of binary masks of the entire head created from the same images but using different thresholds, one suitable and the other unsuitable. In row (A), the threshold used to create the left-hand mask is too low, resulting in noise in the vocal tract (indicated by blue arrow). In row (B), the threshold used to create the right-hand mask is too high, resulting in the jaw being divided into two regions. In row (C), the threshold used to create the left-hand mask is too low, resulting in the tip of the epiglottis being artificially in contact with the tongue.

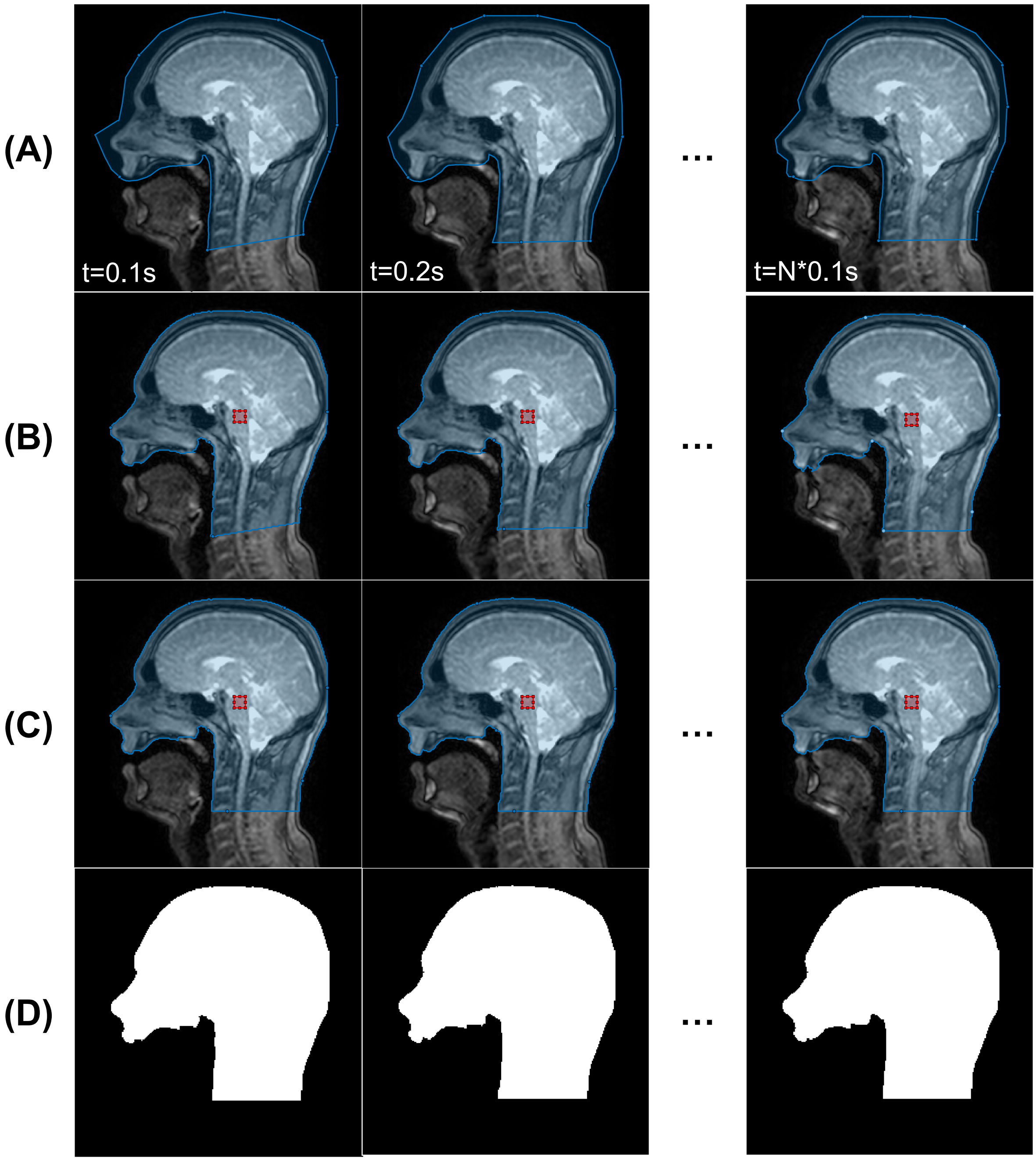

Figure 33: A series of real-time magnetic resonance images of speech with an approximate manually drawn outline of the head class overlaid in blue (A), with the section of the entire head binary mask (see Figure 31B) contained in the approximate outline overlaid (B), with the manually refined version of the binary mask in (B) overlaid (C). (D) shows the binary mask in (C) only.

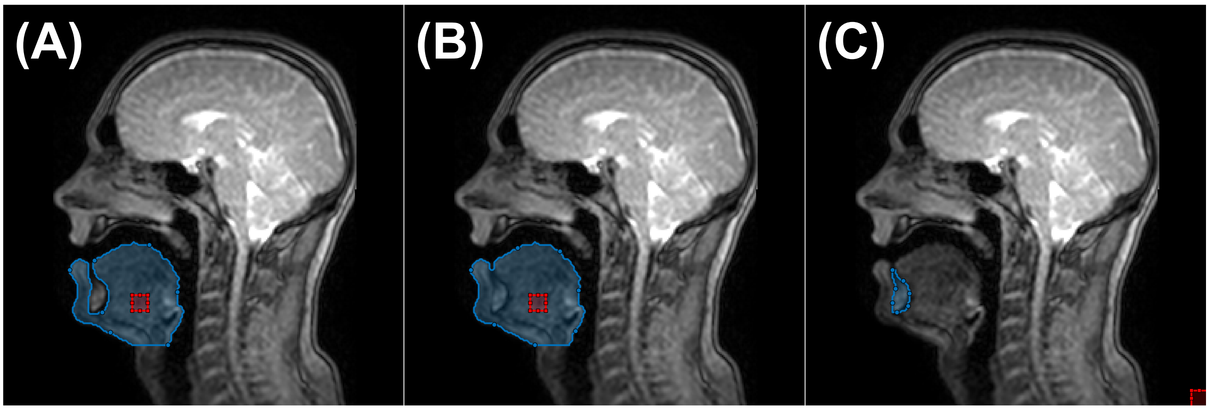

Figure 34: The process to create a binary mask of the tooth space class. (A) A real-time magnetic resonance image with a jaw and tongue class binary masks overlaid in blue. (B)The same binary mask except with the tooth space region manually filled. (C) The same image with the tooth space class binary mask overlaid. The binary mask in (C) was created by subtracting the binary mask in (A) from the binary mask in (B).

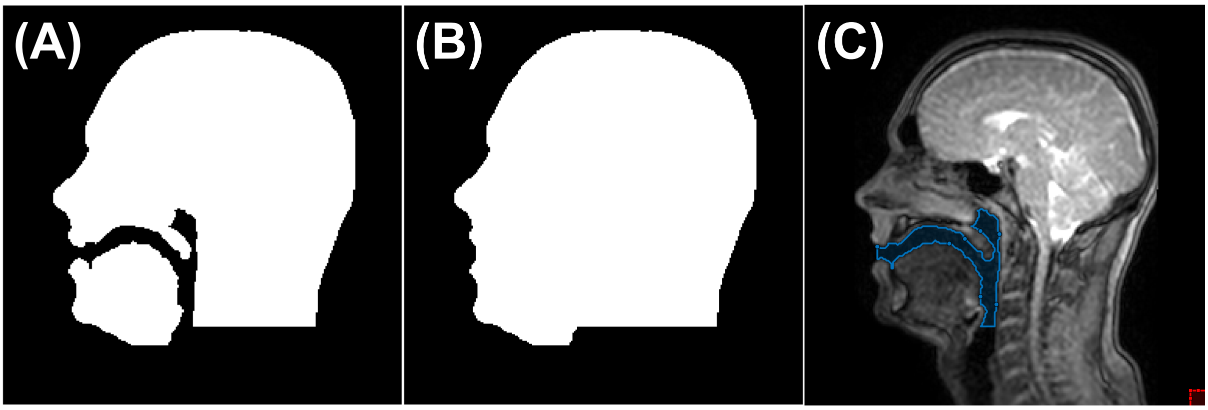

Figure 35: The process to create a binary mask of the vocal tract class. (A) A binary mask of the head, soft palate, jaw, tongue and tooth space classes combined. (B) The same binary mask except with the vocal tract region manually added to it. (C) A real-time magnetic resonance image with the vocal tract binary mask overlaid in blue. The binary mask in (C) was created by subtracting the binary mask in (A) from the binary mask in (B).

4.1.5 Results

4.1.5.1 Velopharyngeal Closure Identification

The GT labels of each image series are shown in Figure 36. Of the 392 images, 230 (58.7%) images were labelled as showing contact between the soft palate and posterior pharyngeal wall, while 162 (41.3%) were labelled as not showing contact. As shown in Figure 37, in three image series two thirds of the images were labelled as showing contact, while in the other two image series approximately half of the images were labelled as showing contact.

The GT numbers of velopharyngeal closures shown in the image series are listed in Table 3. In total, 30 velopharyngeal closures were shown in the image series.

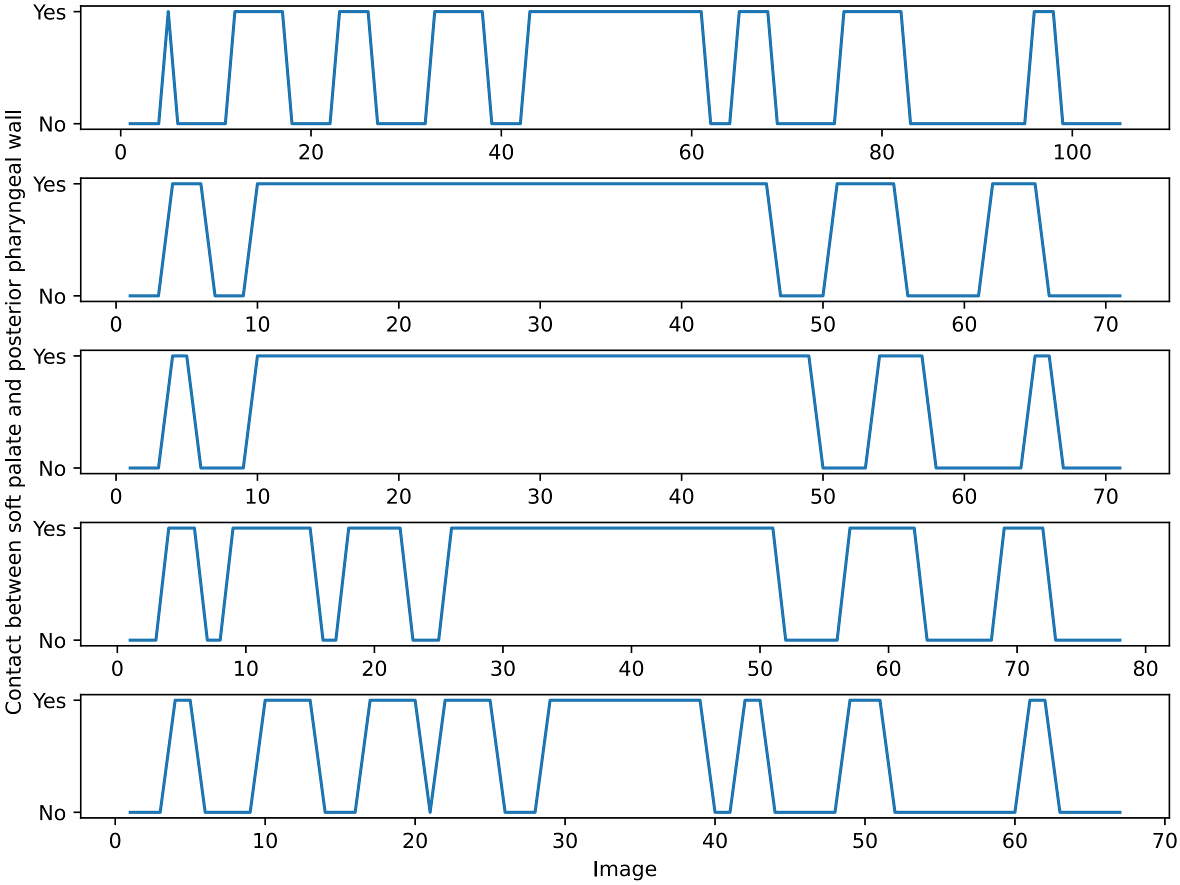

Figure 36: The ground-truth labels of the five real-time magnetic resonance image series. Each line chart represents a different series and has different x-axes. Each peak in a line chart indicates a velopharyngeal closure.

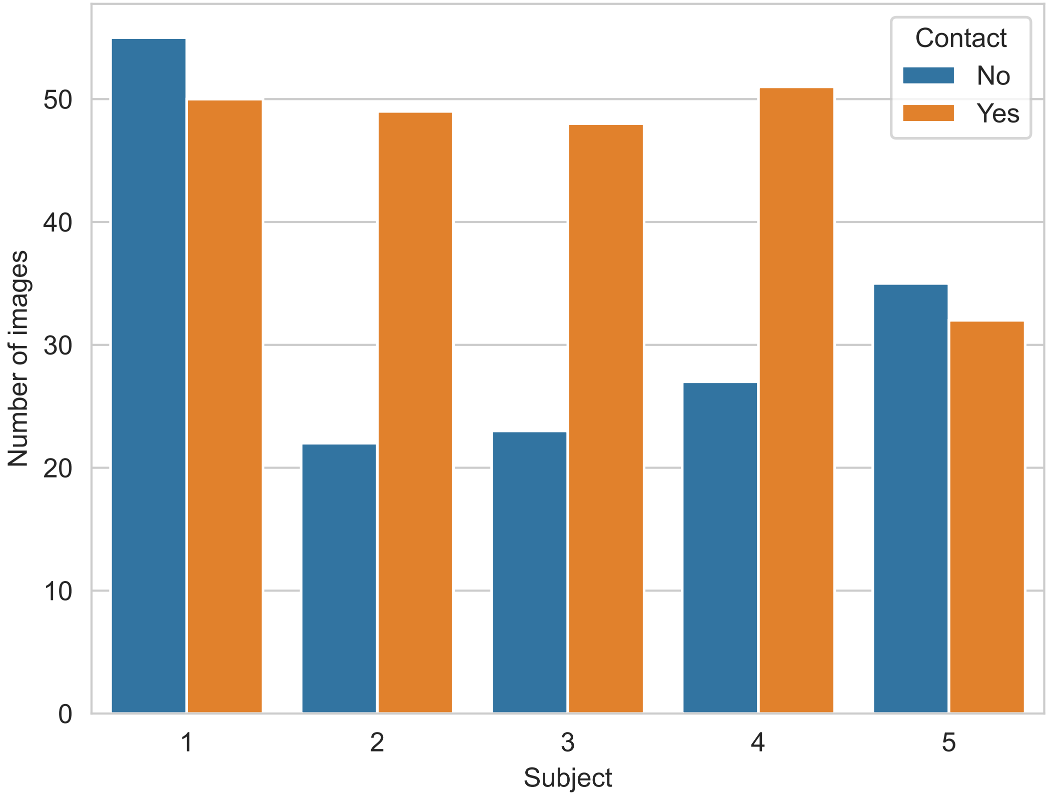

Figure 37: The number of real-time magnetic resonance images showing contact between the soft palate and posterior pharyngeal wall, complementary information to that provided in Figure 36.

Table 3: The number of velopharyngeal closures shown in the image series and Figure 36.

Subject	Velopharyngeal closures
1	8
2	4
3	4
4	6
5	8

As shown in Figure 38, there was intra-rater agreement in the labels of 385 of 392 (98.2%) images and in all 30 velopharyngeal closures. In three image series, intra-rater agreement in the labels was 100% (220 of 220) images, while in the other two image series intra-rater agreement in the labels was 97.0% (65 of 67) and 95.2% (100 of 105) of images respectively. All label differences were for images at the start or end of a velopharyngeal closure, where the soft palate is close to or in contact with the posterior pharyngeal wall.

Such discrepancies affected the durations of velopharyngeal closures but not the number of velopharyngeal closures.

There was complete inter-rater agreement in the labels of 357 of 392 (91.1%) images and in 25 of 30 (83.3%) velopharyngeal closures. All label differences were for images where the soft palate was close to or in contact with the posterior pharyngeal wall. In two image series, there was complete inter-rater agreement in all 12 velopharyngeal closures. In the other three image series, there was complete inter-rater agreement in 5 of 6 (83.3%), 3 of 4 (75.0%) and 5 of 8 (62.5%) velopharyngeal closures respectively. As shown in Figure 38, raters one and two (the two raters with the most experience of rtMRI of speech) had the highest inter-rater agreement, with agreement in the labels of 384 of 392 (98.0%) images and in all 30 velopharyngeal closures. There was inter-rater agreement between at least three raters in the labels of 385 of 392 (98.2%) images and in all 30 velopharyngeal closures. Figure 39 shows images where inter-rater agreement in labels was lower. In all five cases where there was inter-rater disagreement in a velopharyngeal closure, one rater considered there to be two closures instead of one.

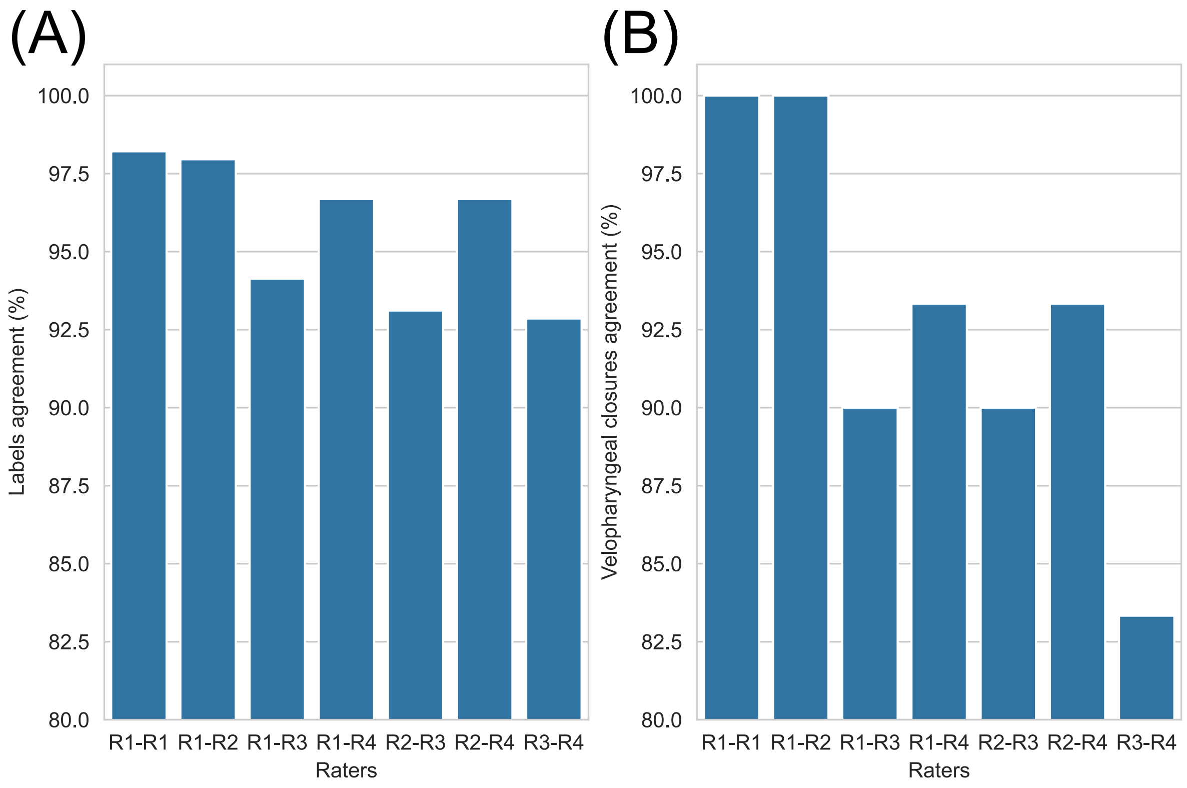

Figure 38: The intra- and inter-rater agreement in the labels of the 392 images (A) and in the velopharyngeal closures (B).

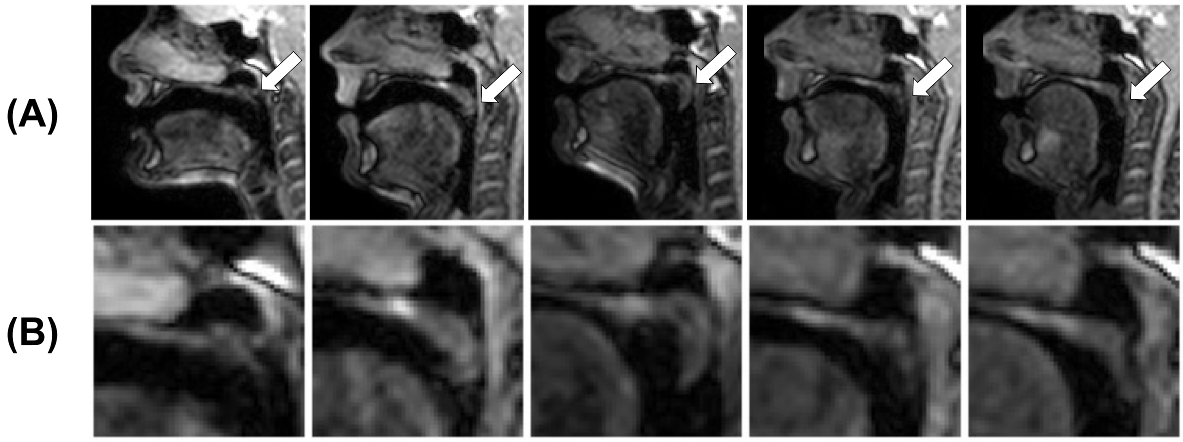

Figure 39: Real-time magnetic resonance images cropped to only show the vocal tract (A) and soft palate (B) where only two out of four raters agreed on the label.

4.1.5.2 Ground-Truth Segmentation Creation

GT segmentations for one of the image series are shown in Figure 29. In terms of number of pixels, as shown in Figure 40, the largest class was the head class with a median of 23633 pixels per segmentation, while the smallest class was the tooth space class with a median of 164 pixels per segmentation, closely followed by the soft palate class with a median of 277 pixels per segmentation.

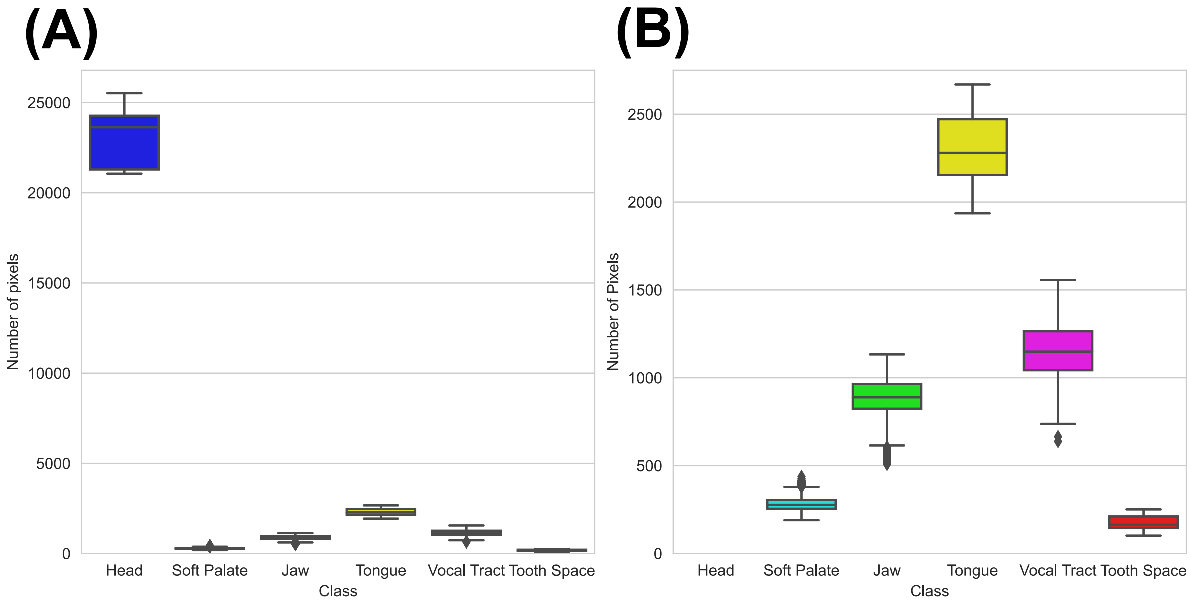

Figure 40: The number of pixels of each class per ground-truth segmentation. (B) is identical to (A) except the y-axis maximum value has been reduced to 2750.

Quantified using the DSC and HD, the median intra-rater agreement was 0.97 and 1.4 pixels respectively. As shown in Figure 41, inter-rater agreement was highest for segmentations of the head class with a median DSC of 1.0 and a median HD of 1.2 pixels, while inter-rater agreement was lowest for segmentations of the tooth space and soft palate classes, with median DSCs of 0.95 and 0.97 respectively, and a median HD of 1.4 pixels. Segmentations of the soft palate class had the largest range in DSC, closely followed by segmentations of the tooth space. A small number of segmentations of the tongue and vocal tract classes had larger HDs. Two of these larger distances were caused by the epiglottis being included in one of the segmentations of the tongue class but not the other (see Figure 42). The other larger distance was caused by contact between the head and tongue classes in one of the segmentations but not in the other (see Figure 42).

As shown in Figure 43, intra-rater agreement was consistently lower in the segmentations of images showing contact between the soft palate and posterior pharyngeal wall, across all classes and metrics. Figure 44 shows images where intra-rater agreement in segmentations was low.

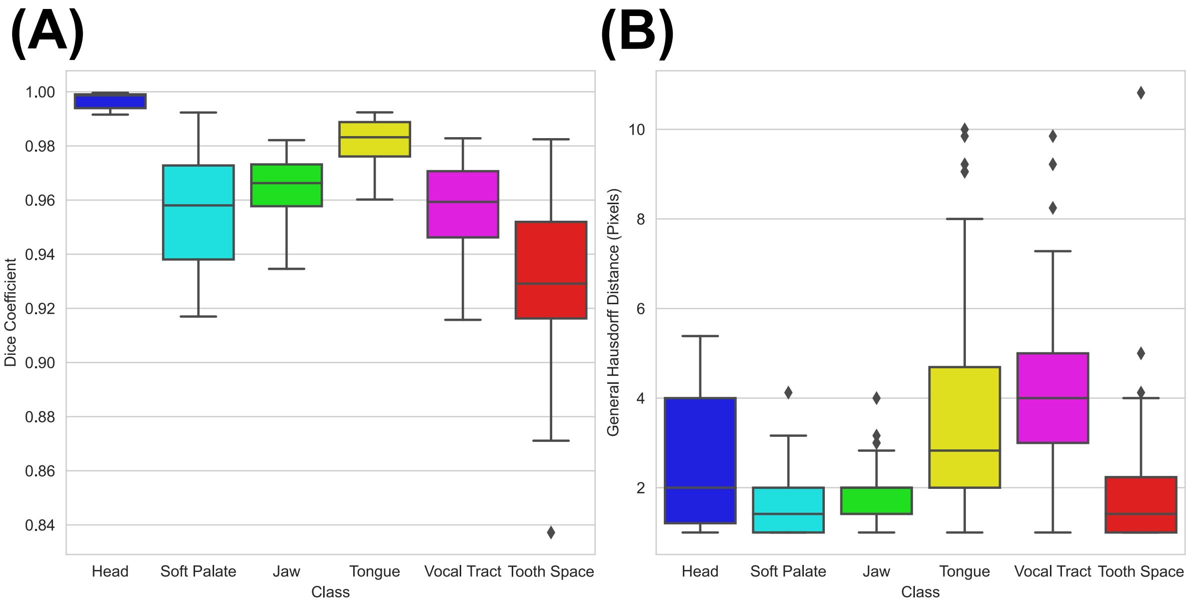

Figure 41: The intra-rater agreement in the ground-truth segmentations, evaluated using the Dice coefficient (A) and general Hausdorff distance (B).

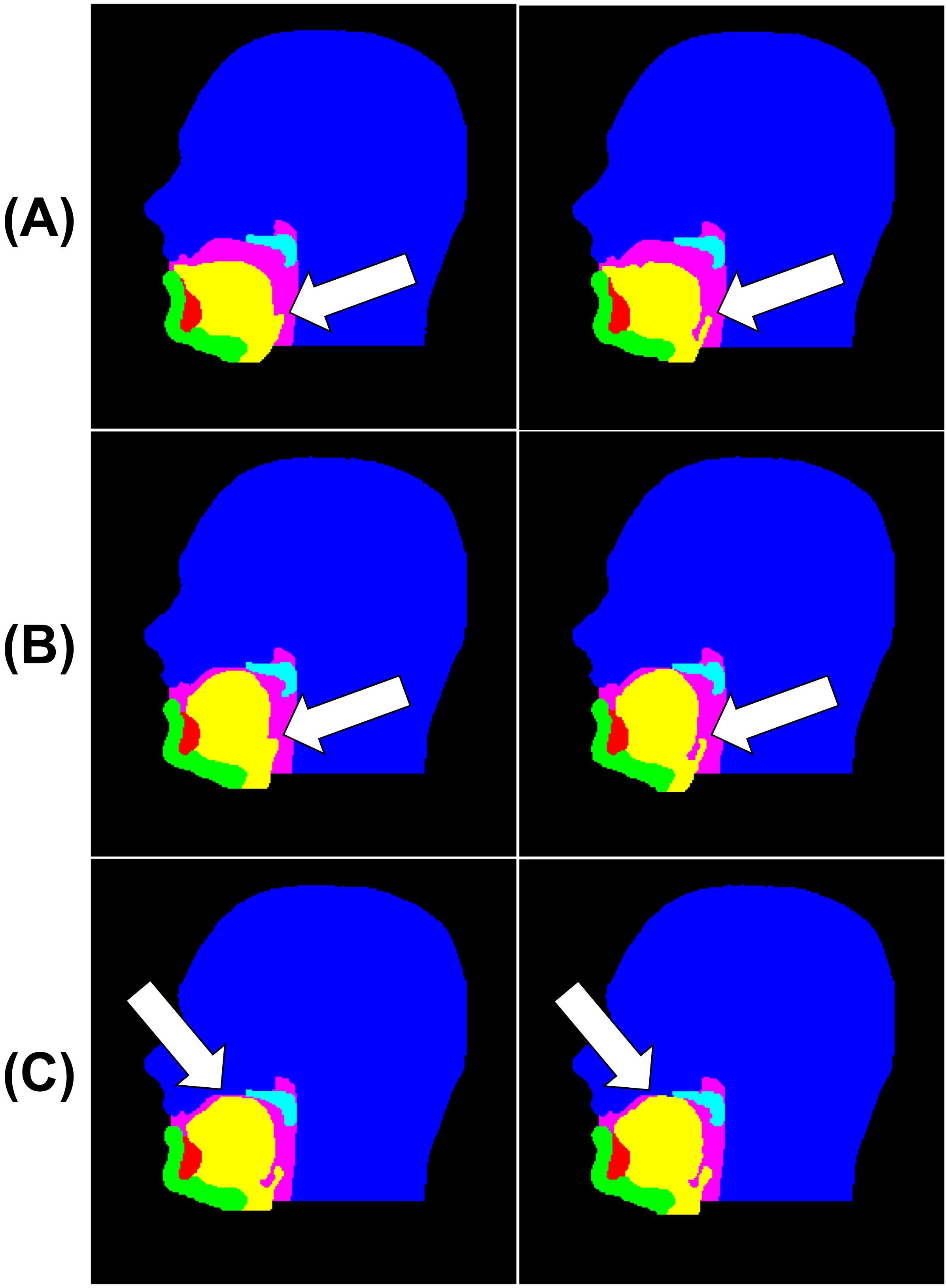

Figure 42: Pairs of ground-truth segmentations with large intra-rater differences. In rows (A) and (B), the vocal tract between the epiglottis and the anterior surface of the tongue has been included in the tongue class in the left-hand segmentation (first attempt) but not in the right-hand one (second attempt). In row (C), the head and tongue classes are in contact in the right-hand segmentation but not in the left-hand one. The ground-truth segmentations are of the head (dark blue), soft palate (light blue), jaw (green), tongue (yellow), vocal tract (pink) and tooth space (red) classes.

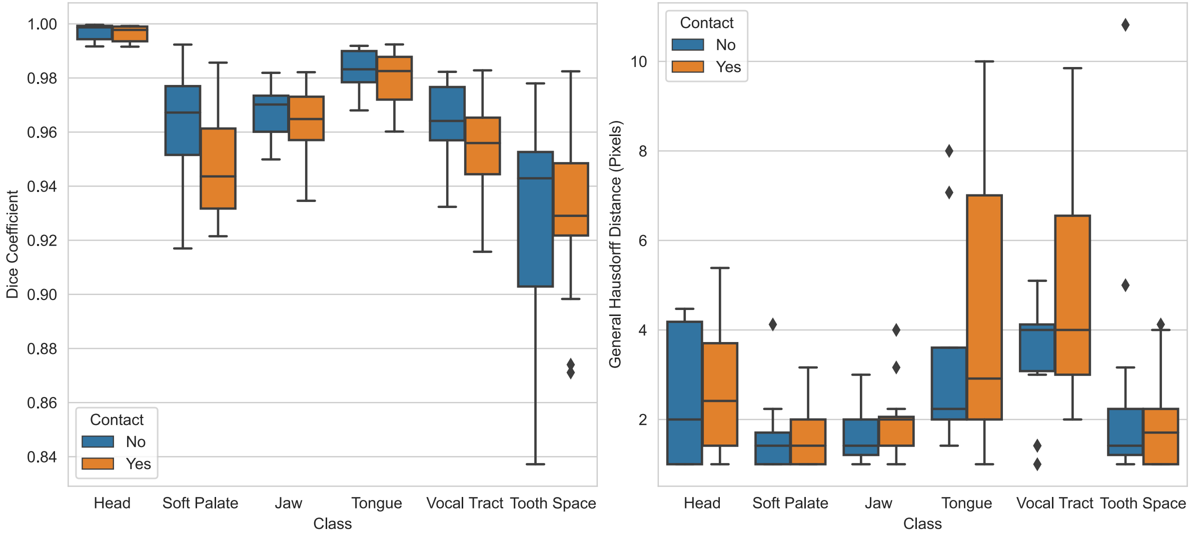

Figure 43: The intra-rater agreement in the ground-truth segmentations, evaluated using the Dice coefficient (A) and general Hausdorff distance (B), and grouped according to whether there is contact between the soft palate and posterior pharyngeal wall or not.

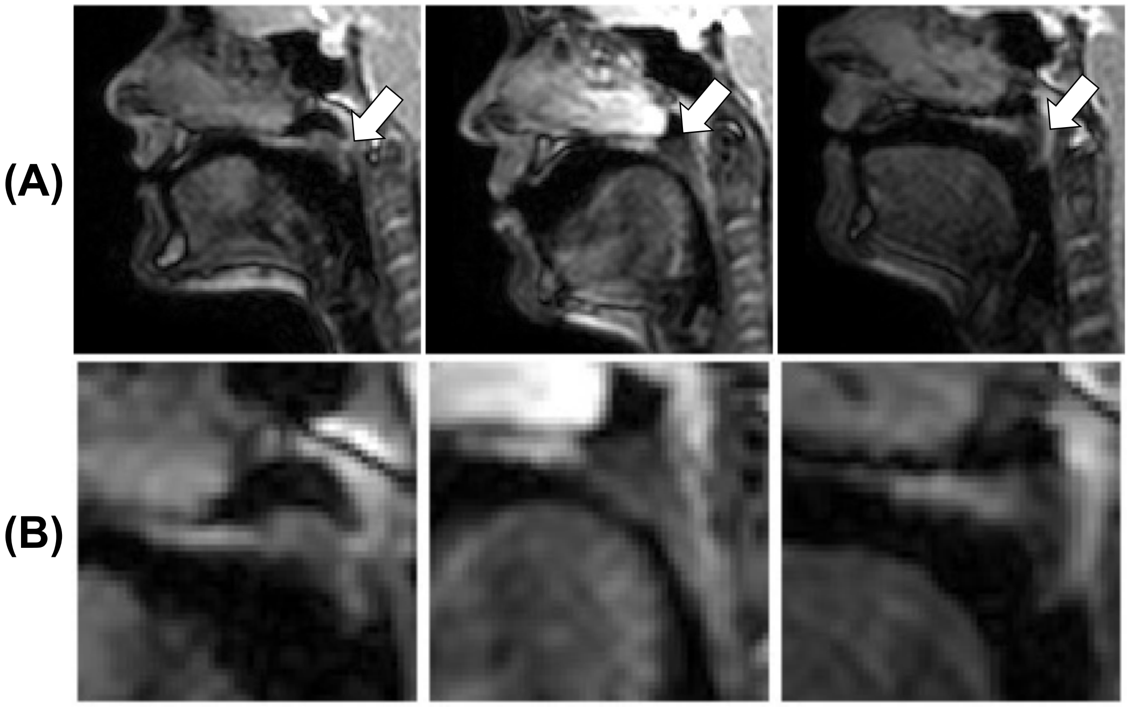

Figure 44: Real-time magnetic resonance images cropped to only show the vocal tract (A) and soft palate (B) whose ground-truth segmentations had lower intra-rater agreement. The images show examples (indicated by white arrows) of the three main image quality related challenges faced by the MRI Physicist while creating the segmentations. In the left-hand image pair, there is fluid between the soft palate and posterior pharyngeal wall. In the central image pair, there is fluid in the vocal tract and also blurring of the soft palate-vocal tract boundary as a result of motion. In the right-hand image pair, the boundary between the soft palate and posterior pharyngeal wall is unclear.

4.1.6 Discussion

4.1.6.1 Velopharyngeal Closure Identification

Labelling each image as either showing contact between the soft palate and posterior pharyngeal wall or not enabled identification of the number of velopharyngeal closures shown in the image series. Labelling of the images by multiple raters gave an indication of the subjectivity of the labels, and enabled this subjectivity to be reduced. Complete inter-rater agreement in the labels of 357 of 392 (91.1%) images demonstrates that the majority of the images clearly showed if there was contact or not. In all 35 images whose labels had lower intra-rater agreement, the soft palate was very close to the posterior pharyngeal wall, making it challenging to distinguish if there was contact or not. In 28 (80%) of these images, there was intra-rater agreement between the majority of the raters and therefore a clear consensus on what the labels for these images should be. The other seven images whose labels had the lowest intra-rater agreement were all at the start or end of a velopharyngeal closure. As a result, they only affected the duration of velopharyngeal closures and not the number of velopharyngeal closures. This suggests that there is minimal subjectivity in the identified number of velopharyngeal closures. Comparison of inter-rater agreement in image labels with other studies is not possible as there is currently no published work reporting such agreement.

The number of velopharyngeal closures shown in the image series ranged from four to eight. This range is consistent with the expected number of velopharyngeal closures for the speech task that the volunteers performed: between four and nine, depending on the rate of speech. Assuming normal speech, the start and end points of the velopharyngeal closures were also consistent with the expected points for the speech task: new velopharyngeal closures should always start when the volunteer begins saying “one”, “two”, “eight” and “ten”, and always end while the volunteer is saying “one”, “seven”, “nine” and “ten” as these four words contain the speech sound [n] whose production requires no contact between the soft palate and posterior pharyngeal wall. Depending on the rate of speech, there can be a velopharyngeal closure during production of each of the following numbers in the speech task: “two”, “three”, “four”, “five”, “six” and “seven”.

4.1.6.2 Ground-Truth Segmentation Creation

GT segmentations of six regions in rtMR images of the vocal tract during speech were successfully created. The six regions were chosen because of their relevance to speech scientists as well as clinicians assessing speech. Particularly important for this study, segmentations of the soft palate and posterior pharyngeal wall (part of the head class) were created. Automatic segmentation of these anatomical features would enable automatic measurement of the distance between the soft palate and posterior pharyngeal wall as well as automatic soft palate shape and size analyses. These are measurements and analyses that clinicians are increasingly interested in performing to investigate if these can inform treatment decisions.

In every segmentation, the head class has a much larger number of pixels than all the other classes combined. This difference in the number of pixels should be a key consideration when developing DL-based methods to segment images, as the performance of these methods can be detrimentally affected by such differences. Strategies to compensate for this difference will therefore need to be found. A strategy to compensate for this difference was used in the work presented in chapter 5 of this thesis.

Intra-rater agreement in segmentations, quantified using the DSC, was highest for segmentations of the head class, and lowest for segmentations of the soft palate and tooth space classes. This result is unsurprising as the head class has a much larger number of pixels than the soft palate and tooth space classes, therefore the effect of a pixel label discrepancy on the DSC is much larger for the latter two classes. Intra-rater agreement, quantified using the HD, was similar for all three classes with a median value of 1.4 pixels. Since the HD measures discrepancies between boundaries, this result suggests that the class boundaries including the artificial ones such as the inferior boundary of the head class in the neck (see Figure 30) were usually reproducible to within a pixel or two.

The Physicist faced three main image quality related challenges while creating the segmentations, examples of which are shown in Figure 44. First, in images in which the soft palate and posterior pharyngeal wall were in contact, there was often no clear boundary between these two anatomical features. Second, distinguishing between fluid and soft tissue in the vocal tract was challenging as both have similar intensities in the images. The third challenge was the blurring of air-tissue boundaries in the images as a result of articulator motion during image acquisition. In images with these issues, the Physicist used knowledge about the shape and position of the soft palate and posterior pharyngeal wall in earlier images to help to infer the boundaries. These three challenges are likely to be the reason why intra-rater agreement was consistently lower in segmentations of images showing contact between the soft palate and posterior pharyngeal wall, across all classes and metrics.

A few segmentations of the tongue and vocal tract classes had much larger HDs than average. These larger distances highlight two limitations of the segmentations. First, that the epiglottis was not always accurately segmented. Second, that contact between the tongue and head classes was not always consistent in the segmentations. These limitations need not be addressed for the work presented in chapters 5 and 6, as for these experiments the main requirement is that the segmentations of the soft palate and posterior pharyngeal wall are as accurate as possible. However, these limitations should be addressed for work whose main requirement is that segmentations of the tongue are as accurate as possible.

Comparison of intra-rater agreement in segmentations with other studies is not possible as there is currently no published work reporting such agreement.

Creation of the dataset is an important step towards addressing the unmet need for automatic methods to quantify the vocal tract and articulators in 2D rtMR images of the vocal tract, as it allows investigation of the feasibility of developing vocal tract and articulator segmentation methods. While the dataset is appropriate for demonstrating the feasibility of automatic vocal tract and articulator segmentation, further work is required to address two limitations of the dataset.

First, a larger and more diverse dataset, both in terms of subjects and image acquisition, and one that is more representative of the target patient population is required to develop methods suitable for clinical practice, especially given that DL-based methods usually perform poorly on data with different characteristics to the datasets used to train them. More specifically, since the target patient population primarily consists of children, the dataset must contain images of children. In addition, since velopharyngeal closure does not occur as expected in some of the speech of patients with VPI, the dataset must contain image series where velopharyngeal closure does not occur as well as image series where it does. The dataset must also be balanced in terms of gender and ethnicity, to avoid developing biased quantification methods. Regarding image acquisition parameters, the dataset used in this work consisted only of images acquired using a single MRI scanner and pulse sequence. Consequently, all the images had a very similar image contrast. Again, while using such a dataset is appropriate for demonstrating the feasibility of segmenting 2D rtMR images of the vocal tract during speech, a range of different pulse sequences have been proposed for dynamic 2D imaging of the vocal tract during speech [12,27,61]. A dataset with images acquired using many different MRI scanners and pulse sequences is therefore required to ensure that methods developed using the dataset are generalisable and perform well on images from different sources. While there are publicly available 2D rtMR image sets of the vocal tract during speech [18,19], these do not have corresponding GT segmentations thus limiting their use for training supervised DL-based segmentation methods.

4.1.7 Conclusions

GT labels and segmentations were successfully created for five series of 2D rtMR images of speech. Such segmentations are a prerequisite for the development of DL-based methods to analyse this type of image.

The GT labels enabled identification of the number of velopharyngeal closures shown in the series. Inter-rater agreement between labels was high in almost all the images. The seven images where inter-rater agreement was lower were all at the start or end of a velopharyngeal closure. As a result, they only affected the duration of velopharyngeal closures and not the number of velopharyngeal closures.

Intra-rater agreement between the GT segmentations was also high, suggesting that the process described in section 4.1.4 results in reproducible creation of segmentations. One class in the segmentations has a much larger number of pixels than all the others. This imbalance in the number of pixels should be taken into consideration when developing DL-based methods to analyse the images, as otherwise the performance of the methods may be compromised.

In the next chapter, work in which the rtMR images and their corresponding labels and GT segmentations were used for the development of an automated DL-based segmentation tool is presented.

4.2 3D Vocal Tract MRI Dataset

4.2.1 Introduction

As explained in section 3.1.9, use of MRI to visualise the LVP is increasing due to the growing availability of MRI scanners and the unique ability of MRI to non-invasively acquire 3D images with excellent soft tissue contrast and a high spatial resolution. As the LVP and the soft tissue that surrounds it have very similar tissue properties, a challenge when imaging the LVP is ensuring that the image contrast between the LVP and the surrounding soft tissue is sufficient to discriminate between the two. Previous work has predominantly acquired T_2-weighted 3D images of the LVP at 3.0 T using TSE pulse sequences [25,126,139,140,127–129,131,134,136–138]. In addition, a recommendation to acquire T_2-weighted images for assessing the LVP in clinical practice was recently made [8]. However, the results of recent work which investigated the optimal image contrast for identification of LVP landmarks in 3D images acquired at 1.5 T suggest that T_1- or PD-weighted images may enable more accurate identification [13]. However, the literature contains no reports of equivalent investigations into the optimal image contrast for 3D LVP visualisation at 3.0 T. This lack of consensus on the optimal MR image contrast for 3D LVP visualisation at 3.0 T motivated the image optimisation experiment presented in this section.

To verify the optimal image contrast for LVP visualisation, a dataset with the following properties is required. First, to increase the generalisability of the results, the dataset should consist of images of multiple subjects. Second, to enable comparison of different image contrasts, the dataset should consist of multiple images per subject. For each subject, these images should be acquired using the same set of pulse sequences. Third, for fair comparison between images acquired using different pulse sequences, all images should have the same spatial resolution. While a dataset with these properties has been acquired [13], unfortunately this dataset is not publicly available. There are publicly available MRI datasets that include 3D images of the vocal tract [18,19,117,118,220,221], however, these datasets were not intended to be used for verifying the optimal image contrast for LVP visualisation and therefore do not have the second and third properties that are required for this purpose. Consequently, a new dataset is required to enable verification of the optimal contrast for LVP visualisation.

As explained in section 3.3.5, there is increasing interest in measuring aspects of the LVP in MR images [13,25,130–139,122,140–143,123–129]. In all previous work [13,25,130–139,122,140–143,123–129], measurements such as the length and thickness of the LVP were manually obtained from MR images. However, obtaining measurements in this way is time-consuming, requires input by specialists and is prone to intra- and inter-observer variability. To avoid the burden of manual measurements and to facilitate LVP measurement on a larger scale, there is currently an unmet need for automatic LVP measurement methods. A common approach for automating the measurement of anatomical features in biomedical images is to first segment the features and then perform measurements using the segmentations. As a first step towards developing an automatic LVP measurement method, in very recent work [17], four state-of-the-art DL-based methods were used to segment the LVP in 3D T_1-weighted MR images. However, there are no reports in the literature of any methods for segmenting the LVP in 3D T_2-weighted MR images. In order to develop such a method GT segmentations are required.

GT segmentations of the LVP have been created in previous work [17,144,146], however, these segmentations have not been made publicly available. While there are publicly available MRI datasets that include 3D images of the vocal tract [18,19,117,118,220,221], these datasets either do not include GT segmentations of anatomical features [18,19,117,118] or only include GT segmentations of the vocal tract [220,221]. The current lack of publicly available LVP GT segmentations is a barrier to the development of DL-based methods to segment (and ultimately quantify aspects of) the LVP in MR images, in addition to preventing rigorous comparison of segmentation methods.

The work presented in this section makes two main contributions. First, via an image optimisation experiment, it provides new evidence on the optimal image contrast for LVP visualisation in 3D MR images acquired at 3.0 T. Second, it creates the first dataset consisting of 3D T_2-weighted MR images and GT segmentations of the LVP, a key step towards addressing the unmet need for methods to automatically segment the LVP in such images.

4.2.2 Image Optimisation Experiment

4.2.2.1 Methods

4.2.2.1.1 Image Acquisition

Five healthy adult volunteers (two females, three males; age range 21 to 31 years) participated in the experiment, after providing informed consent in accordance with ethics committee requirements. The volunteers were imaged in a supine position using a 3.0 T SIGNA Architect MRI scanner, a 45-channel head and neck receive coil (both GE HealthCare, Milwaulkee, WI) and 3D TSE pulse sequences. Three 3D images were acquired per volunteer, using three TSE pulse sequences (more specifically, CUBE pulse sequences) with parameters that resulted in the acquisition of one T_1-weighted image, one PD-weighted image and one T_2-weighted image. CUBE pulse sequences were chosen as these are already highly optimised to enable acquisition of images with specific contrasts and a high spatial resolution. In total, 15 images were acquired in the experiment. Pulse sequence parameters and scan durations are listed in Table 4. In total, 15 images of the entire head (example images are shown in Figure 45) were acquired in the experiment. The reason why images of the entire head were acquired, rather than images of a smaller volume centred on the LVP and pharynx, was to avoid “phase wrap-around” artefacts in the images, as these artefacts can obscure anatomical features of interest [273].

Table 4: Parameters of the three pulse sequences used in the experiments. PD: proton-density; FOV: field of view; TR: repetition time; TE: echo time.

Parameter	**Pulse sequence**		
	T_1-weighted	PD-weighted	T_2-weighted
FOV (mm^3)		256×243×168	
Acquired voxel size (mm^3)		0.8×0.8×1.2	
Reconstructed voxel size (mm^3)		0.5×0.5×0.6	
Signal averages		1	
TR (ms)	550	3000	3000
TE (ms)	16	60	100
Echo train length	22	130	130
Bandwidth per pixel (Hz)		390.6	
GRAPPA factor		2 (in both phase and slice encoding directions)	
Scan time (s)	231	208	208

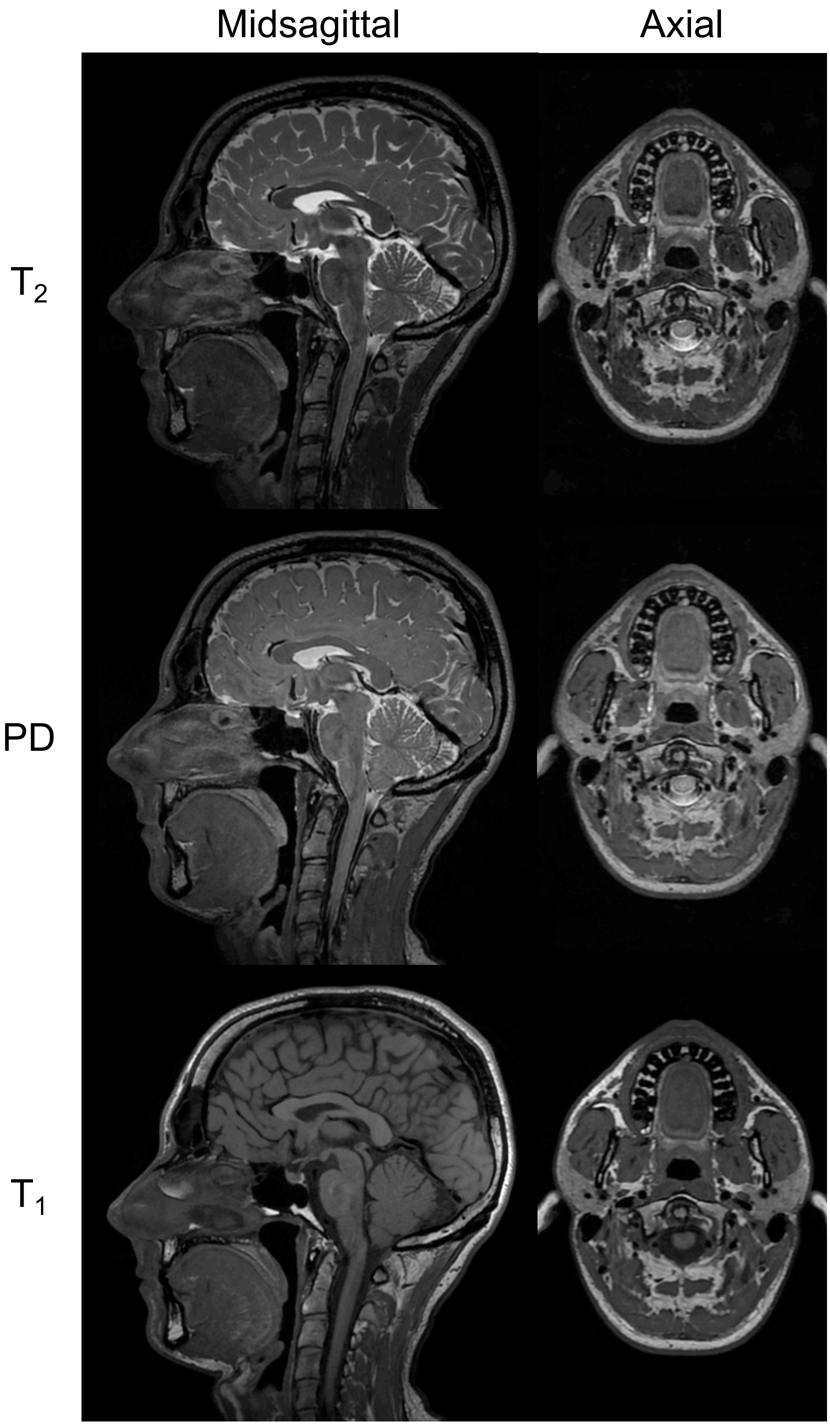

Figure 45: Example midsagittal and axial slices from the three-dimensional images acquired in the image optimisation experiment. T_1, PD (proton density) and T_2 indicate the contrast weighting of the images.

4.2.2.1.2 Image Analysis

The image contrast at the location where the LVP is connected to the soft palate was assessed as this location is the most relevant for clinical teams treating patients with VPI. More specifically, the location of the connection and the structure of the muscle are factors that affect VPI treatment decisions. To assess its contrast, each 3D image was analysed in the following way:

1. The image was visually inspected and an axial slice in which the LVP was clearly visible was identified.
2. Two regions of interest (ROIs) were manually drawn on the image. One was drawn on the LVP while the other was drawn in the adjacent soft tissue. Example ROIs are shown in Figure 46.
3. The mean voxel value in each ROI was calculated.
4. The contrast between the LVP and the adjacent soft tissue in an image was quantified using the following equation [274]:

$$C = \left|\frac{\sigma_{LVP} - \sigma_A}{\sigma_{LVP} + \sigma_A}\right| \tag{19}$$

where σ_{LVP} and σ_A are the mean voxel intensities in the LVP and adjacent soft tissue ROIs respectively. C was used as an indicator of the ease with which the LVP could be distinguished from the soft tissue surrounding it: a higher value indicated that the LVP was more easily distinguishable.

To identify the optimal contrast for visualising the LVP, the values of C of the three images of a subject were compared.

The visual inspection in step 1 and manual ROI drawing in step 2 was performed by an MRI Physicist with five years of experience of speech MRI. The image analysis was performed using the medical image viewer Horos v3.3.6 [275].

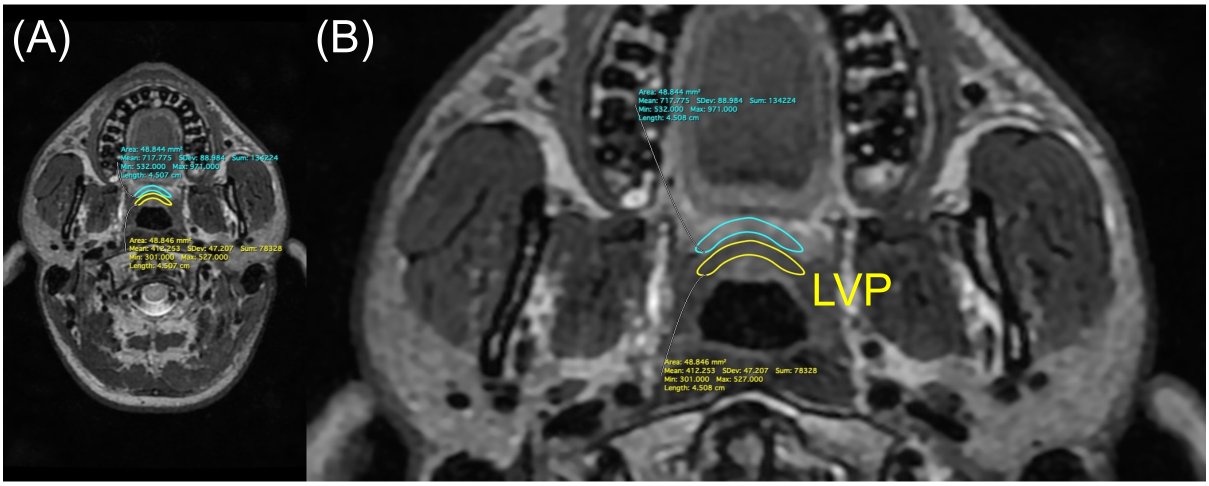
(A)
Area: 48.844 mm²
Mean: 717.775 SDev: 88.984 Sum: 134224
Min: 532.000 Max: 971.000
Length: 4.507 cm
Area: 48.846 mm²
Mean: 412.253 SDev: 47.207 Sum: 78328
Min: 301.000 Max: 527.000
Length: 4.507 cm
(B)
Area: 48.844 mm²
Mean: 717.775 SDev: 88.984 Sum: 134224
Min: 532.000 Max: 971.000
Length: 4.508 cm
LVP
Area: 48.846 mm²
Mean: 412.253 SDev: 47.207 Sum: 78328
Min: 301.000 Max: 527.000
Length: 4.508 cm

Figure 46: Example regions of interest (ROIs). Image (B) is a cropped version of image (A). The yellow ROI is on the levator veli palatini (LVP) while the light blue ROI is on the soft tissue adjacent to the LVP.

4.2.2.2 Results

Example images are shown in Figure 45, while Figure 47 shows the values of C in the images. In four of five subjects, C was greatest in their T$_2$-weighted image, while in the other subject, C was greatest in their T$_1$-weighted image. In four of five subjects, C was lowest in their PD-weighted image, while in the other subject, C was lowest in their T$_1$-weighted image.

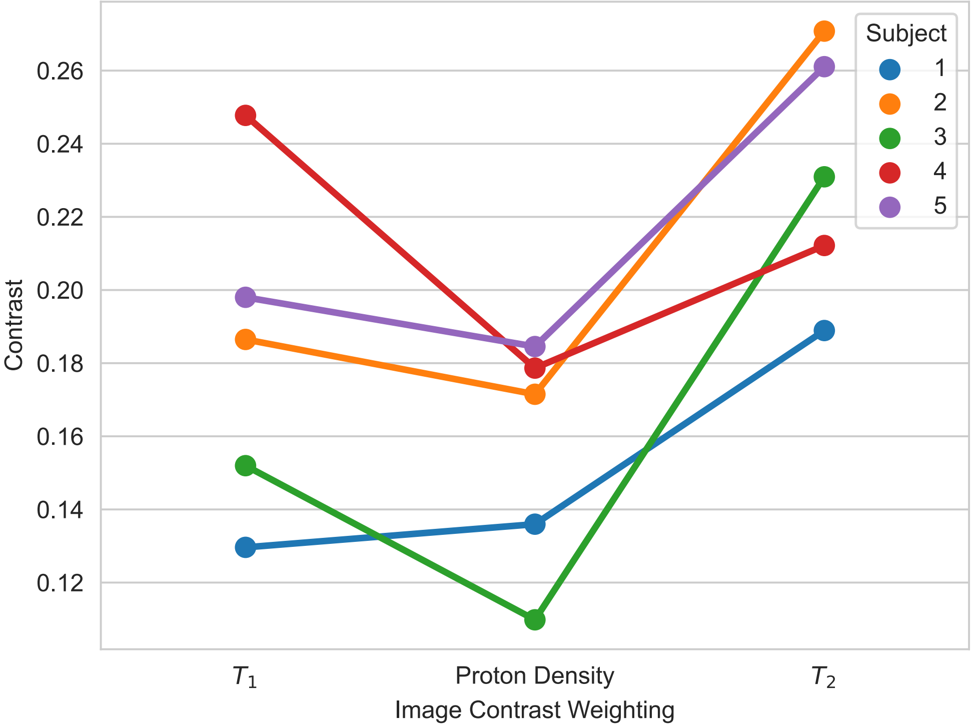

Figure 47: The contrast between the levator veli palatini and the adjacent soft tissue in the T$_1$-, proton-density- and T$_2$-weighted magnetic resonance images.

4.2.2.3 Discussion

Only T_2-weighted images of the LVP were acquired in almost all previous work involving 3D imaging of the LVP at 3.0 T [13,25,137–142,126–129,131–133,136]. However, the results of recent work which investigated the optimal image contrast for identification of LVP landmarks suggest that T_1- or PD-weighted images may enable more accurate identification [13]. The aim of the image optimisation experiment was to quantitatively compare the contrast between the LVP and the adjacent soft tissue in T_1-, PD- and T_2-weighted images, to identify the type of image with optimal contrast for visualising the LVP.

As shown in Figure 47, in four of five subjects, the contrast between the LVP and adjacent soft tissue was greatest in their T_2-weighted image. This result shows that the difference in voxel intensities was largest in T_2-weighted images, suggesting that the LVP is more easily distinguishable in these images than in T_1- and PD-weighted images acquired at 3.0 T. This finding provides evidence to support the recently-made recommendation to acquire T_2-weighted images for assessing the LVP in clinical practice [8] and the choice made in all previous work involving 3D imaging of the LVP at 3.0 T to acquire T_2-weighted images [13,25,137–142,126–129,131–133,136].

Conversely, in four of five subjects, the contrast between the LVP and adjacent soft tissue was lowest in their PD-weighted image. This result shows that the difference in voxel intensities was smallest in PD-weighted images, suggesting that the LVP would be more challenging to distinguish in these images than in T_1- and T_2-weighted images acquired at 3.0 T.

The main limitations of the image optimisation experiment are its small sample size (15 images of five subjects), the assessment of contrast at a single location only and the limited number of different contrasts that were investigated. Regarding the sample size, further work is required to increase the sample size and verify the findings of this image optimisation experiment. Regarding contrast assessment location, the image contrast was assessed at the location where the LVP is connected to the soft palate as this location is the most relevant for clinical teams treating patients with VPI. More specifically, the location of the connection and the structure of the muscle at this location are factors that affect VPI treatment decisions. However, further work is required to identify the optimal contrast for visualising other sections of the LVP using 3D MRI. Regarding the limited number of contrasts, while the investigation provides an indication of the optimal image contrast for

LVP visualisation in 3D MR images of the vocal tract, further work is required to pinpoint the key pulse sequence parameters (i.e. TR and TE) that result in optimal image contrast. This pinpointing could be achieved by, for example, acquiring a wider range of T_2-weighted images and analysing these images in the way described in section 4.2.2.1.2.

4.2.2.4 Conclusions

The visibility of the LVP relative to adjacent soft tissue was found to be greatest in T_2-weighted images. Based on this finding, a larger dataset of T_2-weighted images was created and then used in the development of a DL-based method to segment the LVP, work described in chapter 7 of this thesis.

4.2.3 Image and GT Segmentation Dataset Creation

The results of the image optimisation experiment described in section 4.2.2 suggest that T_2-weighted images are optimal for visualising the LVP. Based on this result, T_2-weighted images were acquired and then manually segmented to create a dataset to enable the development of automatic LVP segmentation methods.

4.2.3.1 Methods

4.2.3.1.1 Image Acquisition

Fifteen healthy volunteers (eight females, seven males; age range 21 to 31 years) participated in the experiment, after providing informed consent in accordance with ethics committee requirements: the five volunteers from the image optimisation experiment and 10 additional volunteers. The additional volunteers were imaged in the same way as described in section 4.2.2.1.1, but only using the TSE pulse sequence with parameters that resulted in the acquisition of T_2-weighted 3D images. One T_2-weighted image per volunteer was included in the dataset. The dataset therefore included 15 images in total. Pulse sequence parameters and scan durations are listed in Table 4. Example images are shown in Figure 45.

4.2.3.1.2 GT Segmentation Creation

GT segmentations were created by manually labelling voxels in the images. The segmentations consisted of three classes: LVP, pharynx and background. There was no overlap between the classes: a pixel could not belong to more than one class. The two reasons for including the pharynx in the GT segmentations were as follows. First, to provide information about the orientation of the LVP relative to the soft palate. A segmentation of the pharynx can provide such information as the anterior boundary of the pharynx is the superior surface of the soft palate. Second, to enable measurement of its volume and shape, aspects of the pharynx that are clinically relevant for VPI treatment planning.

The boundaries of a large section of the pharynx are clearly defined: they are the pharyngeal wall and the superior surface of the soft palate. However, the superior and inferior boundaries of the pharynx are not so clear. Instead, the following artificial boundaries were devised for these sections. The two main goals when devising these boundaries were firstly to include only relevant anatomical features and secondly to make the boundaries as easily reproducible as possible. The superior boundary of the pharynx was defined as the axial slice at the level of the hard palate (see Figure 49D), while the inferior boundary was defined as the axial slice level with the tip of the soft palate (see Figure 49D). These definitions were considered to provide an acceptable trade-off between reproducibility and inclusion of relevant sections of the pharynx.

GT segmentations were created by an MRI Physicist with six years of speech MRI experience using 3D Slicer version 4.11.20210226 [276]. The process for creating a GT segmentation of the LVP in a 3D image was as follows:

1. **Oblique axial slice identification:** an oblique axial slice of the 3D image showing a longitudinal section of the LVP was identified via visual inspection (see Figure 48A).
2. **Initial manual segmentation:** voxels showing the LVP in the oblique axial slice identified in step 1 and in adjacent oblique axial slices were manually labelled (see Figure 48B).
3. **Axial slice identification:** an axial slice of the 3D image showing the LVP was identified via visual inspection (see Figure 48C).
4. **Initial segmentation refinement:** in the axial slice and in adjacent axial slices showing the LVP, the initial LVP segmentation was manually refined to fill “holes” in it (see Figure 48D).

5. **Segmentation post-processing:** the LVP segmentation was morphologically closed using a kernel size of 3×3×1 to fill any remaining “holes” in it.

The process for creating a GT segmentation of the pharynx in a 3D image was as follows:

1. **Initial binary mask creation:** a binary mask of the entire head was created by applying a threshold to the image (see Figure 49B).
2. **Binary mask cropping:** the binary mask created in step 1 was manually cropped to a smaller cuboid containing the pharynx. The superior surface of the cuboid was the axial slice at the level of the hard palate, while the inferior surface was the axial slice level with the tip of the soft palate (see Figure 49C).
3. **Binary mask refinement:** the binary mask was manually refined. More specifically, voxels corresponding to fluid on the surface of the soft palate and pharyngeal walls in the image were identified via visual inspection and then manually removed from the binary mask.
4. **Pharynx segmentation creation:** the voxel values in the binary mask created in step 2 were switched (i.e. voxels labelled as 0 were changed to voxels labelled as 1 and vice versa). This created several connected components, the largest of which was the pharynx GT segmentation (see Figure 49D).

The number of voxels in the LVP and pharynx GT segmentations was calculated, to enable comparison of the size of the two classes.

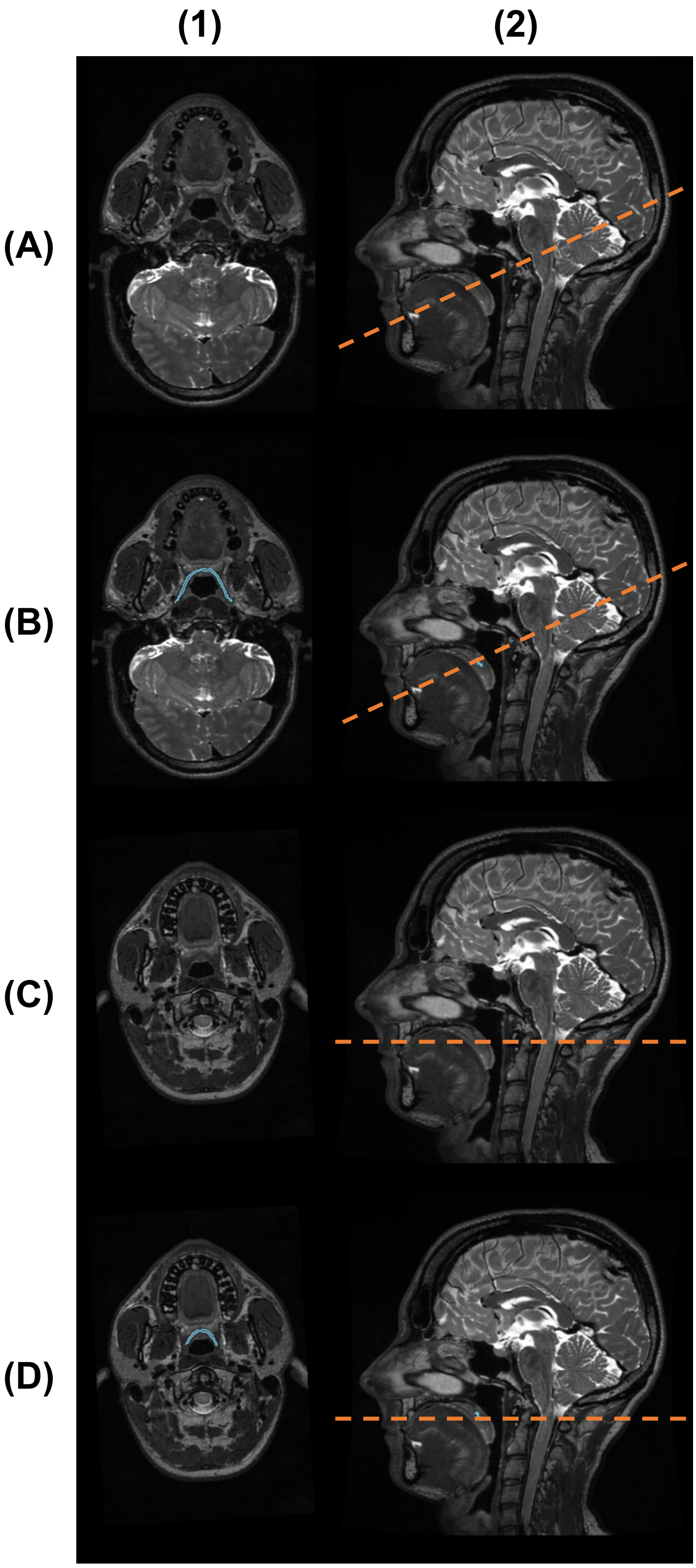

Figure 48: Two-dimensional slices of one of the three-dimensional magnetic resonance images. Column (1) shows an oblique axial slice (A and B) and an axial slice (C and D), while column (2) shows a midsagittal slice. In the slices, light blue shading indicates the levator veli palatini (LVP) ground-truth (GT) segmentation, while the orange dashed lines in column (2) indicate the plane of the slice shown in column 1. In column (1), rows (A) and (B) show the same oblique axial slice without and with the LVP GT segmentation overlaid, while rows (C) and (D) show the same axial slice without and with the LVP GT segmentation overlaid.

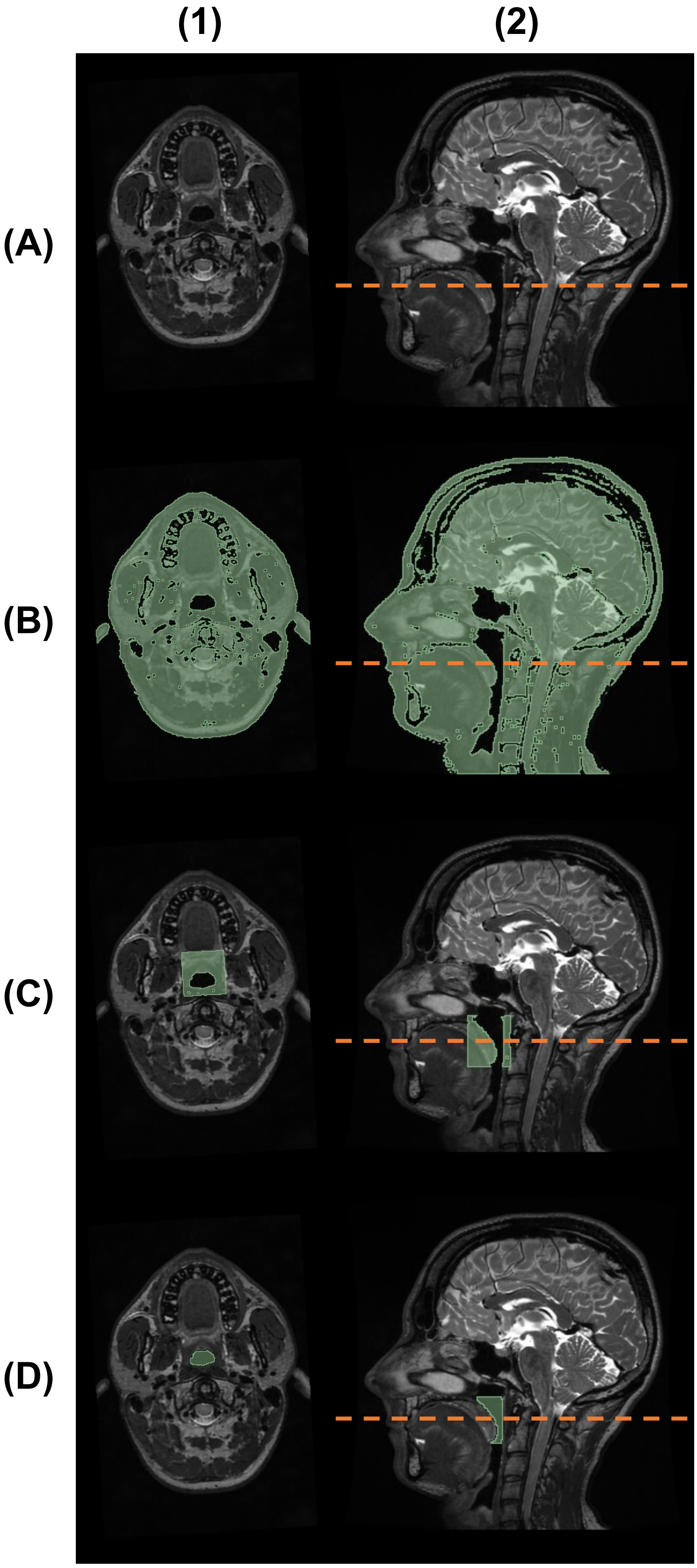

Figure 49: Two-dimensional slices of one of the three-dimensional magnetic resonance images. Column (1) shows an axial slice, while column (2) shows a midsagittal one. Row D shows slices with the pharynx ground-truth segmentation overlaid, while rows B and C show preliminary segmentations. Green shading indicates a segmentation, while the orange dashed lines in column (2) indicate the plane of the slice shown in column (1).

4.2.3.2 Results and Discussion

A dataset consisting of 15 3D MR images of the entire head, each of a different healthy adult volunteer, and GT segmentations of the LVP and pharynx in the images was successfully created. Similarly to previous work [13,25,137–141,126–129,131–133,136], in this work imaging was at 3.0 T and a TSE pulse sequence was used to acquire T_2-weighted images of the entire head at a spatial resolution of $0.8\times0.8\times0.8$ mm^3. The reason why images of the entire head were acquired, rather than images of a smaller volume centred on the LVP and pharynx, was to avoid "phase wrap-around" artefacts in the images, as these artefacts can obscure anatomical features of interest [273].

The dataset includes GT segmentations of the LVP and pharynx, thus enabling the development of methods to automatically segment these anatomical features, key steps towards addressing the current unmet need for automatic methods to measure the LVP in 3D MR images. GT segmentations of the LVP have been created in previous work, however, only for single images [144,146] or for T_1-weighted images [17].

Figure 50 shows all 15 GT segmentations, while Figure 51 shows the number of voxels per class in the GT segmentations. As shown in Figure 51, there were more voxels of the pharynx than the LVP in all GT segmentations: the median number of voxels of the LVP and pharynx was approximately 10,000 and 43,000 respectively. Given that the images each consisted of 512×512×272 voxels, only a small proportion of their voxels corresponded to a segmentation class: approximately 0.01% and 0.06% corresponded to the LVP and pharynx respectively. Cropped versions of the images were used in the development of a DL-based method to segment the LVP in 3D MR images (work described in chapter 7), for two main reasons. First, to reduce the computational burden of the method development. Second, to reduce the complexity of the segmentation task by reducing the number of anatomical features in the image and by increasing the proportion of voxels corresponding to the LVP and pharynx.

There is increasing interest in visualising and measuring aspects of the LVP, to better understand variations in its shape and configuration [25,122,131–140,123,141–143,124–130], to aid planning of surgical treatment of VPI [144,145], and for medical education purposes [146]. While the dataset presented in this work was primarily created to enable the development of automatic LVP measurement methods, it also provides opportunities to develop novel LVP visualisation methods such as patient-specific computer or physical 3D models of the LVP for use in VPI treatment planning. Further work is required to explore these opportunities with clinical teams and investigate if segmentations of the LVP and pharynx provide enough anatomical context for use in VPI treatment planning. For example, additional segmentations of anatomical features such as the soft palate may be required to provide enough anatomical context. Creating segmentations of the soft palate is not essential for enabling automatic LVP measurement and was therefore not prioritised in this work. However, a key challenge that would need to be addressed to reproducibly create GT segmentations of the soft palate is the lack of clear anatomical boundaries between the lateral sections of the soft palate and the adjacent soft tissue. Instead, artificial boundaries that are easily reproducible would need to be devised for these sections of the soft palate.

Creation of the dataset is an important step towards addressing the unmet need for automatic methods to measure the LVP in 3D MR images, as it allows investigation of the feasibility of developing automatic methods to segment the LVP in these images. While the dataset is appropriate for demonstrating the feasibility of automatic LVP segmentation, further work is required to address two limitations of the dataset.

First, a larger and more diverse dataset, both in terms of subjects and image acquisition, and one that is more representative of the target patient population is required to develop automatic LVP measurement methods suitable for use in clinical practice. More specifically, since the target patient population primarily consists of children, the dataset must contain images of children. In addition, since LVP anomalies are prevalent in the target population, the dataset must contain images of LVPs with anomalies as well as LVPs without. The dataset must also be balanced in terms of gender and ethnicity, to avoid potentially developing biased LVP measurement methods. Regarding image acquisition, the images in the dataset were all acquired using the same MRI scanner and pulse sequence. A dataset of images acquired using many different MRI scanners and pulse sequences is required to ensure that methods developed using the dataset perform well on images from different sources. This generalisability is a key requirement for methods suitable for use in clinical practice.

Second, the dataset presented in this work includes GT segmentations created by a single expert only. Future work should investigate the intra- and inter-rater reliability of the GT segmentations for two main reasons. First, to verify that the GT segmentation creation process described in section 4.2.3.1.2 is reproducible. Second, to provide insights into the accuracy and reliability of manual LVP segmentation (for example, through identification of sections of the LVP where agreement between raters is lower) and thus provide information about the maximum accuracy that can be achieved by automatic LVP segmentation methods. Such an investigation would require manual creation of GT segmentations by several experts and would therefore be very time-consuming.

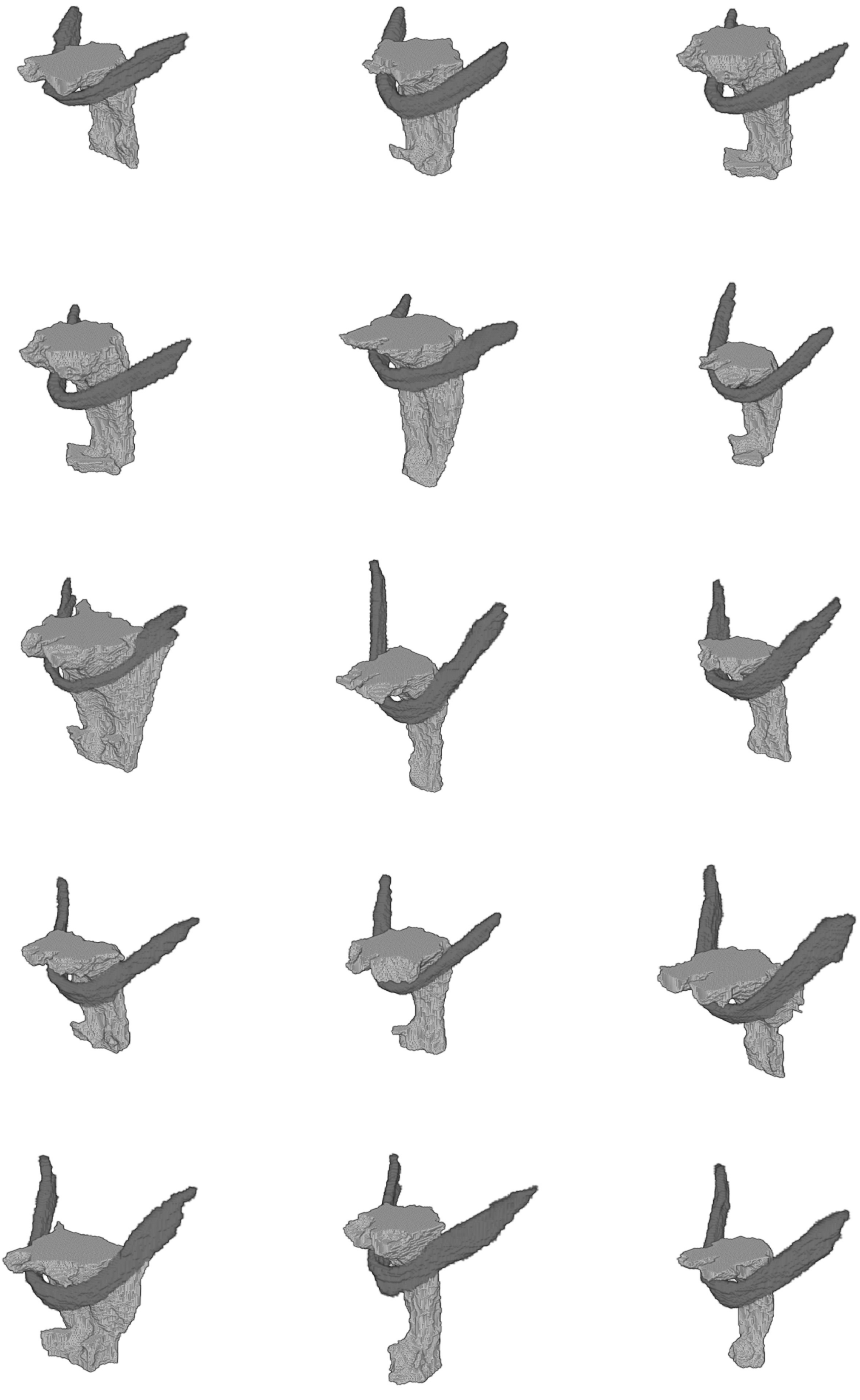

Figure 50: Ground-truth segmentations of the levator veli palatini (dark grey) and pharynx (light grey) in each of the 15 images.

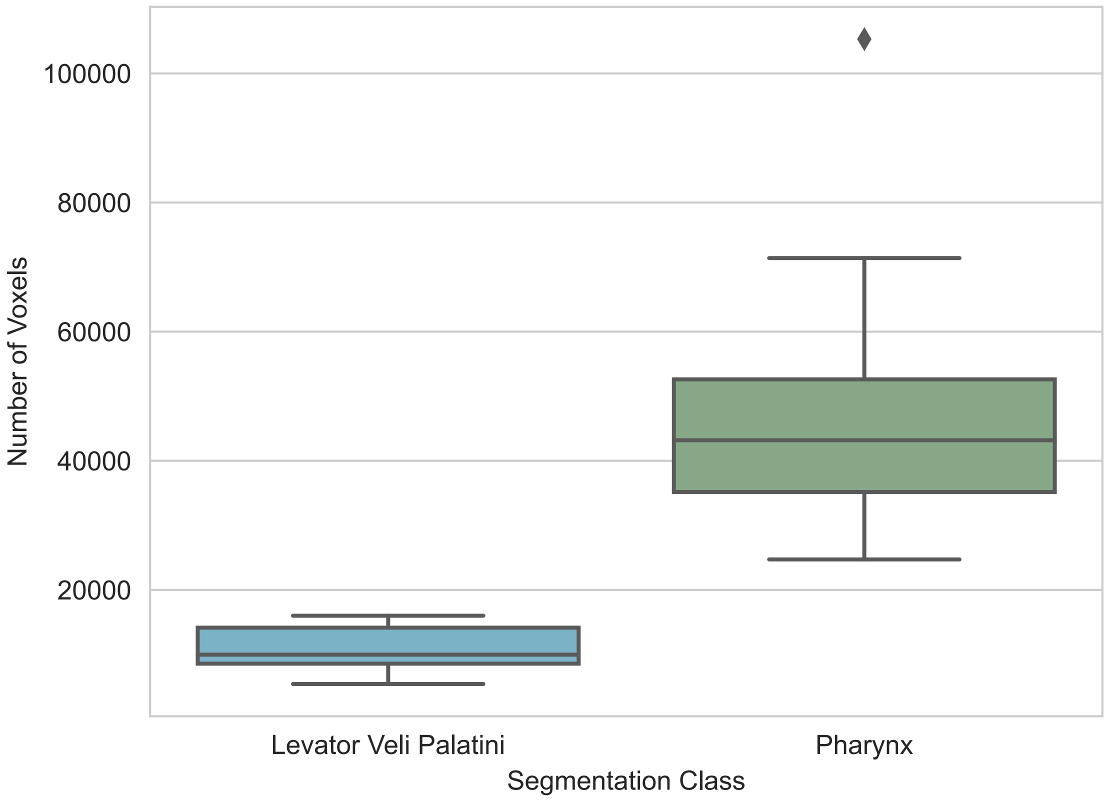

Figure 51: Number of voxels per segmentation class in the ground-truth segmentations of the 15 magnetic resonance images.

4.2.3.3 Conclusions

For the first time, a dataset consisting of 3D T_2-weighted MR images of the vocal tract and GT segmentations of the LVP and pharynx was created. This dataset enabled the development of automatic methods to segment the LVP in 3D MR images (work described in chapter 7), a key step towards addressing the current unmet need for automatic methods to measure the LVP in such images.

Chapter 5: Articulator Segmentation in MR Images of Speech

5.1 Introduction

5.1.1 Motivation

As explained in section 3.1.7, use of rtMRI to visualise articulators and the vocal tract during speech is increasing in both research and clinical settings [12,27,61]. This increase is a result of the development of rtMRI techniques with relatively high spatio-temporal resolutions and the unique ability of rtMRI to noninvasively acquire images of any view without using ionising radiation [12,27,61]. Visualisation of articulators and the vocal tract during speech provides information about their shape, size, position and motion. This information is helping researchers to answer open questions about speech production [14,18,196,197,277–279,19,27,61,68,69,72,75,76], while in the clinical speech assessment of patients with VPI this information aids clinicians to diagnose the cause(s) of VPI and then make treatment decisions [3,4,8,280].

Typically during rtMRI of speech, series of 2D images of a midsagittal slice of the vocal tract are acquired. As explained in section 3.3.3, there is increasing interest in extracting quantitative information about the vocal tract and articulators from such images [7,14,75,76,84,88,195–198,20,68–74]. More specifically, there is interest in measuring the size and shape of the vocal tract [14,68,197,198,70,74–76,84,88,195,196], the size, shape and motion of the soft palate [72,73,75,84,198], lip motion [69,75,84], tongue motion [20,84] and the distance between the soft palate and the posterior pharyngeal wall [7,72,198]. Manual measurement to obtain this information is time-consuming, requires input by specialists and is prone to intra- and inter-observer variability. The increasing interest in extracting quantitative information, in combination with the need to avoid manual measurement, have motivated the development of a range of methods to (semi-)automatically extract this information [15,16,206,207,20,199–205]. Almost all these methods are segmentation based [15,16,207,199–206].

5.1.2 Related Work

Several methods to semi-automatically measure the shape of the vocal tract in 2D rtMR images of speech have been developed [15,16,199–204,208]. One of these methods segmented the entire vocal tract [208], while the others labelled pixels at air-tissue boundaries between the vocal tract and adjacent tissues and therefore only created partial contours of articulators [15,16,199–204]. The methods have been based on a variety of approaches. In [203], the air-tissue boundaries between the vocal tract and adjacent tissues were automatically labelled using an optimisation algorithm to adjust an anatomically informed synthetic image of the vocal tract until the *k*-space of the synthetic image was as similar as possible to the *k*-space of the MR image. Other methods performed the labelling by analysing pixel values along gridlines superposed on the MR image [199] or by using active shape models [15,204].

More recently, DL-based methods have been developed to automatically label air-tissue boundaries between the vocal tract and adjacent tissues [16,200–202,208]. In [16] and [201], FCNs with an architecture similar to SegNet [209] were developed to label the air-tissue boundaries and, in [16], identify which articulators the boundary pixels belonged to. In [200] and [202], FCNs with architectures similar to the original FCN [162] and the FCN in [210] respectively were developed to label the air-tissue boundaries. An FCN with an architecture similar to the original U-Net [175] was developed in [208] to segment the entire vocal tract, not just its boundaries with adjacent tissues.

However, none of the methods described in the two paragraphs above segment entire articulators. Such segmentation is desirable as it would enable measurement of articulator shape, size, position and motion.

This chapter presents a DL-based framework to address this limitation, and is based on a peer reviewed and published journal article [205]. Since the publication of [205], three further related works have been published by other researchers. Two of these works are methods to analyse 2D rtMR images of speech [72,206], while the other is a method to segment the pharynx and entire articulators (tongue and soft palate) in 2D static MR images of the vocal tract [218]. In [72], a method that used generalised additive mixed models [196] to measure the distance between the soft palate and posterior pharyngeal wall was developed, while in [206] a DL-based method to segment entire articulators (soft palate, hard palate, tongue, jaw and upper lip) was developed.

5.1.3 Clinical Considerations

As explained in section 2.1, velopharyngeal closure is a key requirement in the production of the majority of speech sounds [1,22]. During velopharyngeal closure, the soft palate elevates and comes into contact with the pharyngeal wall. However, in patients with VPI, velopharyngeal closure does not always occur, thus causing speech problems [1]. There are large variations in the speech sounds where velopharyngeal closure does not occur in patients with VPI: in some patients velopharyngeal closure does not occur in a few speech sounds only, while in others velopharyngeal closure never occurs. An important consideration when making treatment decisions is the speech sounds where velopharyngeal closure does not occur. Therefore, to be suitable for use in clinical practice, a key requirement for articulator analysis methods is the accurate detection of any velopharyngeal closures that occur. In addition, it is important that articulator analysis methods do not falsely detect velopharyngeal closures when none have occurred. However, standard metrics for evaluating segmentation accuracy do not provide information about the accurate detection of velopharyngeal closure. By comparing the velopharyngeal closures in GT segmentations with those in segmentations estimated by an automated method, the ability of the method to accurately capture velopharyngeal closures can be assessed.

5.1.4 Contributions

The work presented in this chapter makes two contributions. First, it develops a fully automatic DL-based method for segmenting entire articulators and the vocal tract in 2D rtMR images of speech. The method also includes an extension to automatically measure the minimum distance between the soft palate and the posterior pharyngeal wall. This contribution is a step towards addressing the unmet need of automatic measurement of articulator shape, size, position and motion in 2D rtMR images of speech. Second, this work proposes a new metric for evaluating the accuracy of estimated segmentations. This metric is based on velopharyngeal closure, a quantifiable and clinically relevant aspect of articulator motion.

As stated earlier, part of the work presented in this chapter was published as a journal article [205]. Two extensions to the published work are presented in this chapter. First, an extension to the segmentation method in order to enable automatic measurement of the minimum distance between the soft palate and the posterior pharyngeal wall is presented. Second, an additional loss function weighting investigation is presented.

5.2 Method

5.2.1 Overview

Figure 52 shows an overview of the proposed DL-based method. Given a 2D rtMR image of the vocal tract, the method will estimate segmentations for six different anatomical features in the image and then measure the minimum distance between the soft palate and the posterior pharyngeal wall. Segmentations are estimated using a CNN with a similar architecture to the original U-Net [175]. The estimated segmentations are then post-processed to remove anatomically impossible regions in the images. Finally, the minimum distance between the soft palate and the posterior pharyngeal wall is measured from the post-processed segmentations.

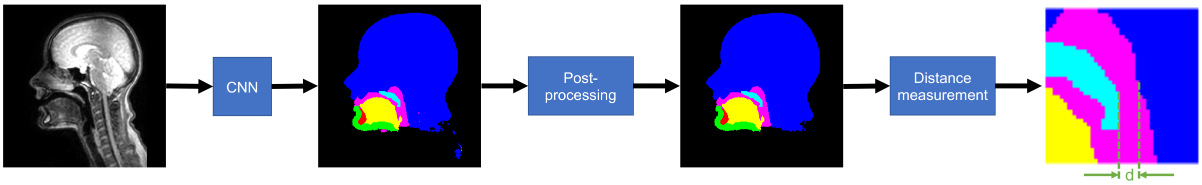

Figure 52: An overview of the proposed deep-learning-based segmentation method. The method consists of three steps: (1) a convolutional neural network (CNN) for estimating segmentations of seven different classes; (2) a post-processing step to remove anatomically impossible regions in the estimated segmentations; (3) further post-processing steps to measure the minimum distance between the soft palate and the posterior pharyngeal wall. The input to the method is a two-dimensional real-time magnetic resonance image of the vocal tract.

5.2.2 CNN Architecture, Implementation and Training

Segmentations were estimated using a CNN with a similar architecture to the original U-Net [175]. The CNN had a five-layer encoding path followed by a four-layer decoding path. More information on its architecture is provided in Figure 53. Dropout (introduced in section 3.2.4) with a probability of 0.5 was included in the fourth and fifth encoding layers. The outputs of the network were seven probability maps, one for each class. The network was implemented using PyTorch 1.4.0 [281] and training was performed on a NVIDIA TITAN RTX graphics card. Cross entropy was used as the loss function during network training. The

Adam optimiser [151] with hyperparameters β_1=0.9, β_2=0.999 and ε=1e-8 was used to adjust network weights. In each experiment, the network was trained for 200 epochs. Data augmentation (introduced in section 3.2.4) was performed to increase the number of images in the training dataset by a factor of four. Augmented images were created by randomly translating, rotating, cropping and rescaling the original images. Translation was by between -30 and 30 pixels in the x-direction and between 0 and -30 in the y-direction. Rotation was by between -10° and 30° clockwise. The reason for the asymmetric ranges of augmentation parameters was to avoid causing anatomically implausible artefacts in the image such as a gap between the base of the neck and the edge of the image. Image cropping was to a matrix size of either 220×220 if followed by rescaling or between 210×210 and 255×255 if followed by zero padding. All augmented images had the same matrix size as the original images. This was achieved by cropping and then zero padding the translated images and the rotated images, and rescaling or zero padding the cropped images.

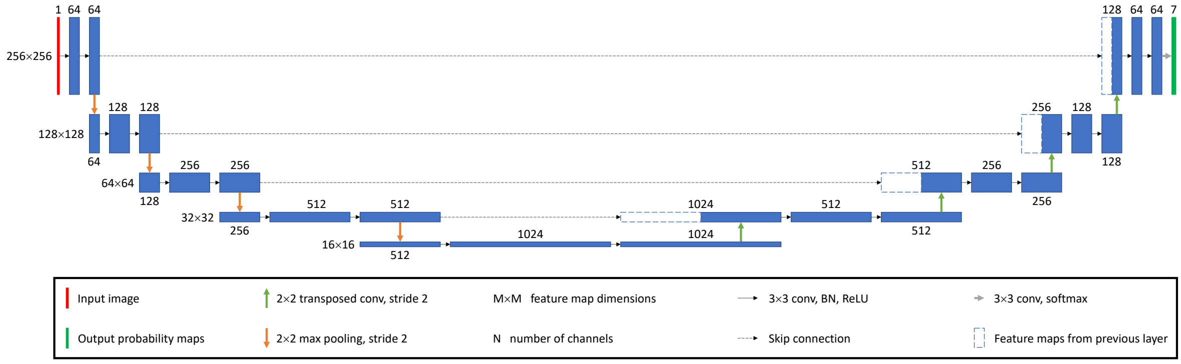

Figure 53: The architecture of the convolutional neural network of the proposed method [205]*. BN: batch normalisation, ReLU: rectified linear unit, conv: convolution.*

5.2.3 Loss Function Weighting

Use of training datasets with large imbalances in the number of pixels of each class is known to detrimentally affect the accuracy of CNNs [282,283]. A wide variety of approaches have been proposed to compensate for such imbalances [153,175,283–285], the majority of which involve weighting loss functions according to class frequency. To compensate for the class frequency imbalance in the training dataset, the loss function used to train the CNN (cross entropy) was weighted according to class frequency. More specifically, inspired by

[175,284,285], the losses of pixels of class $k \in \{1, 2, \dots, 7\}$ were multiplied by the following weight:

$$w_k = \frac{\sum_k N_k}{N_k} \tag{20}$$

where N_k is the number of pixels of class k in the training dataset. The motivation for compensating for the class frequency imbalance was to improve the accuracy with which the method segmented the soft palate.

Inspired by [175] and motivated by a desire to improve the accuracy with which velopharyngeal closures were captured in segmentations estimated by the method, an additional experiment was performed where the loss function was weighted according to both class pixel frequency and the minimum pixel distance from the nearest class boundary. In this experiment, the loss of a pixel was multiplied by both the weight in Equation (20) and the following weight:

$$w_b = \frac{1}{d^3} \tag{21}$$

where d is the minimum Euclidean distance of the pixel from the nearest boundary. The rationale for including a boundary distance weight was to encourage accurate segmentation of pixels at the boundaries between classes. This weight is similar to the one used in the training of the original U-Net [175].

5.2.4 Segmentation Post-Processing

At test time, connected-component-analysis-based post-processing was performed on each segmentation estimated by the CNN in order to remove anatomically impossible regions. More specifically, each region (i.e. connected component) in the estimated segmentation was automatically analysed in the following way:

1. The classes of the regions in contact with it were identified.
2. If the region was surrounded by another region, its class was changed to that of the surrounding region.

3. If the region was either in contact with an anatomically impossible region (for example, if a jaw region was in contact with a soft palate region) or not in contact with anatomically expected regions (for example, if a tooth space region was not in contact with a jaw region and a tongue region), the classes of the pixels surrounding the region were identified and the class of the region was changed to the mode class of these pixels. The rules for determining if a region was anatomically plausible are listed in Table 5.

This analysis was performed using MATLAB R2019b.

Table 5: The rules for determining the anatomical plausibility of a region in a segmentation. "Class" indicates the segmentation class, "Forbidden Contact" indicates segmentation classes that the region must not be in contact with to be anatomically plausible, while "Required Contact" indicates the segmentation classes that the region must be in contact with to be anatomically plausible.

Class	Forbidden Contact	Required Contact
Head	Tooth space	Soft palate, vocal tract
Soft palate	Jaw	Head, vocal tract
Jaw	Soft palate	Tongue, tooth space
Vocal tract	N/A	Head, soft palate, tongue
Tooth space	Soft palate	Jaw, tongue

5.2.5 Distance Measurements

Additional post-processing steps to automatically measure the minimum distance between the soft palate and the posterior pharyngeal wall in each estimated segmentation were performed. Figure 54 shows an overview of the steps. The steps were as follows and were implemented in MATLAB R2019b:

1. The coordinates of the centroid of the soft palate were identified.
2. Soft palate and head pixels with x-coordinates less than that of the centroid were removed.
3. For each head pixel, the Euclidean distance to the nearest soft palate pixel was calculated.
4. The minimum distance was identified and converted from pixel dimensions to mm by multiplying it by a scaling factor.

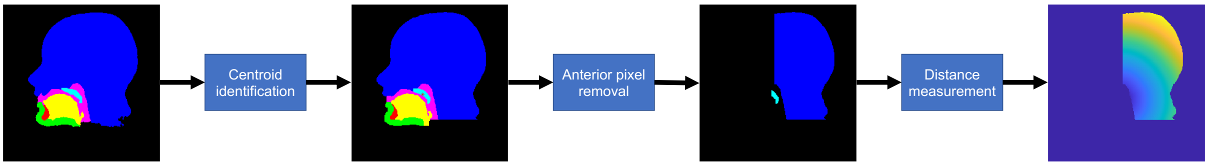

Figure 54: An overview of the post-processing steps to measure the minimum Euclidean distance between the soft palate and the posterior pharyngeal wall.

5.3 Experiments

5.3.1 Data

The five MR image series described in section 4.1.2 were used in the experiments, along with their corresponding GT segmentations described in section 4.1.4.

5.3.2 Segmentation Accuracy Assessment

The segmentation accuracy of the proposed method was assessed using two metrics. First, the DSC [194] was used to quantify the overlap between the GT segmentations and the segmentations estimated by the method. Second, the HD was used to quantify the maximum discrepancy between the boundaries of the GT and estimated segmentations.

The segmentation accuracy of the proposed method was also indirectly assessed by comparing the minimum distance between the soft palate and the posterior pharyngeal wall in corresponding estimated and GT segmentations. For each corresponding pair of segmentations, the absolute difference between the minimum distances was calculated:

$$d_{diff} = |d_{GT} - d_{estimated}| \tag{22}$$

where d_{GT} is the minimum distance in the GT segmentation, while $d_{estimated}$ is the minimum distance in the estimated segmentation. Minimum distances were calculated in the way described in section 5.2.5.

5.3.3 Velopharyngeal Closure Assessment

The accuracy with which the estimated segmentations showed velopharyngeal closures was assessed by manually comparing the closures in the GT and estimated segmentations. To enable the comparison, each segmentation in a series was visually inspected and labelled as either showing contact between the soft palate and posterior pharyngeal wall or not showing contact. Contact was defined as three or more soft palate pixels in contact with head pixels in the head class region corresponding to the posterior pharyngeal wall. Sequences of labels were then plotted (see Figure 55 for an example) and manually compared to identify the number of:

- “Correct” closures: closures that were shown in both the GT and estimated segmentations.
- “Additional” closures: closures that were shown in the estimated segmentations but not the GT segmentations.
- “Merged” closures: one or more consecutive closures that were shown as separate closures in the GT segmentations and a single closure in the estimated segmentations.
- “Missed” closures: closures that were shown in the GT segmentations but not in the estimate segmentations.

An example of each type of closure is shown in Figure 55.

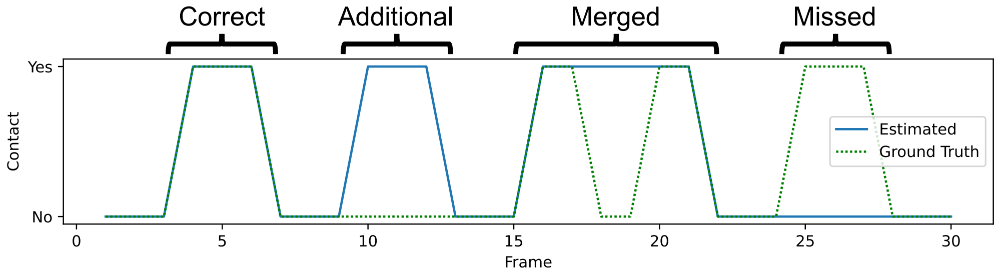

Figure 55: Examples of each type of velopharyngeal closure. On the y-axis, “Yes” indicates contact between the soft palate and posterior pharyngeal wall, while “No” indicates no contact.

5.3.4 Cross-Validation

To evaluate the generalisability of the proposed method, a five-fold cross-validation was performed with the dataset of each subject being left out once. Hyperparameter optimisation was achieved by carrying out a nested cross-validation for each fold of the main cross-validation, in the way described in [286]. This nested cross-validation was a four-fold cross-validation with the dataset of each of the remaining four subjects being left out once. Six different learning rate {0.003, 0.0003, 0.00003} and mini-batch size {4, 8} combinations were evaluated in this way, and the hyperparameter combination that resulted in the highest mean DSC on the left-out dataset (of the nested cross-validation) after post-processing was chosen as the optimal hyperparameter combination. Once the optimal hyperparameter combination had been identified for a fold of the main cross-validation, the CNN of the proposed method was trained using all the datasets except the left-out dataset for that fold, and then the entire method (including the post-processing steps) was tested using the left-out dataset.

5.3.5 Unseen Vocal Tract Shape Investigation

Different vocal tract shapes and articulator positions are required to produce different speech sounds. The data used to train the CNN of the proposed method does not contain images of all the different possible vocal tract shapes in English. To investigate the ability of the method to segment vocal tract shapes not present in the training dataset, 15 additional rtMR images were segmented using the method. The images (three per subject) were of the same five subjects described in 4.1.2 of this thesis producing three sounds which require vocal tract shapes not present in the training dataset: /ɒ/ and /b/ in “Bob” and /a/. The accuracy of the segmentations was assessed as described in section 5.3.2. The images were acquired and GT segmentations created in the ways described in sections 4.1.2 and 4.1.4 respectively.

5.4 Results

The hyperparameter combinations that resulted in the highest segmentation accuracy in the nested cross-validations are listed in Table 6.

Examples of segmentations estimated by the class frequency weighted version of the proposed method are shown in Figure 56. Figure 56A(2), Figure 56B(2) and Figure 56C(2) show estimated segmentations with relatively low, average and high DSCs respectively, while Figure 56D(2), Figure 56E(2) and Figure 56F(2) show estimated segmentations with relatively large, average and small HDs respectively. Column 3 in Figure 56 shows the estimated segmentations after post-processing.

Table 6: Optimal hyperparameter combinations. The 'Fold' column indicates the fold of the cross-validation, while the 'Loss Function Weighting' column indicates the way the loss function of the convolutional neural network of the proposed method was weighted during training. 'CF' indicates class frequency weighting, while 'CF and BD' indicates class frequency and boundary distance weighting.

Fold	Loss Function Weighting	Learning Rate	Mini-Batch Size
1	CF	0.0003	4
	CF and BD	0.0003	8
2	CF	0.0003	4
	CF and BD	0.00003	8
3	CF	0.0003	4
	CF and BD	0.00003	4
4	CF	0.0003	4
	CF and BD	0.0003	8
5	CF	0.003	8
	CF and BD	0.0003	4

Figure 57 shows the accuracy of the segmentations estimated by both versions of the proposed method. Figure 57A shows the DSCs of each class in the estimated segmentations, while Figure 57B shows the HDs of each class. The median DSC of the segmentations estimated by the version of the method where the loss function was weighted using class frequency only was 0.96, while the median HD was 5 mm. In 93% of segmentations (365 of 392 images in the test dataset), the DSCs of all the classes were above 0.85. The median DSC of the segmentations estimated by the version of the method where the loss function was weighted using both class frequency and boundary weighting was 0.95, while the median HD was 5 mm. In 86% of segmentations (339 of 392 images in the test dataset), the DSCs of all the classes were above 0.85.

Figure 58 shows the minimum distances between the soft palate and the posterior pharyngeal wall measured in the GT segmentations and corresponding segmentations estimated by both versions of the proposed method, while Figure 59 shows the absolute differences between the measurements in GT and corresponding estimated segmentation pairs.

The velopharyngeal closures in the GT and estimated segmentations of all the subjects are shown in Figure 60, while the number of each type of velopharyngeal closure (“correct”, “merged”, “additional” and “missed”) in the segmentations is summarised in Table 7. Figure 61 shows rtMR images whose estimated segmentations incorrectly showed velopharyngeal closure.

Five examples of segmentations estimated by the proposed method when it was inputted with additional rtMR images of vocal tract shapes that were not present in the training dataset are shown in Figure 62. The median DSC of the estimated segmentations of the 15 additional rtMR images was 0.96, while the median HD was 6 mm.

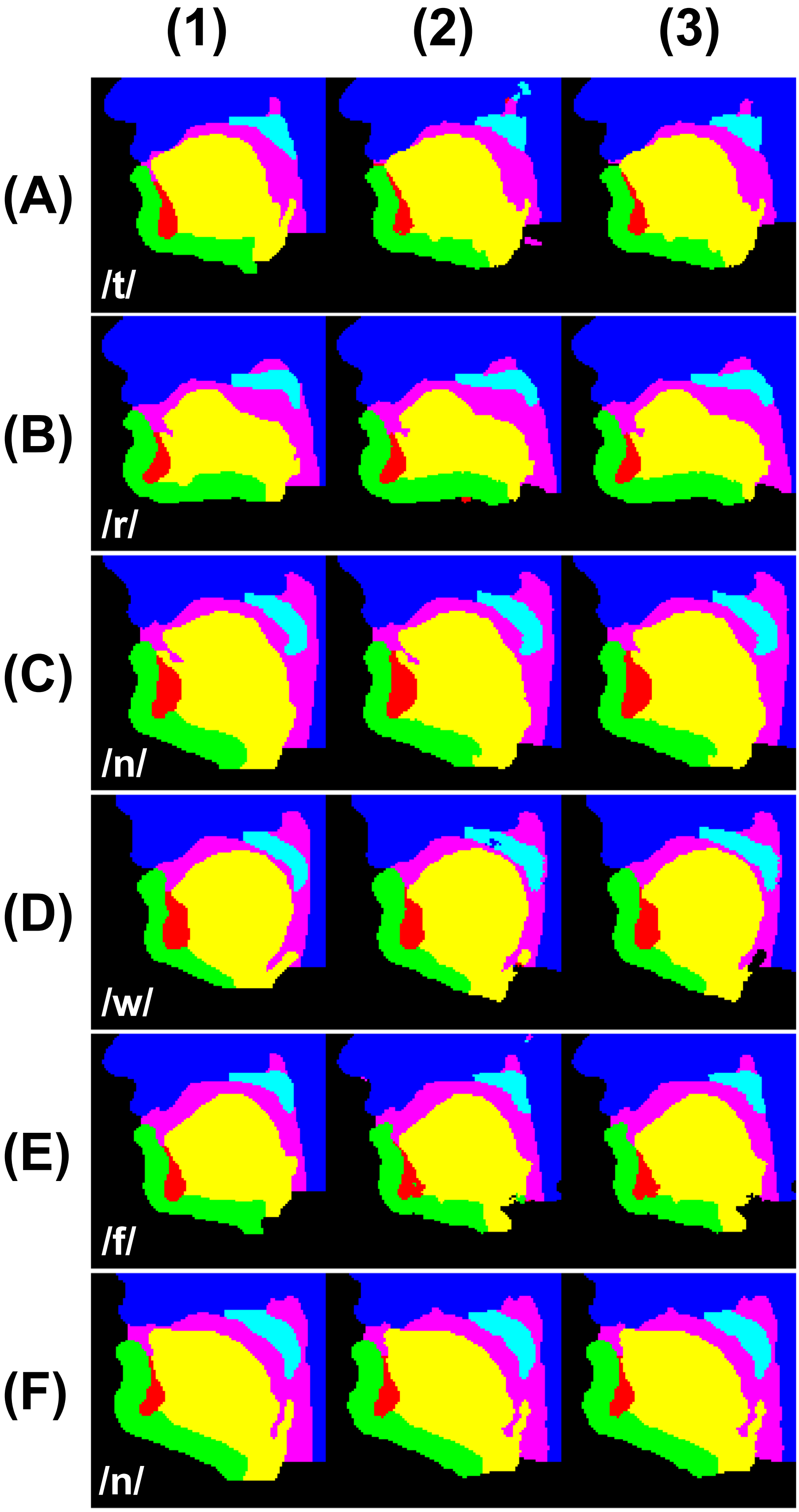

Figure 56: Examples of ground-truth segmentations (column 1) and corresponding segmentations estimated by the class frequency weighted version of the proposed method before and after the post-processing step (columns 2 and 3 respectively). Rows A to C show estimated segmentations with low, average and high Dice coefficients respectively. Rows D

to F show estimated segmentations with large, average and small general Hausdorff distances respectively. The sounds being produced by the subjects are /t/ in "two" (row A), /r/ in "three" (row B), /n/ at the end of "nine" (row C), /w/ in "one" (row D), /f/ in "four" (row E) and /n/ in "ten" (row F). The segmentations have been cropped to only show the vocal tract region. Image source: [205].

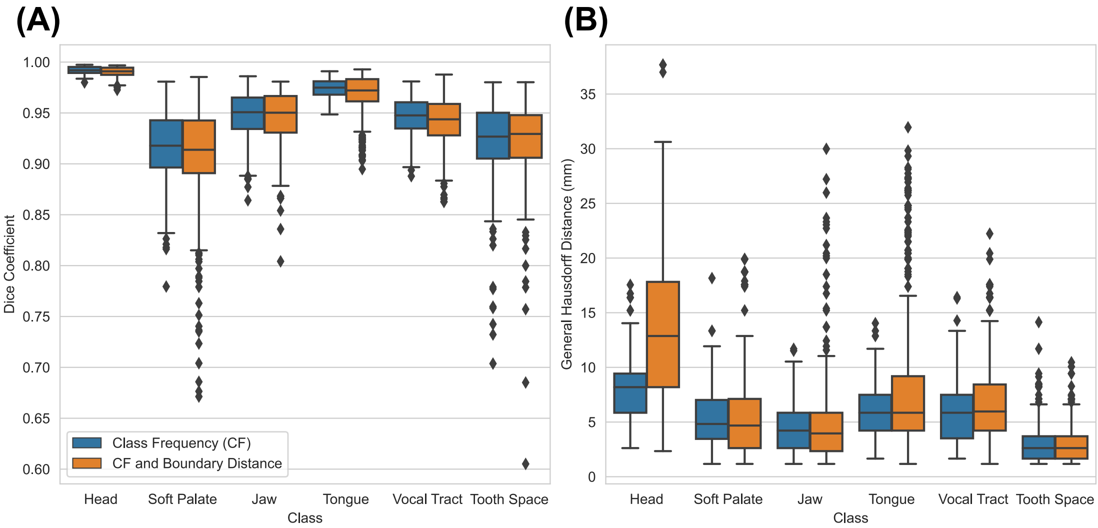

Figure 57: (A) Dice coefficients and (B) general Hausdorff distances of the segmentations estimated by both versions of the proposed method. In the Figure legend, 'Class Frequency (CF)' and 'CF and Boundary Distance' indicate the loss function weighting used during the training of the proposed method.

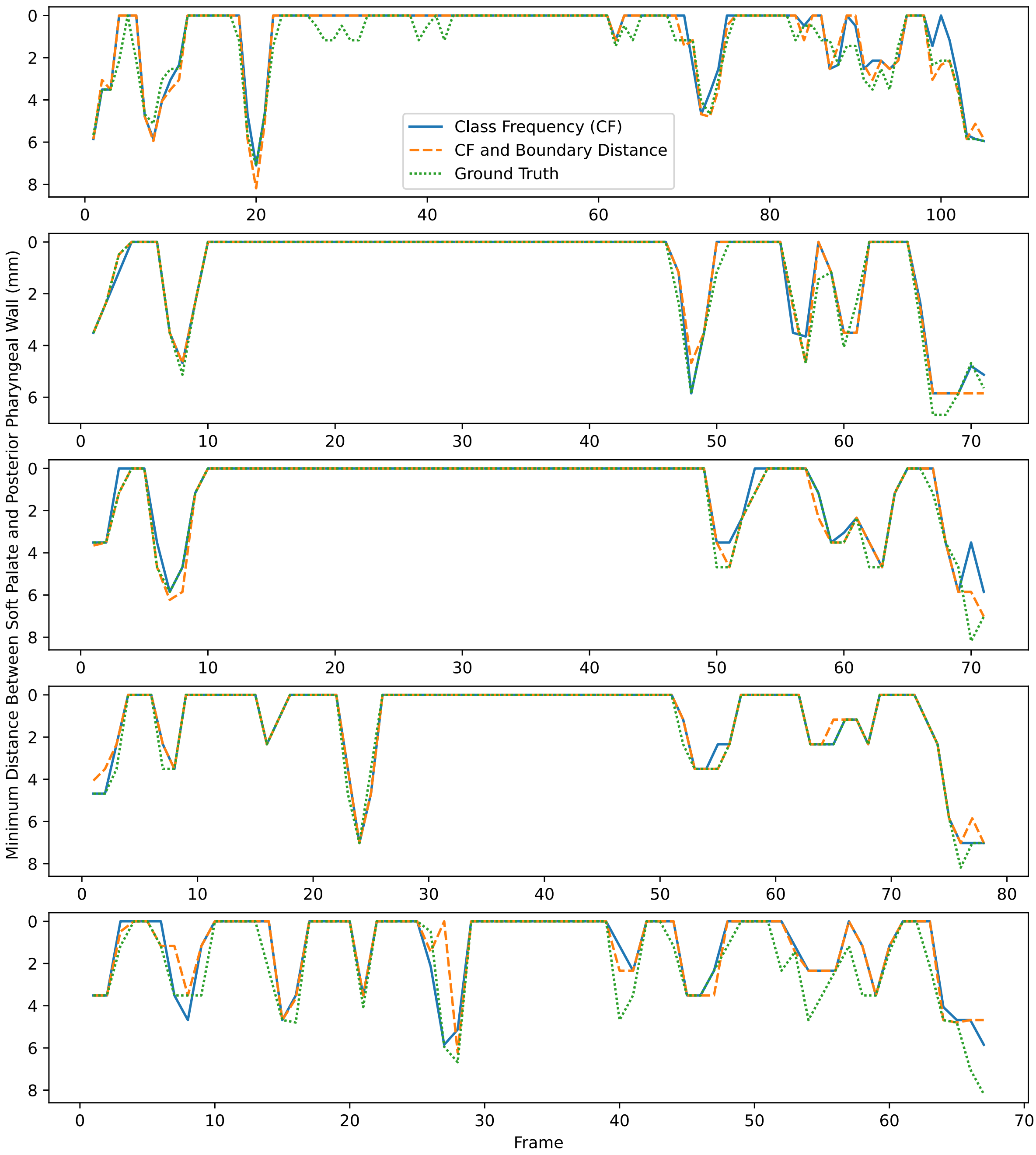

Figure 58: Minimum distances between the soft palate and posterior pharyngeal wall measured in the ground-truth segmentations and segmentations estimated by both versions of the proposed method. Each row corresponds to a different subject. In the Figure legend, 'Class Frequency (CF)' and 'CF and Boundary Distance' indicates the loss function weighting used during the training of the proposed method.

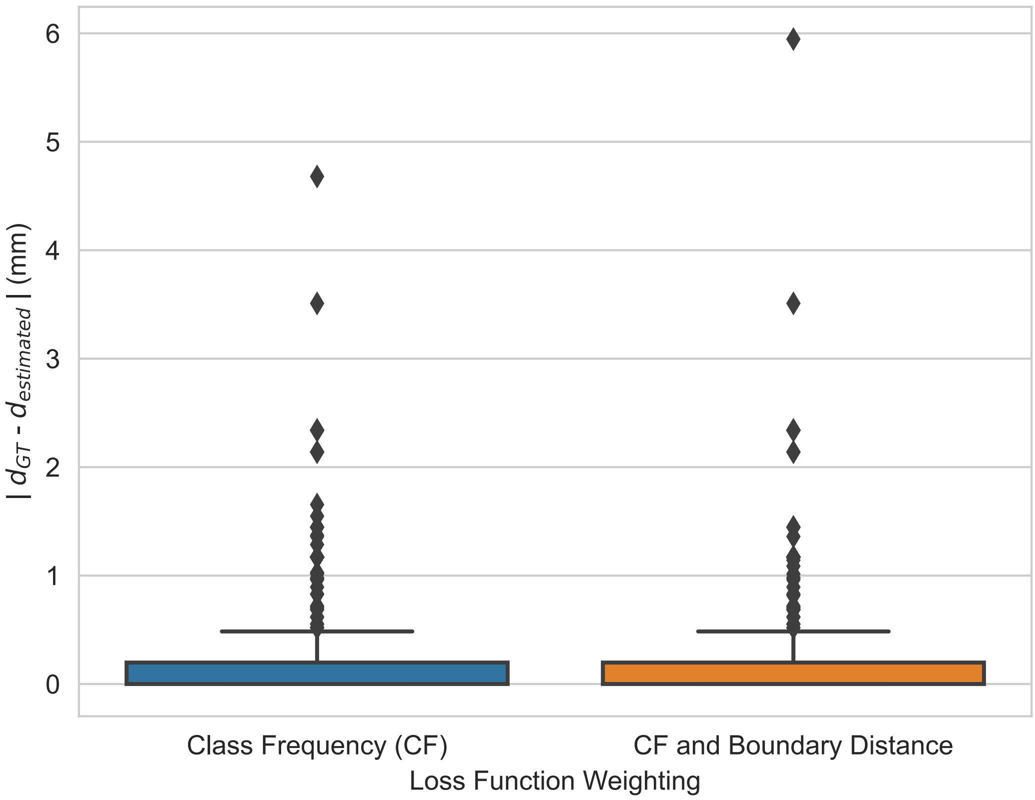

Figure 59: Absolute differences in the minimum distance (between the soft palate and posterior pharyngeal wall) measured in the ground-truth segmentations (d_{GT}) and corresponding segmentations estimated by both versions of the proposed method ($d_{estimated}$). The x-axis label 'Loss Function Weighting' indicates the loss function weighting used during the training of the proposed method.

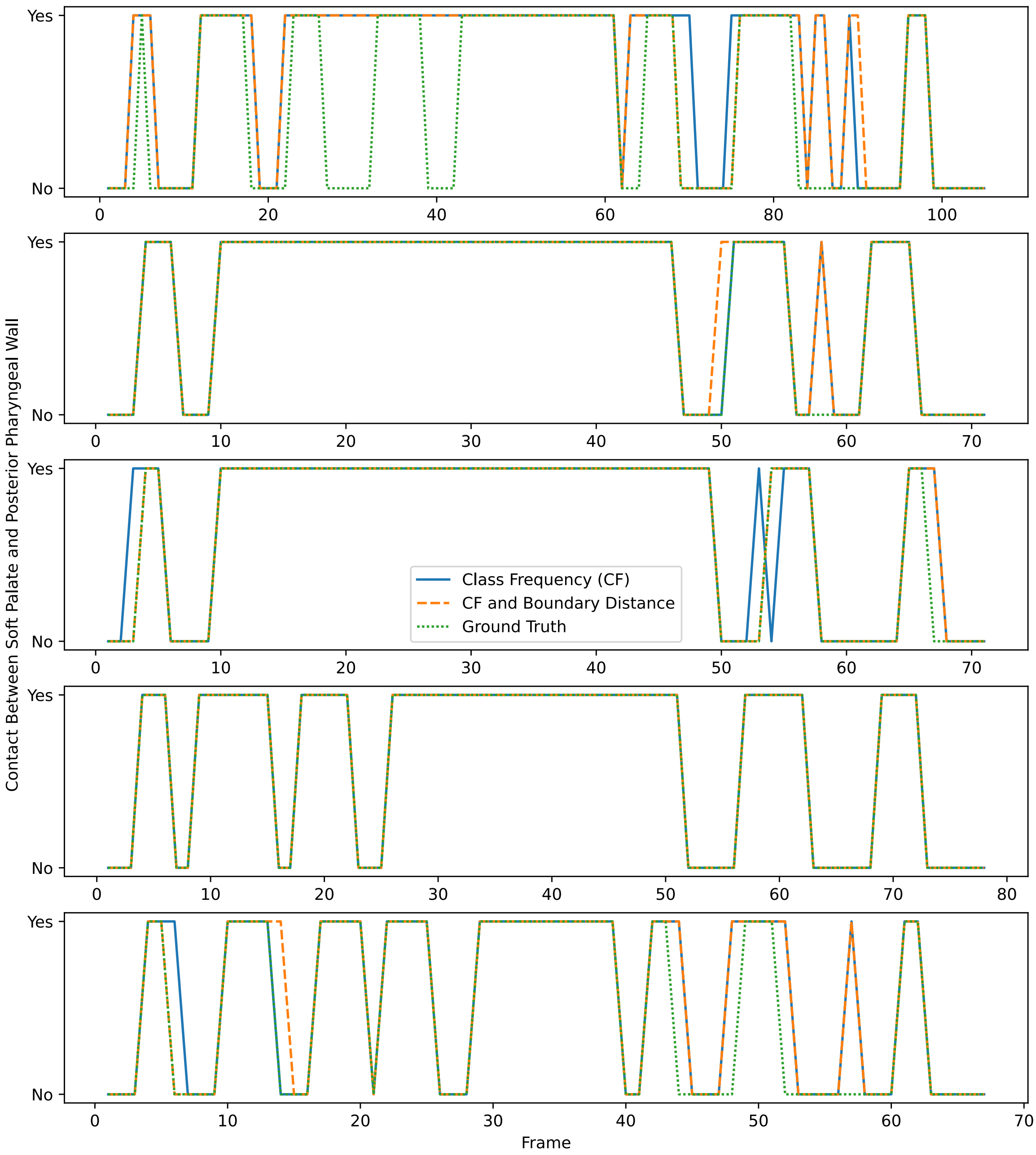

Figure 60: Velopharyngeal closures in the ground-truth segmentations and segmentations estimated by both versions of the proposed method. Each row corresponds to a different subject. In the Figure legend, 'Class Frequency (CF)' and 'CF and Boundary Distance' indicates the loss function weighting used during the training of the proposed method.

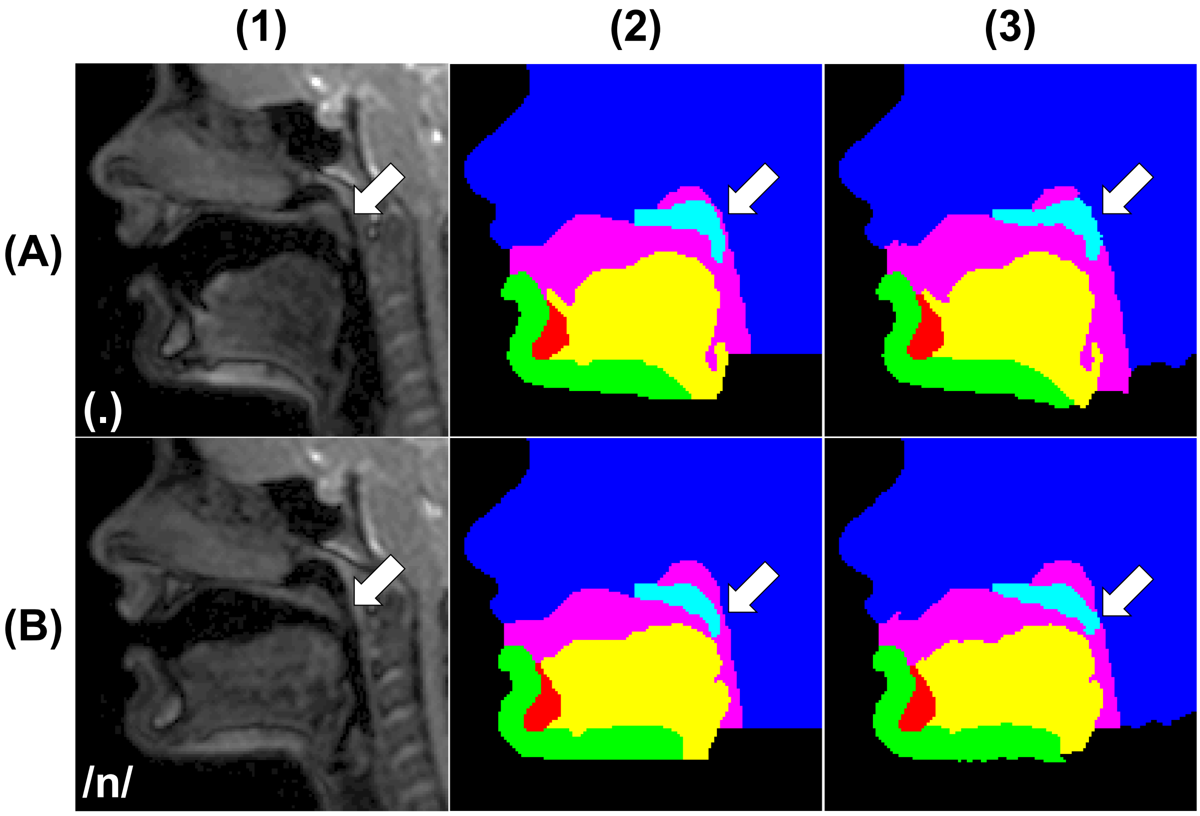

Figure 61: Magnetic resonance images (column 1) whose estimated segmentations after post-processing (column 3) incorrectly showed velopharyngeal closure. Column 2 is the ground-truth segmentation of the images. In both images, the soft palate is close to the posterior pharyngeal wall but not in contact with it. Row A shows the subject pausing between saying "four" and "five", while row B shows the subject producing the sound /n/ at the end of "nine". The images and segmentations have been cropped to only show the vocal tract region.

Table 7: Number of velopharyngeal closures in the ground-truth segmentations and segmentations estimated by the proposed method. Total: total number of closures in the segmentations. Correct: closures that were shown in both the ground-truth and estimated segmentations. Additional: closures that were shown in the estimated segmentations but not the ground-truth segmentations. Merged: one or more consecutive closures that were shown as separate closures in the ground-truth segmentations and a single closure in the estimated segmentations. Missed: closures that were shown in the ground-truth segmentations but not in the estimated segmentations. The columns 'Class Frequency' and 'Class Frequency and Boundary Distance' indicate the loss function weighting used during the training of the proposed method.

	Ground truth	**Class Frequency**	**Class Frequency and Boundary Distance**
Total	30	33	32
Correct	30	27	27
Additional	0	5	4
Merged	0	3	3
Missed	0	0	0

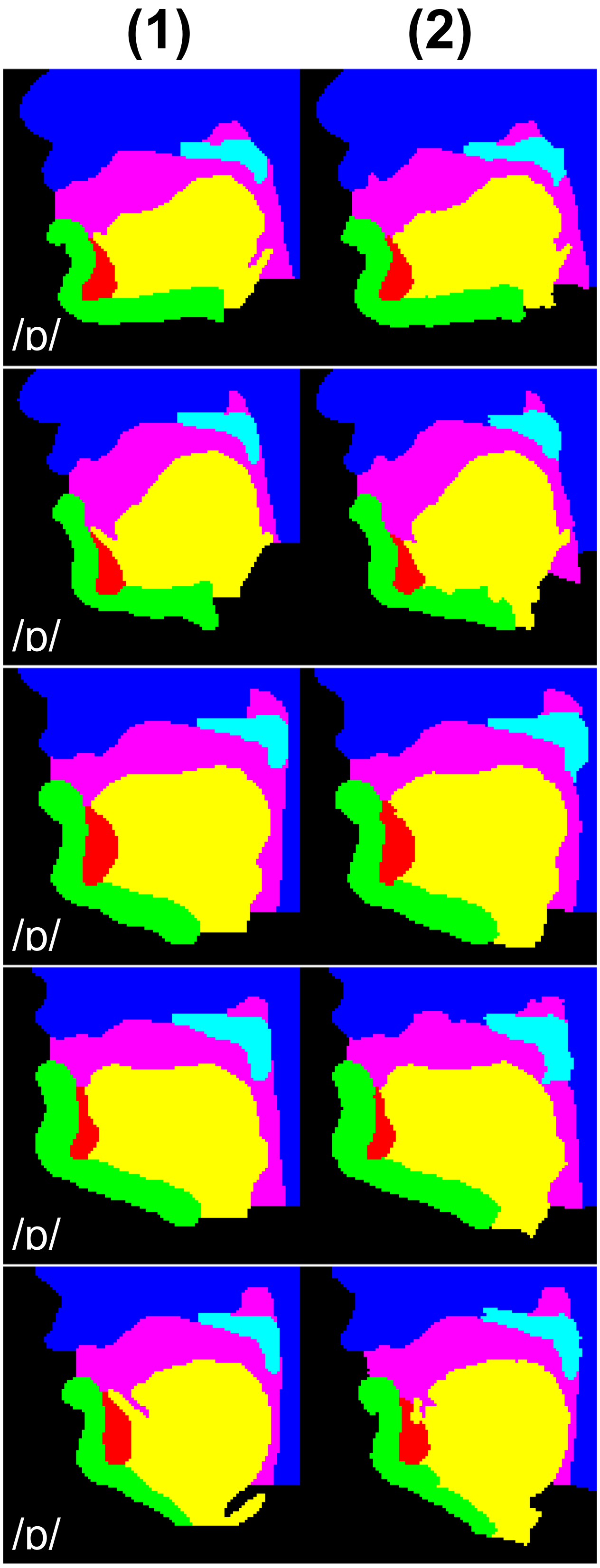

Figure 62: Ground-truth segmentations (column 1) and corresponding segmentations estimated by the proposed method (column 2) when inputted with images of vocal tract shapes that were not present in the training dataset. The segmentations have been cropped to only show the vocal tract region.

5.5 Discussion

At the time the work presented in this chapter was published [205], the main contribution and novelty of the work was the development of an automatic method to fully segment multiple groups of articulators as well as the vocal tract in 2D rtMR images of the vocal tract during speech. This novelty overcame the limitations of existing methods that either only segmented the air-tissue boundaries between the vocal tract and neighbouring tissues [15,16,199–204] or fully segmented the vocal tract only [208]. However, since the work was published three further related works have been completed. Two of these works are methods to analyse 2D rtMR images of the vocal tract during speech [72,206], while the other is a method to segment the pharynx and entire articulators (tongue and soft palate) in 2D static images of the vocal tract [218]. In [206] a DL-based method to segment entire articulators (soft palate, hard palate, tongue, jaw and upper lip) inspired by the method presented in this chapter was developed, while in [72] a method to measure the minimum distance between the soft palate and posterior pharyngeal wall was developed.

Another contribution and novelty of the work presented in this chapter is the development of a clinically relevant metric for assessing the accuracy of segmentations created by vocal tract and articulator segmentation methods. This novel metric was used to assess the accuracy of the segmentations estimated by the proposed method.

The final contribution and novelty of the work presented in this chapter is the extension of the method to enable automatic calculation of the minimum distance between the soft palate and the posterior pharyngeal wall, a measurement of particular interest in clinical speech assessment [7].

The proposed segmentation method is deep learning based and consists of three steps: first, segmentations are created by inputting rtMR images into a trained CNN with a similar architecture to the original U-Net [175]; second, a connected component analysis based post-processing is performed on the segmentations to remove anatomically impossible regions; third, the minimum distance between the soft palate and posterior pharyngeal wall is measured, a measurement that is of growing interest to clinicians managing patients with VPI [7,72,135,198,287,288]. This method is a step towards the ultimate goal of automatic articulator segmentation and measurement in clinical practice.

Two different CNN loss function weightings were investigated. The first was intended to compensate for the class pixel frequency imbalance in the dataset used to train the proposed method, while the second also included a weighting intended to prioritise accurate segmentation of boundary pixels. The two main motivations for this prioritisation were to increase the accuracy with which the proposed method captured velopharyngeal closures and the accuracy of the minimum distance (between the soft palate and posterior pharyngeal wall) measurements. However, as shown in Figure 57, Figure 58, Figure 59, Figure 60 and Table 7, the additional weighting did not improve the segmentation accuracy of the proposed method. A possible explanation for this result is that the additional weighting reduced the number of pixels that had a large effect on the loss function and the CNN struggled to learn to identify this smaller number of pixels.

The proposed method (the version trained without the additional loss function weighting) segmented each class with a high accuracy, as shown by its segmentations achieving a median DSC of 0.96 and a median HD of 5 mm. On average, the head was segmented most accurately (median DSC of 0.99). This result is unsurprising as this class has the largest number of pixels and the least variation in shape and position in the rtMR images. It is therefore the least challenging class for the proposed method to learn to segment. On average, the soft palate and tooth space were segmented least accurately (median DSCs of 0.92 and 0.93 respectively). This result is also unsurprising as these classes have the smallest number of pixels and so small errors at the boundaries will have a bigger impact on the DSC. In addition, the soft palate is the class with the largest variation in shape and position in the rtMR images. It is therefore the most challenging class for the proposed method to learn to segment.

The proposed method (the version trained without the additional loss function weighting) segmented the vocal tract with a higher accuracy (mean DSC of 0.95) than the only other published method for fully segmenting the vocal tract (mean DSC of 0.90) [208]. Both methods are deep learning based and have a similar architecture, therefore one possible reason for the higher accuracy of the proposed method is because it was trained using a larger number of images (up to 1625 images, while [208] was trained using 300 images). A possible reason for the lower accuracy of the other published method could be greater variability in the dataset used to train it. The training dataset of the other published method consisted of images of 10 healthy adult volunteers while the dataset of the proposed method consisted of images of five health adult volunteers. Another possible reason for the higher accuracy of the proposed method could be that the method captured more contextual information as a result of segmenting a larger number of classes, thus improving the accuracy with which it segmented the vocal tract.

The proposed method (the version trained without the additional loss function weighting) segmented the soft palate, jaw and tongue with greater median DSCs (0.92, 0.95 and 0.97 respectively) to the method published in [206], which segmented these articulators with median DSCs of 0.89, 0.91 and 0.97 respectively. In addition, the proposed method segmented the soft palate and tongue with a greater accuracy than the method proposed in [218], which segmented these articulators with a median DSCs of 0.79 and 0.89 respectively. However, the method proposed in [218] was developed for segmenting 2D static MR images of the vocal tract, rather than 2D rtMR images of the vocal tract during speech. All these methods are deep learning based and have a similar architecture, therefore possible reasons for the higher accuracy of the proposed method are similar to those discussed in the preceding paragraph. The proposed method was trained on up to 1625 images, while [206] was trained using 820 images and [218] was trained using 151 images.

In 93% of cases (365 of 392 images), the DSC of each of the six estimated segmentations (one per class) was 0.85 or above. This result suggests that the generalisability of the proposed method is good.

The proposed method includes steps to measure the minimum distance between the soft palate and posterior pharyngeal wall. As shown in Figure 59, the median error in the distances measured in the segmentations estimated was close to zero. Comparison of this error with the only other published method to measure the minimum distance between the soft palate and posterior pharyngeal wall [72] is not possible as no errors were reported. A limitation of the distance measurement steps is that they make large assumptions about the way the head is orientated in the images. For example, the steps assume that the head is facing towards the left of the images. To be suitable for use in clinical practice, the method should be able to identify images where the head is not orientated in this way, to avoid spurious results.

When clinically assessing the speech of patients with speech problems, an important consideration is whether velopharyngeal closure occurs during speech. It is therefore important that segmentation methods intended for use in clinical speech assessment accurately show any velopharyngeal closures that occur, while not artificially creating velopharyngeal closures when these do not occur (i.e. preserve gaps between the soft palate and posterior pharyngeal wall). The segmentations estimated by the proposed method (the version trained without the additional loss function weighting) correctly captured 90% (27 out of 30) of the velopharyngeal closures in the GT segmentations. As shown in Figure 60, three consecutive closures in the GT segmentations were shown as a single closure in the estimated segmentations. It is important to note that the soft palate motion between these three closures was different from the motion between all the other closures: instead of moving to a position far from the pharyngeal wall, the soft palate remained close to the wall (an example is shown in Figure 61A). Consequently, the gap between the soft palate and posterior pharyngeal wall remained small. The estimated segmentations also showed five closures that did not occur in the GT segmentations (two are shown in Figure 60). All five of these additional closures occurred when the soft palate was close to the posterior pharyngeal wall (an example is shown in Figure 61B). The merging of closures and the occurrence of additional closures shows that the proposed method was not always able to preserve small gaps between the soft palate and the pharyngeal wall. Further work is required to improve the ability of the method to preserve such gaps. A factor that can make preservation of such gaps particularly challenging is the presence of fluid within them. In rtMR images, fluid has a similar intensity to the soft palate and posterior pharyngeal wall and can therefore make it appear as though the soft palate and posterior pharyngeal wall are in contact (an example is shown in Figure 61B). This factor should be considered in any future work.

Different vocal tract shapes and articulator positions are required to produce different speech sounds. Our method was trained using 2D rtMR images of vocal tract shapes that occur in counting from one to ten (a speech task commonly performed in clinical speech assessment) rather than using images of all the different possible shapes in English. Nevertheless, the proposed method was able to segment rtMR images of three different vocal tract shapes that were not present in the training dataset with a similar accuracy to images of vocal tract shapes that were present in the training dataset. The median DSC and median HD of segmentations of the former images were 0.96 and 6 mm respectively, while those of the latter images were 0.96 and 5 mm. The similarity of these results suggests that the proposed method is able to segment images of vocal tract shapes that were not present in the training dataset with an accuracy similar to images of vocal tract shapes that were present in the training dataset. However, further work involving images of a larger range of vocal tract shapes is required to investigate the extent to which this finding is true.

The proposed method does not exploit the temporal nature of the image series. In other words, it segments images individually without considering prior or subsequent images in the series. Future work could investigate if exploiting the temporal nature of the image series, for example using RNNs, results in improved segmentation accuracy.

This work is a step towards the ultimate goal of automatic articulator segmentation and measurement in clinical practice. However, a large amount of future work is required to achieve this goal. More specifically, three major challenges must be overcome. One challenge concerns the dataset used to develop the method, while the other two are technical.

First, as explained in section 4.1.6, a larger and more diverse dataset, both in terms of subjects and image contrast, must be created and used to develop and extend the method. More specifically, a dataset more representative of the target patient population is required: since the target patient population primarily consists of children, the dataset must contain images of children. In addition, since velopharyngeal closure does not occur as expected in some of the speech of patients with VPI, the dataset must contain image series where velopharyngeal closure does not occur as well as image series where it does. In addition, a dataset with images acquired using many different MRI scanners and pulse sequences is required to ensure that methods developed using the dataset are generalisable and perform well on images from different sources. While there are publicly available 2D speech MRI dataset [18,19], these do not have corresponding GT segmentations thus limiting their use for training supervised DL-based segmentation methods.

Second, to be suitable for use in clinical practice, the method should be extended so that the shape and size of the soft palate and vocal tract are automatically measured in the images. While the segmentations estimated by the method are a useful step towards achieving this goal, further methods should be developed to automatically measure the size and shape of these segmentations.

Third, to be suitable for use in clinical practice, the method should be extended so that the motion of articulators can be automatically tracked. There is increasing interest in automatic quantification of articulator motion in 2D rtMR image series, for example to facilitate analysis of articulator motion before and after treatment in patients with VPI. As explained in section 3.4.1, an established way to automatically quantify complex motion in an image series is by using a nonlinear image registration method to estimate displacement fields between the images. Accurate displacement fields would enable clinical teams to obtain nearly automatically clinically relevant information such as the direction in which the soft palate elevates during speech, the speed at which it elevates and the distance by which it elevates. Future work should extend the method in partnership with clinical teams to ensure that the measured aspects of motion are clinically relevant.

5.6 Conclusions

A novel automatic method to fully segment multiple groups of articulators as well as the vocal tract in 2D rtMR images of the vocal tract during speech was developed. The method is a step towards the ultimate goal of automatic articulator and vocal tract segmentation and measurement in clinical practice.

At the time it was published [205], the method overcame the limitations of existing methods that either only segmented the air-tissue boundaries between the vocal tract and adjacent tissues or only fully segmented the vocal tract. Since then, three similar works have been developed [72,206,218], however, the proposed method remains the method that achieved the highest accuracy.

In addition to the novel method, a novel clinically relevant metric for assessing the accuracy of vocal tract and articulator segmentation methods was developed and used to assess the accuracy of the novel method.

The next chapter will present work that extends the proposed method to enable tracking of the motion of articulators in 2D rtMR images of the vocal tract during speech.

Chapter 6: Articulator Motion Quantification in MR Images of Speech

6.1 Introduction

6.1.1 Motivation

As explained in section 3.1.6, visualisation of the vocal tract and articulators during speech provides information about the size, shape, motion and position of these anatomical features during speech production. In a research context, primarily in speech science research, this information is desirable as it provides insights into speech production, while, as explained in section 2.2.3, in clinical speech assessment this information is desirable as it enables identification of the defect(s) preventing velopharyngeal closure and consequently informs treatment decisions [1,3,4].

As explained in section 3.1.7, use of rtMRI to visualise the vocal tract and articulators during speech is increasing due to the growing availability of MRI scanners, the development of rtMRI techniques for such visualisation, and the unique ability of MRI to non-invasively acquire images of any orientation without using ionising radiation [12,27,61]. While currently the main application of rtMRI of speech is in speech science research [68–76], the use of rtMRI in clinical speech assessment of patients with VPI is increasing [7,77–82].

Real-time MRI of speech typically involves acquiring series of 2D images of a midsagittal slice of the vocal tract [12,27]. There is increasing interest in automatic quantification of articulator motion in these series, for example to facilitate analysis of articulator motion before and after VPI treatment. An established way to automatically quantify complex motion in an image series is by using a nonlinear image registration method to estimate displacement fields between the images.

During speech, the articulators move in a complex manner. As well as changing shape and position, they come into contact and separate from each other and anatomical structures such as the pharyngeal wall. As described in section 2.2.3, the motion of the soft palate informs VPI treatment decision making in clinical speech assessment. Consequently, a key requirement of articulator motion quantification methods intended for use in clinical speech assessment is that the methods accurately capture soft palate motion. In particular, the methods must capture any velopharyngeal closures that occur.

6.1.2 Related Work

As described in section 3.4, traditional nonlinear registration methods establish nonlinear spatial correspondences (usually displacement vector fields) between two images by iteratively optimising a cost function [225]. Many different types of methods have been developed and used to register a wide variety of medical images [225]. Well-established methods include FFDs [236], demons [239], discrete methods [289] and their extensions such as [240] and [223]. Most traditional nonlinear registration methods are designed to estimate smooth and continuous displacement fields. However, such fields cannot accurately capture certain types of motion such as organs sliding past each other or organs coming into contact and then separating from each other. Instead, displacement fields with discontinuities are required to capture these types of motion. While several methods [235,259–262] have been developed to capture the former type of motion, only one of these [261] can capture the latter type. This method would be particularly suitable for capturing the motion of the articulators during speech, however, unfortunately there is no publicly available implementation of it.

Recently, inspired by the successes of DL-based methods in other medical image analysis tasks, researchers have developed DL-based nonlinear registration methods [228,229,243–246,290]. The latest methods [243–246,290] are unsupervised or weakly-supervised and consist of CNNs (introduced in section 3.2.6) for estimating displacement fields between images and spatial transformers [255] for transforming images and/or segmentations according to the estimated displacement fields. These methods have achieved state-of-the-art accuracy in the registration of MR images of organs including the heart [243,244] and brain [245,246,290].

Registration and segmentation can be related tasks, and there is increasing evidence that including segmentation information during the training of a registration CNN results in more accurate motion estimates [245,247,257,258,248–254,256]. The motivation for including segmentation information in the registration process is usually to provide information about the locations of boundaries between anatomical features in the images and also information about which anatomical features different regions of the displacement fields belong to. Inclusion of segmentation information is typically achieved by including region-overlap-based terms such as the DSC (introduced in section 3.3.2) in the CNN loss function. Joint registration and segmentation frameworks [247–250,252,257,258] have been developed as well as "segmentation-informed" registration frameworks such as VoxelMorph [245]. In fact, VoxelMorph can be trained in two ways: (i) using only the estimated displacement fields and the fixed and transformed moving images in the loss function, and (ii) in a segmentation-informed manner, where fixed and transformed moving segmentations are also used in the loss function.

Segmentation information has also been included in the registration process in two other ways. The first approach is to use segmentations to modify the appearance of the images, in order to optimise the images for the registration task [251,253,256]. In this approach, the images are modified before being used as inputs to the registration CNNs either by multiplying them by binary masks [251,253] or by using a fully convolutional image transformer network whose loss function includes a region-overlap-based term [256]. The second approach is to use segmentations as well as images as inputs to the registration CNN [254]. The rationale for inputting segmentations, even if these are estimates rather than ground-truths, is that they provide information about the positions of anatomical features in the images and would therefore help the CNN to estimate more accurate displacement fields.

Similarly to traditional nonlinear registration methods, currently the majority of DL-based methods are designed to estimate smooth and continuous displacement fields. Three methods have been developed to estimate displacement fields with discontinuities [253,257,267]. [267] is designed to capture sliding motion only, while [253] and [257] are designed to capture cardiac cycle motion and their suitability for capturing motion where organs come into contact and then separate from each other has not yet been investigated.

In previous work, only traditional registration methods have been applied to MR images of the vocal tract [20,21,69,72,75,76,196,268–270]. Rigid methods were used to correct for changes in head position in series of 2D rtMR images acquired during speech [69,72,75,76,196], while nonlinear methods were used to synthesise dynamic image series of speech [21,268–270], create dynamic 3D atlases of the vocal tract during speech [21] and estimate the speed at which the tongue tip moves during speech [20]. More specifically, the diffeomorphic demons method [240] was used in [268–270], the FFD method [236] was used in [21] and the method described in [271] was used in [20]. In [20,268,269], images where articulators were in contact were registered to images where they were not and vice versa. However, the authors did not evaluate if their chosen registration methods captured these changes in contact. In [268], the authors reported that the diffeomorphic demons method did not capture articulators coming into contact (for example, the lips coming into contact). Nevertheless, the authors used the same method in similar subsequent work [269]. In [20,269], the authors did not discuss if their chosen registration methods captured changes in articulator contact. Tongue tip speeds estimated using the nonlinear-registration-based method in [20] were found to be similar to those reported in the literature, suggesting that these methods can accurately estimate the speed at which articulators move during speech.

To accurately represent soft palate motion, displacement fields estimated by nonlinear registration methods must capture any velopharyngeal closures that occur. However, standard metrics such as region-overlap-based terms do not evaluate this. Accurate velopharyngeal closure capture is especially important for methods to analyse the soft palate motion of patients with VPI, as the presence or absence of velopharyngeal closures can affect treatment decisions [1].

In the previous chapter and [205], a metric based on velopharyngeal closure was proposed and used to evaluate the accuracy of a method to segment 2D rtMR images of the vocal tract during speech. This metric quantifies how many of the velopharyngeal closures in the GT segmentations occur in the estimated segmentations and is calculated by comparing corresponding consecutive segmentations in the two series. It could also be used to evaluate the accuracy of a registration method. In this case, the metric would be calculated by comparing the GT segmentations of the fixed images with the transformed GT segmentations of the moving images.

6.1.3 Contributions

The work presented in this chapter makes two contributions and has been peer reviewed and published as a journal article [291]. First, it begins to address the unmet need for automatic articular motion analysis in 2D rtMR images of the vocal tract during speech, by developing a segmentation-informed nonlinear registration framework to estimate articulator-specific displacement fields between these images. This is the first time that segmentation-informed registration has been used for this application. Second, the work uses for the first time a metric based on a quantifiable and clinically relevant aspect of articulator motion (velopharyngeal closure) to evaluate the accuracy of these displacement fields.

The work builds on the work presented in the previous chapter in the following ways. First, it uses the DL-based segmentation method presented in the previous chapter to provide the segmentations used as inputs to the registration CNN. Second, it uses the velopharyngeal closure evaluation metric proposed in the previous chapter to evaluate if the displacement fields estimated by the proposed segmentation-informed nonlinear registration method accurately capture velopharyngeal closures.

6.2 Methods

6.2.1 Proposed Registration Framework

Figure 63 shows an overview of the proposed framework. Given a pair of images from a series of 2D rtMR images of the vocal tract, the framework will estimate a displacement field to align the moving image to the fixed image. The framework is based on the segmentation-informed VoxelMorph framework [245] but features two adaptations. First, it includes a method to segment the images. Second, segmentations as well as images are used as inputs to the registration CNN, in the same manner as the framework of Chen et al. [254]. (In the segmentation-informed VoxelMorph, segmentations are only used to compute part of the loss function during training.) Figure 64 shows the architecture of the registration CNN. The segmentation method included in the framework is the DL-based method presented in the previous chapter and [205] to segment 2D rtMR images of the vocal tract. This method segments six anatomical features in each image. All six segmentations are used as inputs to the registration CNN. The registration CNN therefore has 14 input channels (two for the 2D fixed and moving images, 12 for the 2D fixed and moving segmentations), while the registration CNN of VoxelMorph only has two (for the fixed and moving images).

Like the VoxelMorph frameworks, the proposed framework includes a spatial transformer to transform an image or segmentation according to an estimated displacement field. The spatial transformer is required for framework training and evaluation, but not for framework deployment.

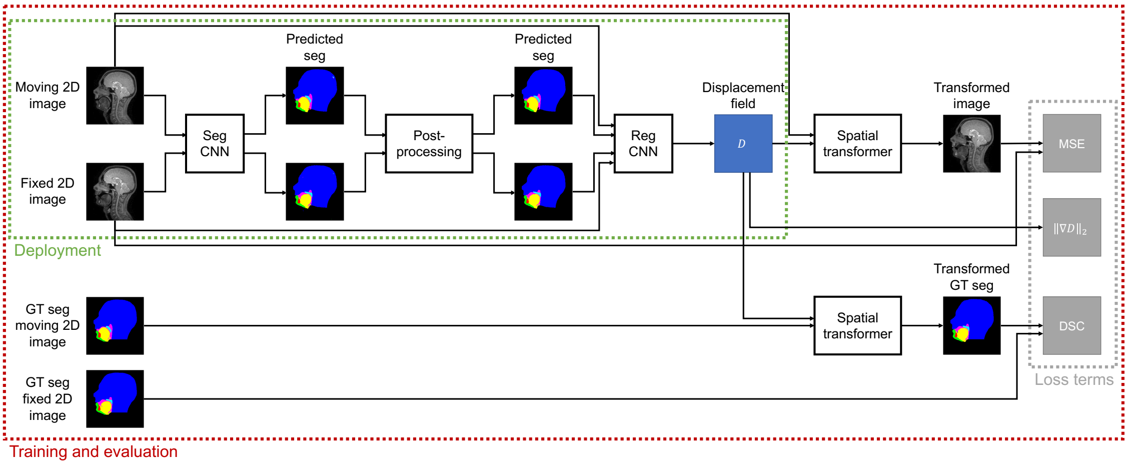

Figure 63: An overview of the proposed framework for segmentation-informed nonlinear registration. A pair of two-dimensional (2D) real-time magnetic resonance images of the vocal tract pass through the framework as follows. First, the image pair are used as inputs to a convolutional neural network (CNN) which estimates segmentations of six different anatomical features in the images. Second, the segmentations are post-processed to remove anatomically impossible regions. Third, the image pair and post-processed segmentations are used as inputs to a registration CNN which estimates a displacement field to align the moving image to the fixed image. Fourth, the moving image and displacement field are used as inputs to a spatial transformer to transform the moving image. During training and evaluation, the spatial transformer is also used to transform the ground-truth (GT) segmentations of the moving image. The red boundary contains the parts of the framework used during training and evaluation, while the green boundary contains the parts used during deployment. The grey boundary contains the terms in the loss function used to train the framework. Image source: [291].

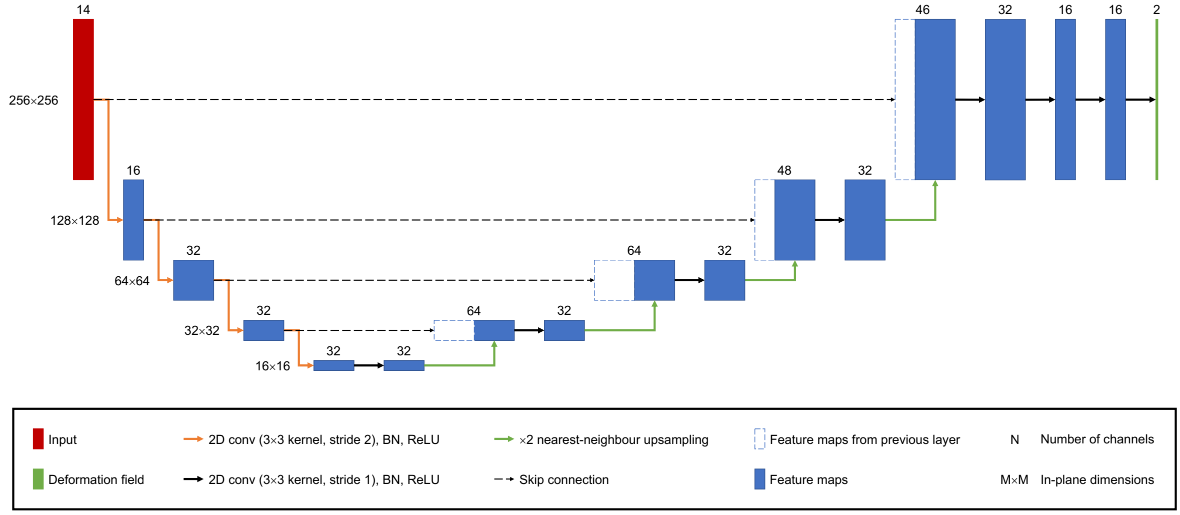

Figure 64: The architecture of the registration convolutional neural network in the proposed framework (i.e. the Reg CNN box in Figure 63). When input with a pair of two-dimensional (2D) real-time magnetic resonance images of the vocal tract and segmentations of six different anatomical features in the pair, the network estimates a displacement field to align one of the images to the other. The network has 14 input channels: two for the image pair, six for the segmentations of the fixed

image and six more for the segmentations of the moving image. The network output has 2 channels: one for displacements in the x-direction and another for displacements in the y-direction. The outputs of each 2D convolution (conv) are batch normalised. Following batch normalisation (BN), the outputs are passed through a rectified linear unit (ReLU). Image source: [291].

6.2.2 Framework Implementation, Training and Evaluation

The segmentation method included in the framework had been trained separately in the way described in the previous chapter and [205]. The registration framework was trained using the same training/validation/test dataset split as the segmentation method. The framework was implemented in PyTorch 1.7.1 [281] and trained for 200 epochs. In each epoch, every image in the training dataset was used once as the fixed image. Each fixed image was randomly paired with another image of the same subject. Each mini-batch consisted of four image pairs. Segmentations of anatomical features in these images were estimated using the segmentation method. The images and estimated segmentations were then used as inputs to the registration CNN. During training and evaluation, GT segmentations of the images were transformed according to the displacement fields estimated by the registration CNN. The Adam optimiser [151] with β_1=0.9, β_2=0.999 and ε=1e-8 was used during training. Data augmentation consisting of random translations, rotations, cropping and rescaling was performed to increase the size of the training dataset by a factor of four. More information about the augmentations is provided in 5.2.2 of this thesis and section 2.3 of [205]. During framework evaluation, every image in the test dataset was used as the fixed image. Each image was paired with the reference image of the dataset.

6.2.3 Loss Function

The proposed framework was trained using the same loss function as the segmentation-informed VoxelMorph framework. This loss function consisted of three terms: an MSE term; an L_2 regularisation of the spatial gradients of the displacement field (D) term and a DSC term. MSE was used to quantify differences in image appearance:

$$MSE = \frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y}(F(x,y) - M_T(x,y))^2 \qquad (23)$$

where $F(x,y)$ is the intensity of pixel (x,y) in the fixed image, $M_T(x,y)$ is the intensity of pixel (x,y) in the transformed moving image, and $X \times Y$ is the image matrix size.

Spatial gradients of the displacement field were approximated using differences between neighbouring pixels:

$$\nabla D(x,y) \approx (D(x+1,y) - D(x,y), D(x,y+1) - D(x,y)) \tag{24}$$

The L₂ regularisation term was therefore calculated using the following equation:

$$\|\nabla D\|_2 = \frac{1}{2}\left(\frac{1}{(X-1)}\frac{1}{Y}\sum_{x=1}^{X-1}\sum_{y=1}^{Y}(D(x+1,y) - D(x,y))^2 + \frac{1}{X}\frac{1}{(Y-1)}\sum_{x=1}^{X}\sum_{y=1}^{Y-1}(D(x,y+1) - D(x,y))^2\right) \tag{25}$$

The full loss function was:

$$L = MSE + \lambda\|\nabla D\|_2 - \gamma DSC \tag{26}$$

where λ and γ are loss weighting terms. Section 6.3.4 provides details on how these hyperparameters were optimised.

6.3 Experiments

6.3.1 Data

The five rtMR image series described in section 4.1.2 were used in the experiments, along with their corresponding GT segmentations described in section 4.1.4. In each series, the first image that met the following criteria was manually chosen as the reference image:

1. Upper and lower lips not in contact.
2. Tongue not in contact with roof of mouth or soft palate.
3. Soft palate not in contact with pharyngeal wall.

Figure 65 shows the reference images. During framework evaluation, these images were used as the moving image for registration purposes.

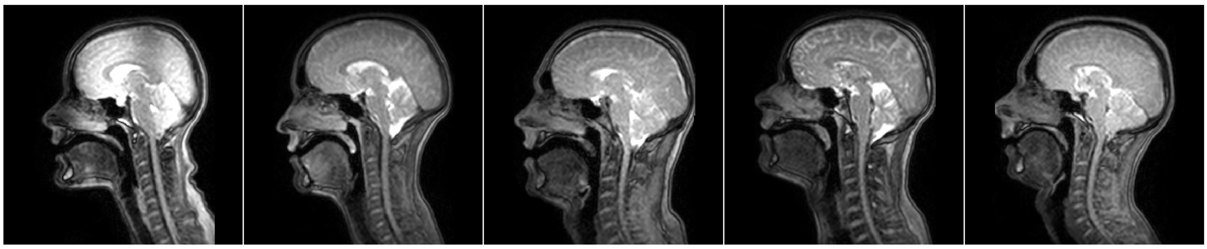

Figure 65: The reference image in each of the five series of two-dimensional real-time magnetic resonance images (image from [291]*). During framework evaluation, these images were used as the moving image for registration purposes.*

6.3.2 Displacement Field Evaluation

Displacement fields estimated by the framework were evaluated by first transforming moving GT segmentations according to the fields and then comparing these with fixed GT segmentations using the three metrics described below.

6.3.2.1 Dice Coefficient and Average Symmetric Surface Distance

The DSC (introduced in section 3.3.2) was used to quantify the overlap of corresponding classes in the fixed and transformed moving GT segmentations, while the ASD (also introduced in section 3.3.2) was used to quantify the average discrepancy between pixels at the surfaces of corresponding classes. Six values of each metric were calculated per moving segmentation: one value per class. The metrics were calculated using the DiceMetric and SurfaceDistanceMetric functions from MONAI 0.9.0 [292].

6.3.2.2 True Velopharyngeal Closures

The third metric (introduced in section 5.3.3) evaluates if velopharyngeal closures are captured by the displacement fields. The number of true velopharyngeal closures captured by the displacement fields was calculated in the following way. First, transformed moving GT segmentations were automatically labelled as showing velopharyngeal closure or not. This enabled the velopharyngeal closures in a series of segmentations to be represented as a series of binary values (one for each frame) with zero indicating no velopharyngeal closure and one indicating velopharyngeal closure. Second, the binary series of the fixed and transformed moving GT segmentations was automatically compared. A velopharyngeal closure was considered to be captured correctly if a series of ones in both binary series overlapped. The software to label segmentations and create and compare binary series was developed in-house and implemented using MATLAB 2019b (MathWorks, Natick, MA). The software determined if a segmentation frame showed velopharyngeal closure by identifying if three or more posterior “soft palate” pixels in the frame were in contact with “head” pixels.

6.3.3 Comparison with State-of-the-Art Methods and Frameworks

The proposed framework was benchmarked against five current state-of-the-art deformable registration methods and frameworks: two traditional methods and three frameworks. The traditional methods were FFD [236] and a segmentation-informed version of FFD (SIFFD) where deformations in certain regions of the moving image are constrained to be rigid [293]. The frameworks were the VoxelMorph (VXM) and segmentation-informed VXM (SIVXM) frameworks [245] and a joint registration and segmentation (JRS) framework [247]. Benchmarking was performed by comparing estimated displacement fields using the two metrics described in section 6.3.2.

6.3.3.1 Free-Form Deformation Methods

Both FFD methods were implemented using NiftyReg version 1.5.39 [237]. The cost function consisted of three terms: a normalised mutual information term (NMI); a bending energy (BE) term and a term based on the symmetric and anti-symmetric parts of the Jacobian (LE) [237]. The full cost function was:

$$C = (1 - \lambda - \gamma)NMI - \lambda BE - \gamma LE \tag{27}$$

where λ and γ are cost weighting terms.

Three iteration levels were used in the optimisation of the cost function, with a maximum of 150 iterations in the final level. In SIFFD, deformations in the region of the image corresponding to the head segmentation estimated by the segmentation method were constrained to be rigid. While it may seem counterintuitive to use rigid constraints, the reason for using these was to prevent the pharyngeal wall (part of the head segmentation class) from being misregistered to the soft palate.

For both methods, several registrations were performed using different combinations of cost weighting term values and spline grid spacings (listed in Table 8), and then evaluated using the metrics described in section 6.3.2, enabling identification of the optimal values and spacings.

6.3.3.2 VoxelMorph Frameworks

The two VoxelMorph frameworks are almost identical; the only difference between them is the loss function used to train them. The SIVXM framework is trained using L (see Equation 26), while the VXM framework is trained using a loss function consisting of two of the three terms in L:

$$L_{VXM} = MSE + \lambda \|\nabla D\|_2 \tag{28}$$

The key difference between L and L_{VXM} is that the former contains a segmentation-dependent term (DSC). Use of L during training therefore results in a segmentation-informed registration framework, while use of L_{VXM} does not. The frameworks were implemented in PyTorch 1.7.1 using the code publicly available at https://github.com/voxelmorph/voxelmorph. Framework training and evaluation was performed as described in section 6.2.2.

6.3.3.3 Joint Image Registration and Segmentation Framework

This framework was implemented in PyTorch 1.7.1 using the code publicly available at https://github.com/cq615/Joint-Motion-Estimation-and-Segmentation. The framework was trained in three stages using three different loss functions, as described in section 2.2 of [247], and for 200 epochs in total. First, the registration CNN was trained for 67 epochs using L_{VXM} (see Equation 28) as the loss function. Second, the segmentation CNN was trained for 67 epochs using cross entropy (CE_{est_seg}) (introduced in section 3.3.2) as the loss function. CE_{est_seg} was calculated by comparing the segmentations estimated by the segmentation CNN to the GT segmentations. Third, both CNNs were jointly trained for 66 epochs using a combination of L_{VXM}, CE_{est_seg} and an additional cross entropy term (CE_{tra_gt}) as the loss function. CE_{tra_gt} was calculated by comparing the fixed and transformed moving GT segmentations. The full loss function was:

$$L_{JRS} = MSE + \lambda \|\nabla D\|_2 + \gamma_1 CE_{est\_seg} + \gamma_2 CE_{tra\_gt} \tag{29}$$

where γ_1 and γ_2 are loss weighting terms. All other aspects of framework training and evaluation were performed as described in section 6.2.2.

6.3.4 Five-Fold Cross-Validation

Four separate five-fold cross-validations were carried out to evaluate the generalisability of the VXM, SIVX, JRS and proposed frameworks respectively. Each cross-validation was carried out as follows. A different image series was left out in each fold. Hyperparameter optimisation was performed as part of the cross-validation, by carrying out a nested cross-validation for each main cross-validation fold. The nested cross-validations were four-fold cross-validations where each of the remaining four image series were left out once. In each nested cross-validation fold, combinations of learning rates and loss term weightings (listed in Table 8) were evaluated. The optimal hyperparameter combination was identified by comparing the number of true velopharyngeal closures captured by the displacement fields estimated for the left-out image series of the nested cross-validation. The combination that resulted in the capture of the largest number of true velopharyngeal closures was chosen as the optimal hyperparameter combination. Once the optimal combination had been identified for a main cross-validation fold, these hyperparameters were used to train the framework. In each main cross-validation fold, the framework was trained using all the image series except the left-out image series for that fold, and then evaluated using the left-out image series.

6.3.5 Ablation Study

Although the segmentations consist of six classes, only the head, soft palate and vocal tract classes are required to determine if there is velopharyngeal closure. An ablation study was performed to investigate the effect of these three classes on the accuracy of the proposed framework. Three experiments were performed where different classes were used as inputs to the registration CNN during the training and evaluation of the framework. In the first, only the soft palate and vocal tract classes were used as inputs. In the second, the head, soft palate and vocal tract classes were used. In the third, all classes except the soft palate and vocal tract were used. In all other respects, the framework was trained and evaluated in the way described in sections 6.2.2, 6.3.2 and 6.3.4.

Table 8: Cost weighting terms, spline grid spacings (GSs in pixels) and hyperparameter combinations that were evaluated. Cost weighting terms (λ, γ) and spline GSs were evaluated when optimising the free-form deformations (FFD) method and segmentation-informed FFD (SIFFD) method. Hyperparameter combinations were evaluated during hyperparameter optimisation of the VoxelMorph (VXM), segmentation-informed VXM (SIVXM), joint registration and segmentation (JRS) and proposed (Proposed) frameworks. N_c indicates the number of combinations. Eight or more combinations of learning rate (LR) and loss weighting terms (λ, γ, γ_1 and γ_2) were evaluated per framework.

Framework	N_c	LR	λ	γ	γ_1	γ_2	GS
FFD & SIFFD	12		{0, 0.001}	{0, 0.01}			{4, 5, 6}
VXM	9	{0.00009, 0.0003, 0.0009}	{0.001, 0.01, 0.1}				
SIVXM	8	{0.0003, 0.0009}	{0.001, 0.01}	{0.1, 1}			
JRS	16	{0.0003, 0.0009}	{0.001, 0.01}		{0.1, 1}	{0.1, 1}	
Proposed	8	{0.0003, 0.0009}	{0.001, 0.01}	{0.1, 1}			

6.3.6 Statistical Tests

Normality of DSC and ASD groups was assessed using a Chi-squared goodness-of-fit test. No groups were found to be normally distributed using a 5% significance level. Groups of DSCs were compared using either a two-tailed Wilcoxon signed-rank test or a two-tailed sign test, depending on whether the distribution of differences between paired data points was symmetric. Groups of ASDs were compared in the same way as groups of DSCs. Numbers of true velopharyngeal closures were compared using McNemar's test. A 5% significance level was used for all tests, corrected using the Holm-Bonferroni method to compensate for multiple comparisons. All statistical tests were performed using MATLAB 2019b (MathWorks, Natick, MA).

6.4 Results

6.4.1 Optimal Parameters

Table 9 lists the optimal parameters for the FFD methods, while Table 10 lists the optimal hyperparameters for training each framework.

Table 9: Optimal parameters for the free-form deformations (FFD) method and segmentation-informed FFD (SIFFD) method. GS indicates spline grid spacing in pixels, while λ and γ are cost weighting terms.

Method	**Subject**	$\boldsymbol{\lambda}$	$\boldsymbol{\gamma}$	**GS**
FFD	1	0.001	0	5
	2	0	0.01	6
	3	0	0	4
	4	0.001	0	6
	5	0	0	6
SIFFD	1	0	0	4
	2	0	0.01	4
	3	0	0.01	4
	4	0.001	0	4
	5	0	0	5

Table 10: Hyperparameters identified as being optimal during hyperparameter optimisation of the VoxelMorph (VXM), segmentation-informed VXM (SIVXM), joint registration and segmentation (JRS) and proposed (Proposed) frameworks. LR indicates learning rate and CV fold indicates the fold of the cross-validation, while λ, γ, γ_1 and γ_2 are loss weighting terms.

Framework	CV fold	LR	λ	γ	γ_1	γ_2
VXM	1	0.0009	0.001			
	2	0.0009	0.001			
	3	0.0009	0.001			
	4	0.0003	0.001			
	5	0.0009	0.001			
SIVXM	1	0.0009	0.001	1		
	2	0.0009	0.001	1		
	3	0.0009	0.001	1		
	4	0.0009	0.001	1		
	5	0.0003	0.001	1		
JRS	1	0.0003	0.001		1	1
	2	0.0003	0.01		1	1
	3	0.0003	0.001		1	0.1
	4	0.0003	0.001		1	1
	5	0.0003	0.001		1	1
Proposed	1	0.0009	0.001	0.1		
	2	0.0009	0.01	1		
	3	0.0003	0.01	1		
	4	0.0009	0.001	1		
	5	0.0009	0.01	1		

6.4.2 Example Images and Segmentations

Figure 66 shows example transformed images and GT segmentations output by each of the methods and frameworks. In Figure 66, the fixed images are consecutive images from one of the image series and show a velopharyngeal closure. This closure is captured by the proposed framework: contact between the soft palate and pharyngeal wall is shown in three of the transformed images and segmentations. However, the closure is not captured by the

FFD methods or the VXM framework: none of the transformed images or segmentations show contact between the soft palate and the pharyngeal wall. The closure is partially captured by the SIVXM and JRS frameworks: two of the transformed images and segmentations output by the former framework show contact between the soft palate and the pharyngeal wall, while one of the transformed images and segmentations output by the latter framework shows such contact.

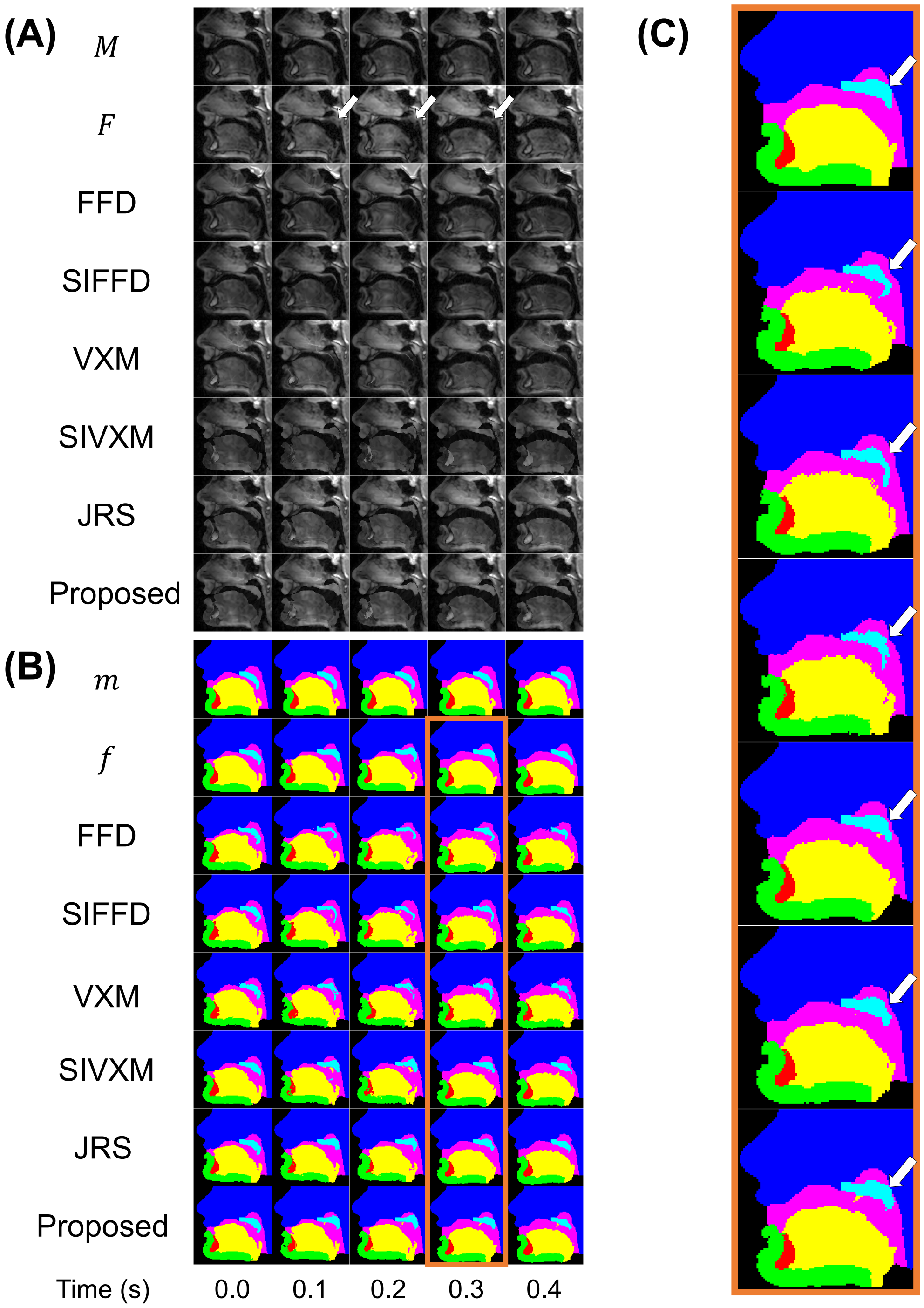

Figure 66: Transformed images and transformed ground-truth segmentations output by each method and framework, cropped to only show the vocal tract region. In (A), the first two rows show the moving image (M) and fixed image (F) pairs. The five fixed images are consecutive images from one of the image series and show a velopharyngeal closure. The white arrows show where the soft palate is in contact with the pharyngeal wall. The moving images are the reference image of the subject. The remaining rows in (A) show the transformed moving images output by the free-form deformations (FFD) and segmentation-informed FFD (SIFFD) methods and the VoxelMorph (VXM), segmentation-informed VXM (SIVXM), joint registration and segmentation (JRS) and proposed (Proposed) frameworks. In (B), the first two rows show the ground-truth segmentations of the moving image (m) and fixed images (f). The remaining rows in (B) show the transformed ground-truth segmentations output by each method or framework. (C) shows enlarged versions of the segmentations outlined in orange in (B). Image source: [291].

6.4.3 Displacement Field Accuracy Evaluation

6.4.3.1 Dice Coefficients and Average Symmetric Surface Distances

Figure 67 and Figure 68 show the DSCs of the transformed GT segmentations output by each of the methods and frameworks, while Figure 69 and Figure 70 show the ASDs of the transformed GT segmentations. In Figure 67 and Figure 69, the evaluation metric is averaged over all segmentation classes, while in Figure 68 and Figure 70 the evaluation metric is averaged over a single segmentation class.

The median DSCs of the segmentation-informed frameworks were consistently higher than those of the FFD methods and VXM framework, both when DSCs were averaged over all six segmentation classes (as shown in Figure 67) and when DSCs were averaged over a single class (as shown in Figure 68). However, as shown in Figure 68 where the DSCs are averaged over a single class, no segmentation-informed framework consistently achieved statistically significantly higher DSCs than the others. Although the SIVXM framework achieved the highest median DSC in three classes (head, soft palate and tooth space), in the soft palate class there was no statistically significant difference between its DSCs and those of the proposed framework, and in the head class there was no statistically significant difference between its DSCs and those of the JRS framework. Similarly, although the proposed framework achieved the highest median DSC in two classes (jaw and vocal tract), in the jaw class there was no statistically significant difference between its DSCs and those of the JRS framework. However, the ranges of the DSCs of the proposed framework were consistently narrower than those of the other frameworks, suggesting improved robustness in registration performance.

As shown in Figure 69 and Figure 70, almost identical trends in framework performance were observed when the frameworks were evaluated using the ASD as when the frameworks were evaluated using the DSC.

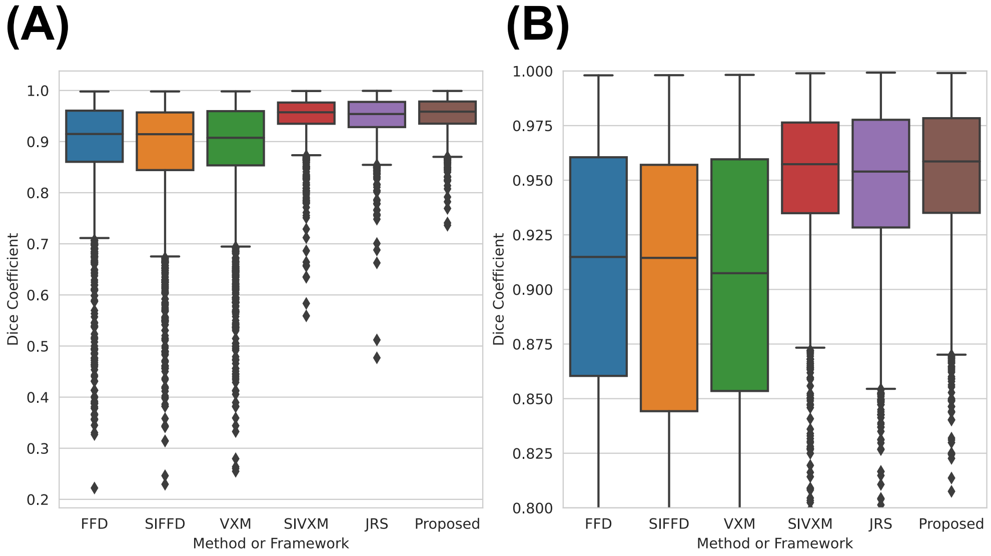

Figure 67: Dice coefficients (DSCs) of the transformed ground-truth segmentations output by the free-form deformations (FFD) and segmentation-informed FFD (SIFFD) methods and the VoxelMorph (VXM), segmentation-informed VXM (SIVXM), joint registration and segmentation (JRS) and proposed (Proposed) frameworks. The DSCs are averaged over the six segmentation classes. (B) shows the section of (A) where the DSCs are between 0.8 and 1.0. There were statistically significant differences between all the DSC groups. Image source: [291].

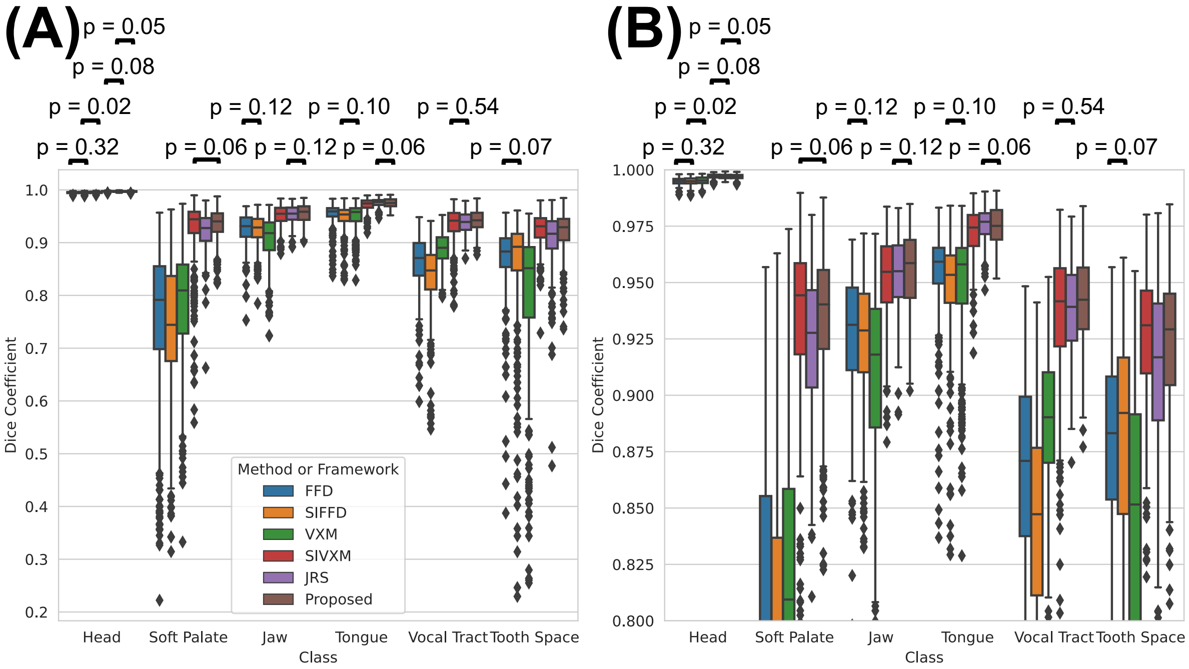

Figure 68: Dice coefficients (DSCs) of the transformed ground-truth segmentations output by the free-form deformations (FFD) and segmentation-informed FFD (SIFFD) methods and the VoxelMorph (VXM), segmentation-informed VXM (SIVXM), joint registration and segmentation (JRS) and proposed (Proposed) frameworks. The DSCs are averaged over a single segmentation class. (B) shows the section of (A) where the DSCs are between 0.8 and 1.0. There were statistically significant differences between all the DSC groups, except between pairs of groups indicated with black bars above the box plots. Image source: [291].

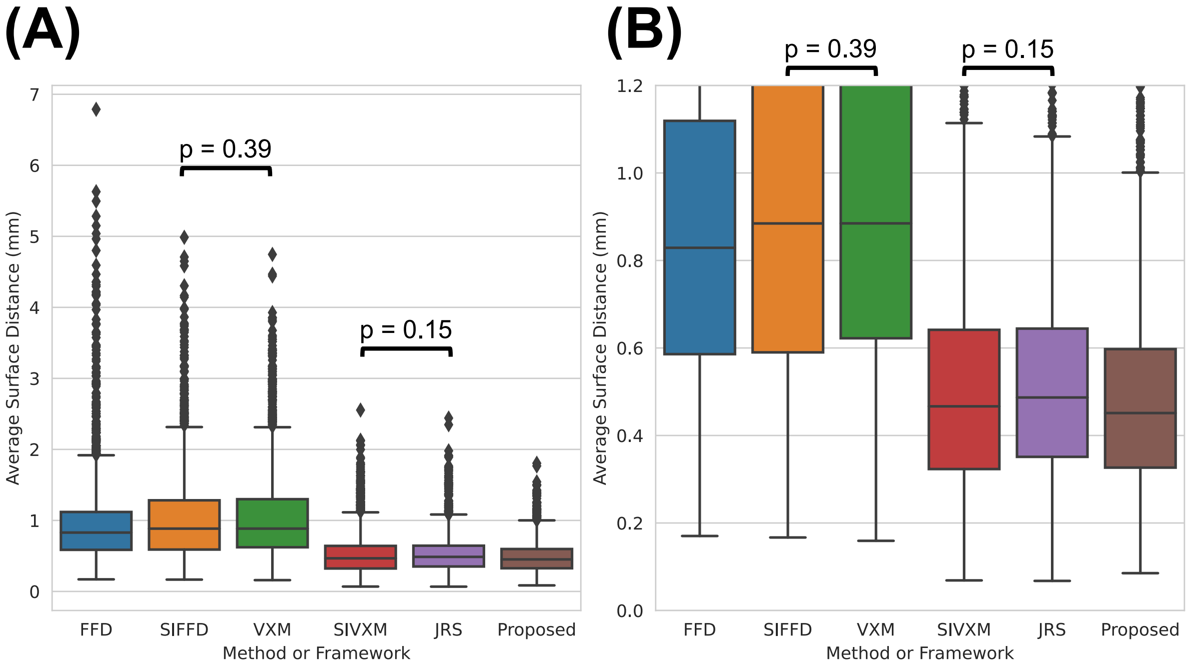

Figure 69: Average symmetric surface distances (ASDs) of the transformed ground-truth segmentations output by the free-form deformations (FFD) and segmentation-informed FFD (SIFFD) methods and the VoxelMorph (VXM), segmentation-informed VXM (SIVXM), joint registration and segmentation (JRS) and proposed (Proposed) frameworks. The ASDs are averaged over all six segmentation classes. (B) shows the section of (A) where the ASDs are between 0.0 and 1.2. There were statistically significant differences between all the ASD groups, except between pairs of groups indicated with black bars above the box plots. Image source: [291].

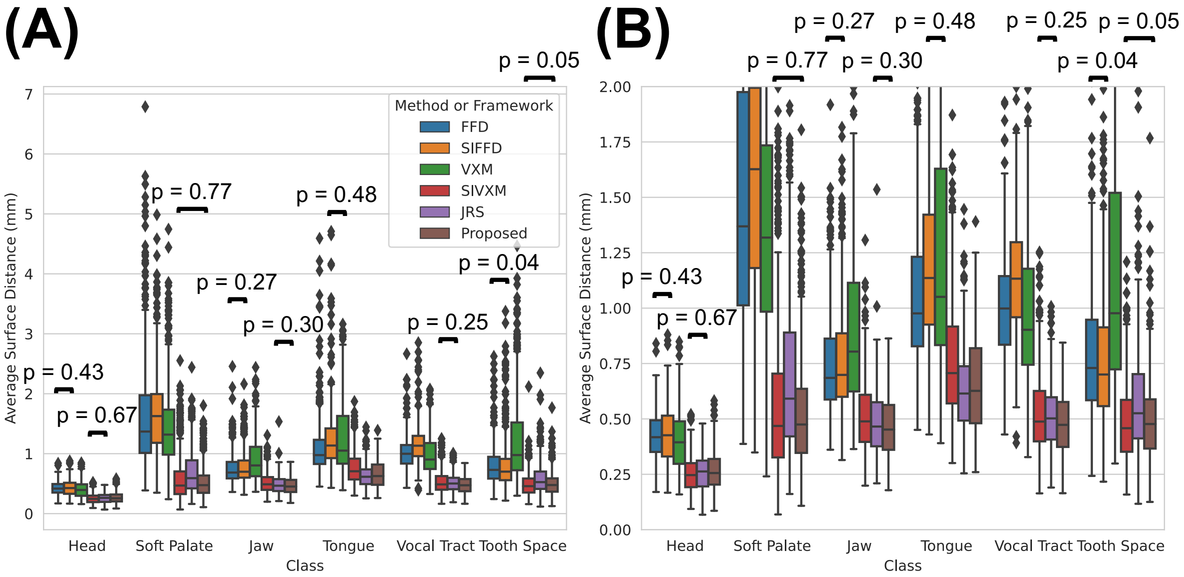

Figure 70: Average symmetric surface distances (ASDs) of the transformed ground-truth segmentations output by the free-form deformations (FFD) and segmentation-informed FFD (SIFFD) methods and the VoxelMorph (VXM), segmentation-informed VXM (SIVXM), joint registration and segmentation (JRS) and proposed (Proposed) frameworks. The ASDs are averaged across a single segmentation class. (B) shows the section of (A) where the ASDs are between 0 and 2. There were statistically significant differences between all the ASD groups, except between pairs of groups indicated with black bars above the box plots. Image source: [291].

6.4.3.2 True Velopharyngeal Closures

Figure 71 shows the number of true velopharyngeal closures in the transformed GT segmentations output by each of the methods and frameworks.

The FFD methods failed to capture any velopharyngeal closures. Comparing the frameworks, the VXM framework captured the smallest number of velopharyngeal closures (3), while the proposed framework captured the largest (27). Furthermore, the proposed framework captured all the closures in four of the five image series, while the SIVXM and JRS frameworks only captured all the closures in one of the series and the VXM framework did not capture all the closures in any of the series. There were statistically significant differences between the true velopharyngeal closures captured by each framework, except between the SIVXM and JRS frameworks.

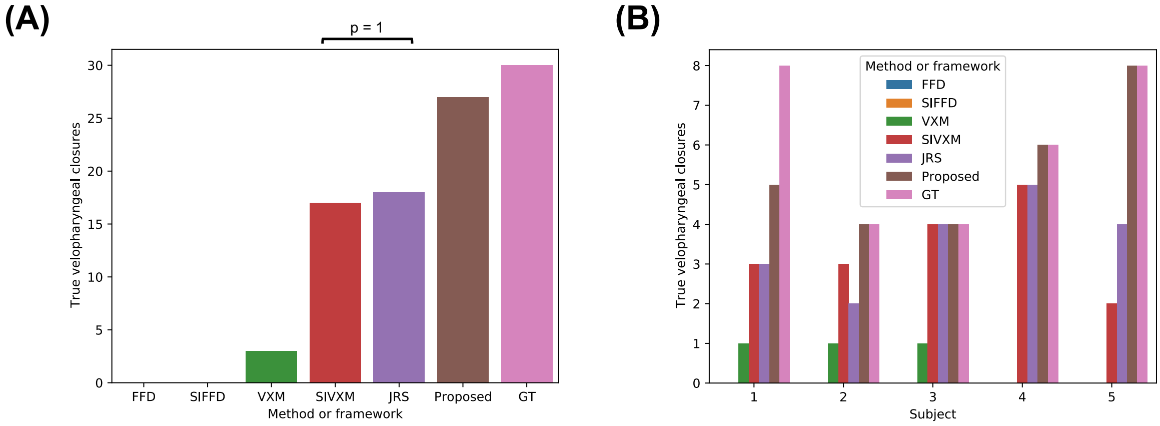

Figure 71: True velopharyngeal closures in the transformed ground-truth (GT) segmentations (of the moving image) output by the free-form deformations (FFD) and segmentation-informed FFD (SIFFD) methods and the VoxelMorph (VXM), segmentation-informed VXM (SIVXM), joint registration and segmentation (JRS) and proposed (Proposed) frameworks. The bars labelled GT indicate the number of velopharyngeal closures in the GT segmentations of the fixed images. In (A) the true velopharyngeal closures are summed across all five subjects, while in (B) the true velopharyngeal closures are summed across a single subject. There were statistically significant differences between the true velopharyngeal closures captured by each framework, except between the frameworks indicated with the black bar in (A). Image source: [291].

6.4.4 Ablation Study

Figure 72 shows the DSCs of all classes in the transformed GT segmentations output by each version of the proposed framework, while Figure 73 shows the ASDs of all classes. The median DSCs of the classes that were used as inputs to the registration CNN of the framework were consistently higher than those of the other classes, while the median ASDs of the classes were consistently lower.

Figure 74 shows the number of true velopharyngeal closures in the transformed GT segmentations output by each version of the proposed framework. The version where the head, soft palate and vocal tract classes were used as inputs to the registration CNN captured the same number of closures as the version where all classes were used as inputs, while the version where the soft palate and vocal tract classes were used as inputs captured one less closure. The version where the soft palate and vocal tract classes were not used as inputs failed to capture any closures.

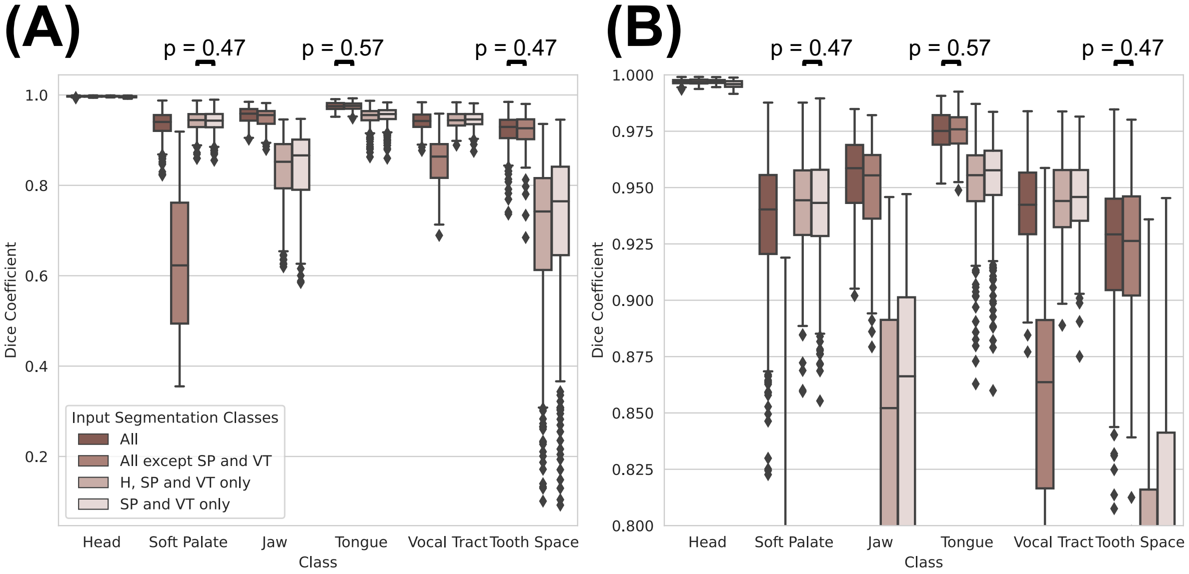

Figure 72: Dice coefficients (DSCs) of the transformed ground-truth segmentations output by the proposed framework, averaged across a single segmentation class. The colour code indicates the segmentation classes used as inputs to the registration convolutional neural network of the proposed framework during training and evaluation. In the Figure legend, 'All' indicates that all six segmentation classes described in section 4.1.4 were used as inputs, while 'H, SP and VT' indicates the head (H), soft palate (SP) and vocal tract (VT) classes. (B) shows the section of (A) where the DSCs are between 0.8 and 1. There were statistically significant differences between all the ASD groups, except between pairs of groups indicated with black bars above the box plots. Image source: [291].

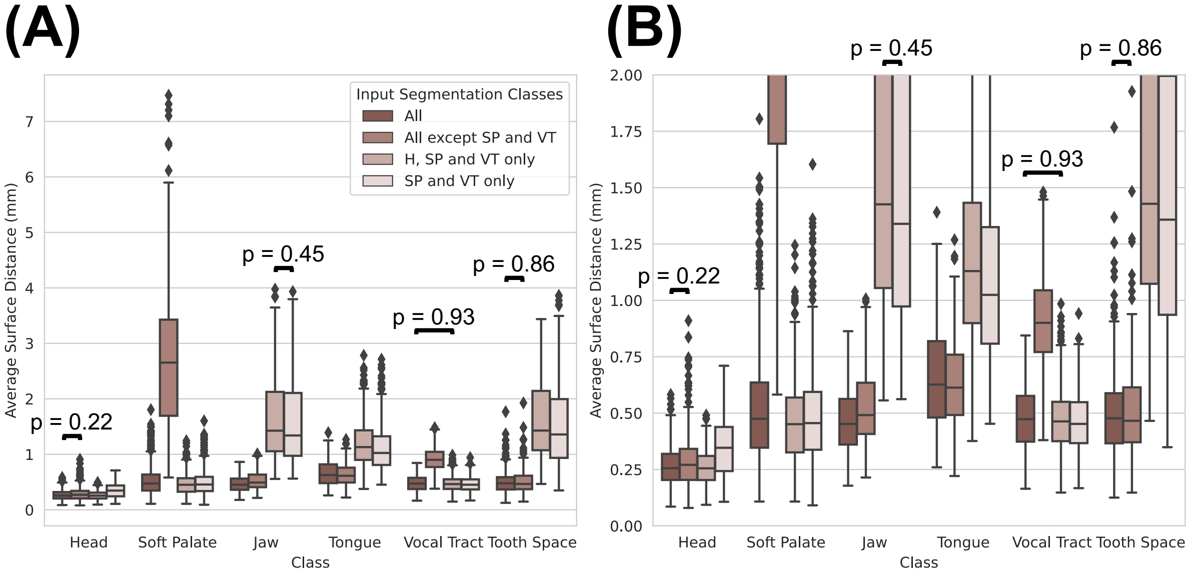

Figure 73: Average symmetric surface distances (ASDs) of the transformed ground-truth segmentations output by the proposed framework, averaged across a single segmentation class. The colour code indicates the segmentation classes used as inputs to the registration convolutional neural network of the proposed framework during training and evaluation. In the Figure legend, 'None (VXM)' indicates the results of the VoxelMorph framework, 'All' indicates that all six segmentation classes described in section 4.1.4 were used as inputs, while 'H, SP and VT' indicates the head (H), soft palate (SP) and vocal tract (VT) classes. (B) shows the section of (A) where the ASDs are between 0 and 2. There were statistically significant differences between all the ASD groups, except between pairs of groups indicated with black bars above the box plots. Image source: [291].

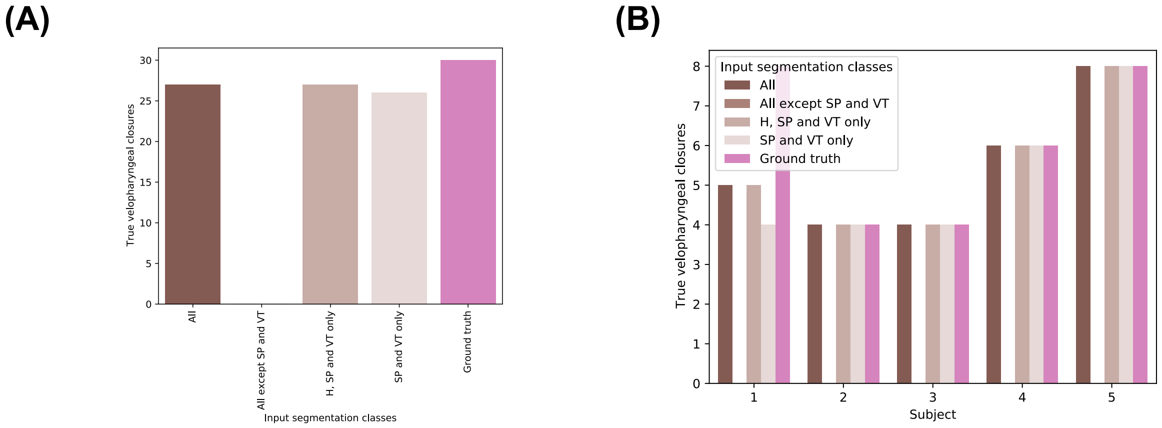

Figure 74: True velopharyngeal closures in the transformed ground-truth segmentations (of the moving image) output by the proposed framework. The label 'Ground truth' indicates the number of velopharyngeal closures in the ground-truth segmentations of the fixed images. In (A) the closures are summed across all five subjects. The label 'All' indicates that all six segmentation classes described in section 4.1.4 were used as inputs to the registration convolutional neural network of the proposed framework, while 'H, SP and VT' indicates the head (H), soft palate (SP) and vocal tract (VT) classes. In (B) the true velopharyngeal closures are summed across a single subject. Image source: [291].

6.5 Discussion

A framework for estimating displacement fields between 2D rtMR images of the vocal tract during speech was successfully developed. The framework is based on the SIVXM framework [245] but features two adaptations. First, the framework includes a method to segment the images. Second, segmentations as well as images are used as inputs to the registration CNN, in the same manner as the framework of Chen et al. [254]. Incorporation of a segmentation method in the framework enables its use when segmentations of the images are not already available. This is the first time DL-based nonlinear registration of MR images of speech has been investigated.

Evaluated using the DSC and ASD, the displacement field estimation accuracy of the proposed framework was superior to two FFD methods and a current state-of-the-art framework (the VXM framework), and very similar to two current state-of-the-art segmentation-informed frameworks (the SIVXM framework and a joint registration and segmentation framework). However, when evaluated using a metric based on velopharyngeal closure, its performance was superior to all five state-of-the-art registration methods and frameworks. In other words, the displacement fields estimated by the proposed framework captured more of the velopharyngeal closures in the image series, and therefore better captured this aspect of articulator motion than the methods and other frameworks.

These results show that metrics based on clinically relevant and quantifiable aspects of organ motion can be used to evaluate the accuracy of registration frameworks and can be more sensitive to differences in accuracy than standard metrics such as the DSC and ASD.

In addition, these results show that registration CNNs input with segmentations as well as images can estimate displacement fields that better capture aspects of articulator motion than registration CNNs input with images only, even if the segmentations are estimates rather than ground truths.

The FFD methods failed to capture any velopharyngeal closures. This result is unsurprising as these methods are designed to estimate smooth and continuous displacement fields, while discontinuous displacement fields are required to capture the complex motion of the articulators. Removing the smooth and continuous displacement field constraints in the cost function did not improve the registration accuracy of the methods, showing that there are additional reasons why they are not appropriate for capturing articulator motion. When registering to fixed images showing velopharyngeal closure, the FFD method consistently misregistered the pharyngeal wall to the soft palate, instead of registering the soft palate to the soft palate. An example of this is shown in Figure 66C. The SIFFD method, which ensured that the head (which includes the pharyngeal wall) deformed in a rigid manner, successfully prevented misregistration of the pharyngeal wall to the soft palate but did not improve the soft palate registration accuracy. Ideally, the proposed framework would have been compared with the FFD-based method developed by Hua et al. [261], as this method was designed to estimate displacement fields with discontinuities. However, unfortunately this was not possible as there is no publicly available implementation of the method.

The results of the ablation study show that unsurprisingly the head, soft palate and vocal tract segmentation classes are crucial for estimating displacement fields that accurately capture soft palate motion. This highlights the importance of using segmentations of the anatomical features whose motions are of interest but also segmentations of neighbouring features that provide information about the positions of the features of interest, for example whether the features of interest are in contact with other features. The results of the ablation study also show that using additional segmentation classes such as the jaw, tongue and tooth space did not affect the number of velopharyngeal closures captured by the framework. However, as shown in Figure 72 and Figure 73, using these additional classes was beneficial as it improved the accuracy with which they were registered by the framework.

To further encourage a CNN to estimate displacement fields that capture velopharyngeal closures, one approach for future investigation would be to use a loss function during CNN training that measures whether the starting points and durations of any velopharyngeal closures captured in a series of estimated displacement fields are correct. However, to be suitable for use in CNN training, this loss term would have to be differentiable. Developing a loss term that meets all these criteria would be challenging. A simpler approach would be to include a loss term based on whether individual transformed segmentations show contact between the soft palate and pharyngeal wall. This could be achieved using a topological loss term such as the one developed by [294] which can identify contact between different segmentation classes in a differentiable manner.

This work is another step towards the ultimate goal of automatic articulator segmentation and measurement in clinical practice. However, a large amount of future work is required to achieve this goal. More specifically, three major challenges must be overcome. One challenge concerns the dataset used to develop the method, while the other two are technical.

First, as explained in section 4.1.6, a larger and more diverse dataset, both in terms of subjects and image contrast, must be created and used to develop and extend the method. More specifically, a dataset more representative of the target patient population is required: since the target patient population primarily consists of children, the dataset must contain images of children. In addition, since velopharyngeal closure does not occur as expected in some of the speech of patients with VPI, the dataset must contain image series where velopharyngeal closure does not occur as well as image series where it does. In addition, a dataset with images acquired using many different MRI scanners and pulse sequences is required to ensure that methods developed using the dataset are generalisable and perform well on images from different sources. While there are publicly available 2D speech MRI dataset [18,19], these do not have corresponding GT segmentations thus limiting their use for training supervised DL-based segmentation-informed registration methods.

Second, to be suitable for use in clinical practice, the method should be extended so that the motion of specific features of articulators such as the tip of the soft palate can be automatically tracked. One way of achieving this would be to require users to manually define a point of interest in one of the images in the series. The method would then use the estimated displacement fields to track the motion of the given point during speech and analyse it to provide information such as the speed and direction of motion. Such an extension to the method would enable clinical teams to obtain nearly automatically clinically relevant information such as the direction in which the soft palate elevates during speech, the speed at which it elevates and the distance by which it elevates. Future work should extend the method in partnership with clinical teams to ensure that the measured aspects of motion are clinically relevant.

Third, to be suitable for use in clinical practice, the method should be robust to changes in head position in the image series due to subject motion. This could be achieved by performing a rigid registration pre-processing step before estimating displacement fields between images. Future work should aim to extend the method by including such a step.

6.6 Conclusions

A framework for estimating displacement fields between 2D rtMR images of the vocal tract during speech was successfully developed and found to more accurately capture aspects of articulator motion than five current state-of-the-art nonlinear registration methods and frameworks. This framework builds on the segmentation method presented in the previous chapter and is another step towards the ultimate goal of automatic articulator motion, shape and size quantification in such image series in clinical practice. In addition, a metric based on a clinically relevant and quantifiable aspect of articulator motion was proposed and shown to be useful for evaluating frameworks for registering 2D rtMR images of speech.

However, three main challenges must be addressed before the method is suitable for use in clinical practice. First, a larger, more diverse and representative dataset of 2D rtMR images of the vocal tract during speech must be created with corresponding GT segmentations and used to train and evaluate the method. Second, the method should be extended to automatically track the motion of specific features of the articulators. Third, the method should be extended to include a rigid registration pre-processing step before estimating displacement fields between images, to ensure it is robust to changes in head position in images as a result of subject motion.

Dynamic 2D imaging of the vocal tract provides clinical teams with 2D information about the motion of the soft palate during speech. However, a key consideration when making VPI treatment decisions is how well the LVP (introduced in section 2.1) is functioning. Important factors that affect LVP function are the shape of the muscle and its orientation relative to the soft palate. While dynamic 2D imaging provides 2D motion information that enables clinical teams to infer how well the muscle is functioning, this type of imaging does not allow visualisation of the muscle. It therefore does not provide clinical teams with information about LVP shape or orientation that could influence VPI treatment decisions and aid treatment planning. Three-dimensional imaging is required to fully visualise the LVP and thus obtain shape and orientation information about the muscle. The next chapter will describe the development of deep learning tools for automating the analysis of the LVP in 3D MR images of the vocal tract.

Chapter 7: Deep-Learning-Based LVP Segmentation in 3D MR Images

7.1 Introduction

The LVP (introduced in section 2.1) plays an essential role in speech production. As explained in section 2.2.1, a poorly functioning LVP can prevent velopharyngeal closure from occurring, leading to speech impairments. Typically in clinical speech assessments, imaging is used to visualise the motion of the soft palate during speech and LVP function is then inferred from the motion. However, the LVP is not visualised. As explained in section 2.2.3, a key factor that influences VPI treatment decisions is the defect(s) preventing velopharyngeal closure. If the defect is a poorly functioning LVP, a surgical treatment that aims to improve LVP function is performed [1].

As explained in section 3.1.8, there is increasing interest in LVP visualisation, to better understand variations in the shape and configuration of the muscle [25,122,131–140,123,141–143,124–130], to aid planning of surgical treatment of VPI [144,145], and for medical education purposes [146]. MRI is predominantly used for LVP visualisation [13,25,130–139,122,140,142,143,123–129], due to its unique ability to acquire images of any orientation with excellent soft tissue contrast without using ionising radiation. As explained in section 3.1.9, due to the small size of the LVP and its 3D structure, 3D imaging at a high spatial resolution is required to fully visualise the muscle. In previous work, 3.0 T MRI at a spatial resolution of 0.8×0.8×0.8 mm^3 has predominantly been used for 3D LVP visualisation [25,126,138–140,127–129,131–133,136,137]. The LVP and the soft tissue that surrounds it have very similar tissue properties. Consequently, a challenge when imaging the LVP is ensuring that the image contrast between the LVP and the surrounding soft tissue is sufficient to discriminate between the two. Previous work has predominantly acquired T_2-weighted 3D images of the LVP at 3.0 T using TSE pulse sequences [25,126,139,140,127–129,131,134,136–138]. In addition, a recommendation to acquire T_2-weighted images for assessing the LVP in clinical practice was recently made [8].

As explained in section 3.3.5, there is increasing interest in quantifying the LVP in MR images [13,25,130–139,122,140–143,123–129]. In all previous work [13,25,130–

139,122,140–143,123–129], measurements such as the length and thickness of the LVP were manually obtained from MR images. However, obtaining measurements in this way is time-consuming, requires input by specialists and is prone to intra- and inter-observer variability. To avoid the burden of manual measurements and to facilitate LVP measurement on a larger scale, there is currently an unmet need for automatic LVP measurement methods. A common approach for automating the measurement of anatomical features in biomedical images is to first segment the features and then perform measurements using the segmentations. As a first step towards developing an automatic LVP measurement method, in very recent work [17], four state-of-the-art DL-based methods were used to segment the LVP and five other anatomical features (adenoids, lateral pharyngeal wall, posterior pharyngeal wall, pterygoid raphe and soft palate) in 3D T_1-weighted MR images. More specifically, two methods based on 3D U-Net [177] (one of which was developed using nnU-Net [187]), the Swin UNETR method [185] and the 3D UX-Net method [219] were used. Evaluated using the DSC, the 3D UX-Net method was found to most accurately segment the LVP and three of the other anatomical features. However, there are no reports in the literature of methods to segment the LVP in MR images with other contrasts such as T_2-weighted images, the contrast that was recently recommended for LVP visualisation in clinical practice [8].

As well as providing a step towards automatic LVP measurement, LVP segmentation offers the opportunity for 3D printing of physical models of the LVP for use in surgical treatment planning and for educational purposes [146]. However, such models would require more anatomical context than simply the LVP. Of particular interest to clinicians is the orientation of the LVP relative to the soft palate and pharynx [144]. Since the soft palate and pharynx are adjacent anatomical features, a segmentation of the latter feature would provide information about the posterior surface of the former.

As explained in section 3.3.1, recently, DL-based methods have achieved state-of-the-art accuracy in the segmentation of 3D images of body organs such as the heart [171,188,191], brain [191] and kidneys [189]. While a wide range of different DL-based segmentation methods have been proposed [295,296], extensions of 3D U-Net [177] such as V-Net [178] and those created by nnU-Net [187] have consistently achieved state-of-the-art accuracy in the segmentation in 3D images of the heart [188,191], brain [191] and kidneys

[189]. In fact, as explained in section 3.3.1, nnU-Net is a framework for configuring and training U-Net effectively.

Typically, CNNs must be trained using large amounts of data to perform a task [153,154]. When large amounts of training data are not available, it is common to use data augmentation methods (introduced in section 3.2.4) to artificially increase the amount of training data [153,154]. Methods to increase the variability of the appearance and layout of medical images are most frequently used [154]. The former methods change the appearance of images by for example adding random noise to the pixel or voxel values, while the latter methods change the layout of images by for example rotating and translating them. However, ideally data augmentation methods should also increase the anatomical variability in the training data. A commonly used way to increase this variability is to augment the training data using random elastic deformations [154], however, such augmentation does not always result in anatomically plausible data [297]. To increase the anatomical variability of training data while maintaining anatomical plausibility, data augmentation methods based on non-linear registration methods have been proposed [297–300]. Data augmentation methods based purely on non-linear registration have been shown to improve the accuracy of DL-based methods to segment the brain [297,298,301] and knee [298] in 3D MR images, while data augmentation methods based on statistical deformation models (SDMs) have been shown to improve the accuracy of DL-based methods to segment the heart in 2D MR images [299,300].

The main contribution of the work presented in this chapter is the development of a method to segment the LVP in 3D MR images of the vocal tract with the contrast that was recently recommended for LVP visualisation in clinical practice [8]. The development of such a method is a step towards the ultimate goal of automatic LVP segmentation and quantification in clinical practice.

7.2 Methods

A DL-based method to automatically segment the pharynx and LVP in 3D MR images of the vocal tract was developed using the nnU-Net framework [187]. The method consists of three sequential steps: image pre-processing, segmentation estimation using a CNN and then segmentation post-processing. The image pre-processing step is identical to that of nnU-Net.

A suitable architecture for the segmentation CNN was identified using the nnU-Net process for this purpose. The nnU-Net CNN training process was then almost fully followed to train the segmentation CNN. Deviations from the nnU-Net CNN training process are described in section 7.2.1. The nnU-Net framework was implemented using the code publicly available at https://github.com/MIC-DKFZ/nnUNet.

7.2.1 Proposed Method Implementation and CNN Training

The image pre-processing step of the method is as follows: each image is normalised independently by first subtracting its mean voxel intensity and then dividing by the standard deviation of its voxel intensities. This pre-processing step is performed during training and at test time.

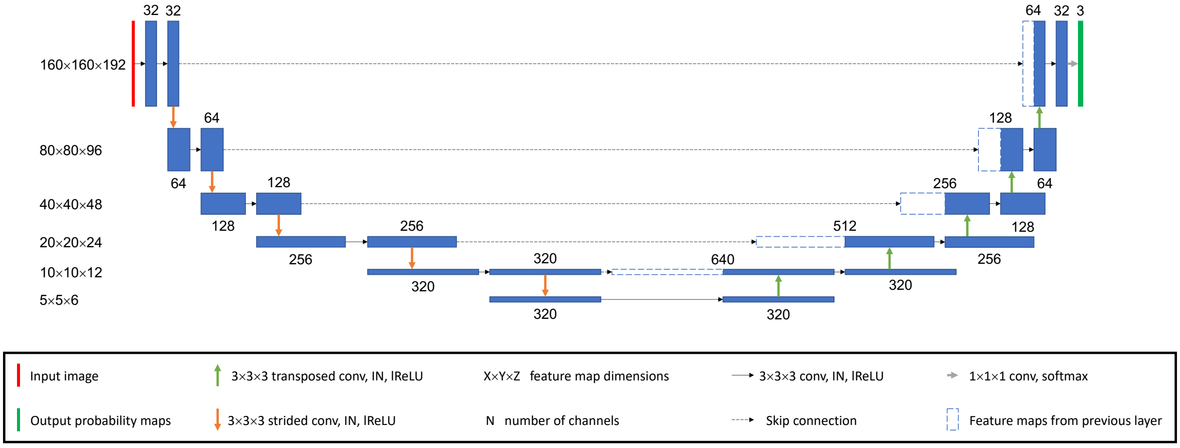

Figure 75: Segmentation convolutional neural network architecture. IN: instance normalization; lReLU: leaky rectified linear unit with negative slope 0.01; conv: convolution.

The CNN architecture is based on that of the 3D U-Net [177] and is depicted in Figure 75. CNN training was performed on a 24 GB NVIDIA TITAN RTX graphics card. During CNN training, the image patch size was 128×128×128 voxels and the mini-batch size was two image patches, and stochastic gradient descent with Nesterov momentum (μ=0.99) and an initial learning rate, LR, of 0.01 were used. Following each epoch, LR was decayed according to the following equation:

$$LR = \left(1 - \frac{epoch}{epoch_{max}}\right)^{0.9} \tag{30}$$

The loss function consisted of the sum of the cross entropy loss and the Dice loss, both introduced in section 3.3.2. Hyperparameter optimisation was not performed as nnU-Net instead identifies suitable hyperparameter values using heuristic rules.

The three deviations from the nnU-Net training process were as follows. First, the default nnU-Net data augmentation methods were not used in all experiments. Second, the CNN was trained for 200 epochs instead of the default of 1000. Third, the number of mini-batches per epoch was 23 instead of the default of 250 in all experiments except the one without data augmentation, where the number of mini-batches per epoch was five. The rationale for the first deviation was to investigate the effect of different data augmentation methods on the segmentation method accuracy, while the rationale for the other two deviations was to avoid the CNN overfitting as a result of the small amount of training data. In addition, a further rationale for the number of mini-batches per epoch was to ensure the number of patches inputted to the CNN per epoch was equal to the number of images in the training dataset. The training process deviations were motivated by the observation that during CNN training the validation loss stabilised after approximately 50 epochs, as shown in Figure 80 in section 7.4.

The segmentation post-processing step of the method is as follows: for each segmentation class, the number of connected components in the segmentation is identified and all regions except the one with the largest number of voxels are removed. This post-processing step is performed at test time only.

7.3 Experiments

7.3.1 Data

Cropped versions of the 15 images and corresponding GT segmentations presented in section 4.2.3 were used in the experiments. Images and GT segmentations were cropped centred on the LVP and pharynx to ensure they only contained relevant anatomy and to also reduce the computational burden of the experiments. All images and GT segmentations were cropped to a size of 160×160×192 voxels. This size was chosen based on the following analysis:

1. For each full-size image and corresponding GT segmentation, the centroid coordinates and dimensions of the smallest 3D bounding box that fully contained the LVP and pharynx was identified.
2. The dimensions of the bounding boxes were compared, to identify the largest x-, y- and z-dimensions (128, 132 and 172 voxels respectively).
3. The largest dimensions were slightly increased to 160, 160 and 192 voxels respectively to ensure that the cropped images and corresponding GT segmentations included a buffer region around the LVP and pharynx.
4. Each full-size image and corresponding GT segmentation was cropped to a size of 160×160×192 voxels centred on the corresponding centroid coordinates identified in step 1.

In all experiments, nine of the 15 images were used either as training data for the proposed method or as the data used to synthetically create new training data. Of the remaining six images, three were used as validation data and three as test data. More details about the train/validation/test dataset splits used in the experiments are provided in section 7.3.5.

7.3.2 Data Augmentation

The effect of different data augmentation methods on the accuracy of the segmentation method was investigated. To achieve this, separate experiments were performed where the segmentation method was developed from training data augmented using different methods. The accuracy of the segmentation methods was then compared using the evaluation metrics described in section 7.3.3. Three augmentation methods along with combinations of these methods were investigated. The methods are described in the sections 7.3.2.1, 7.3.2.2 and 7.3.2.3. Data augmentation was used to synthesise 45 images from the nine original images in the training dataset, and then only the 45 synthesised images were used to train the segmentation CNN of the proposed method. When training the segmentation CNN of the proposed method without data augmentation, the training dataset consisted of nine original images

7.3.2.1 Default nnU-Net Augmentation

By default, nnU-Net applies the augmentations listed in Table 11 to the images in the training dataset and modifies the corresponding GT segmentations accordingly.

Table 11: Default augmentations applied to training data by nnU-Net. $x \sim U(a, b)$ indicates that x is sampled from a uniform distribution with lower limit a and upper limit b. For the brightness and contrast augmentations, effectively the image patch is multiplied pixel-wise with a mask of random values sampled from $U(a, b)$.

Augmentation	**Probability**	**Description**
Rotation	0.2	Image rotated by angle in range $U(-180°, 180°)$
Scaling	0.2	Image scaled by factor in range $U(0.7, 1.4)$
Gaussian noise	0.15	Zero-centred Gaussian noise with variance in range $U(0, 0.1)$ added to voxel intensities
Gaussian blur	0.1	Gaussian blur with kernel width in voxels in range $U(0.5, 1.5)$ applied to image
Brightness	0.15	Voxel intensities multiplied by $x \sim U(0.7, 1.3)$
Contrast	0.15	Voxel intensities multiplied by $x \sim U(0.65, 1.5)$ and then clipped to the original intensity range
Mirroring	0.5	Image is mirrored along an axis

7.3.2.2 Registration-Based Augmentation

This method was inspired by a registration-based interpolation method [302] and was used to create a new and larger training dataset of 45 images from the nine images in the original training dataset. The method consisted of the following steps, as shown in Figure 76:

1. A pair of images, M and F, was randomly chosen from the nine images in the original training dataset.
2. A vector displacement field, D, mapping how the voxels in M should be displaced to align them with corresponding voxels in F was estimated using affine followed by non-linear image registration. Section 7.3.3 provides more details about the image registration and its optimisation.
3. D was interpolated by multiplying it by a value randomly sampled from a continuous uniform distribution with lower limit 0.2 and upper limit 0.8:

$$D_{interp} = \alpha D \tag{31}$$

where α is a weighting term between 0.2 and 0.8.

4. An augmented image and corresponding GT segmentation was created by transforming M and its corresponding GT segmentation according to D_{interp}.

The following conditions were imposed when randomly choosing the image pairs:

- Each image in the original training dataset must be:
  - ○ M in five image pairs
  - ○ F in five image pairs
- Every pair must be unique
- The images in a pair must be different

Examples of GT segmentations created using the registration-based augmentation method are shown in Figure 77.

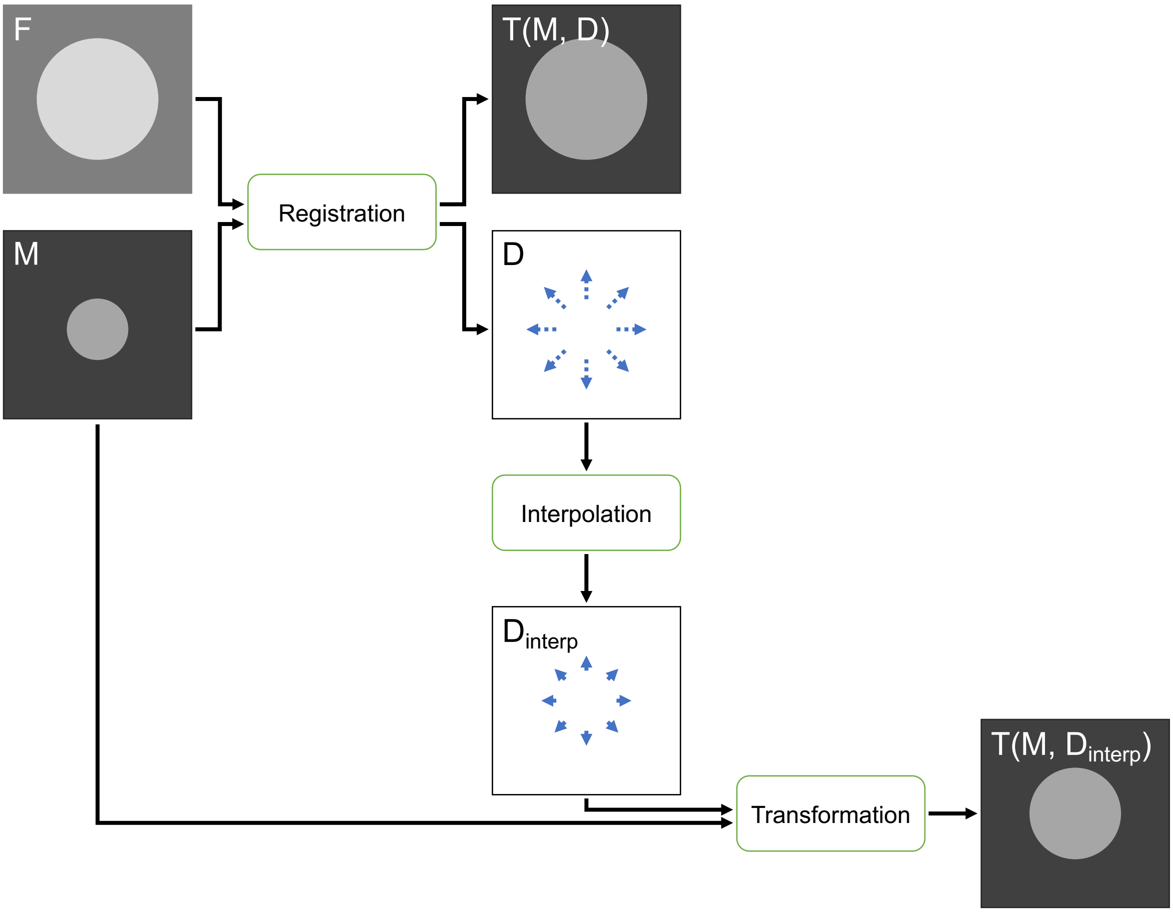

Figure 76: An overview of the registration-based augmentation method. A moving image (M) is nonlinearly registered to a fixed image (F). The resulting displacement field, D, is interpolated to D_{interp} by multiplication by a weighting term between 0.2 and 0.8. An augmented version of M is created by transforming M according to D_{interp}.

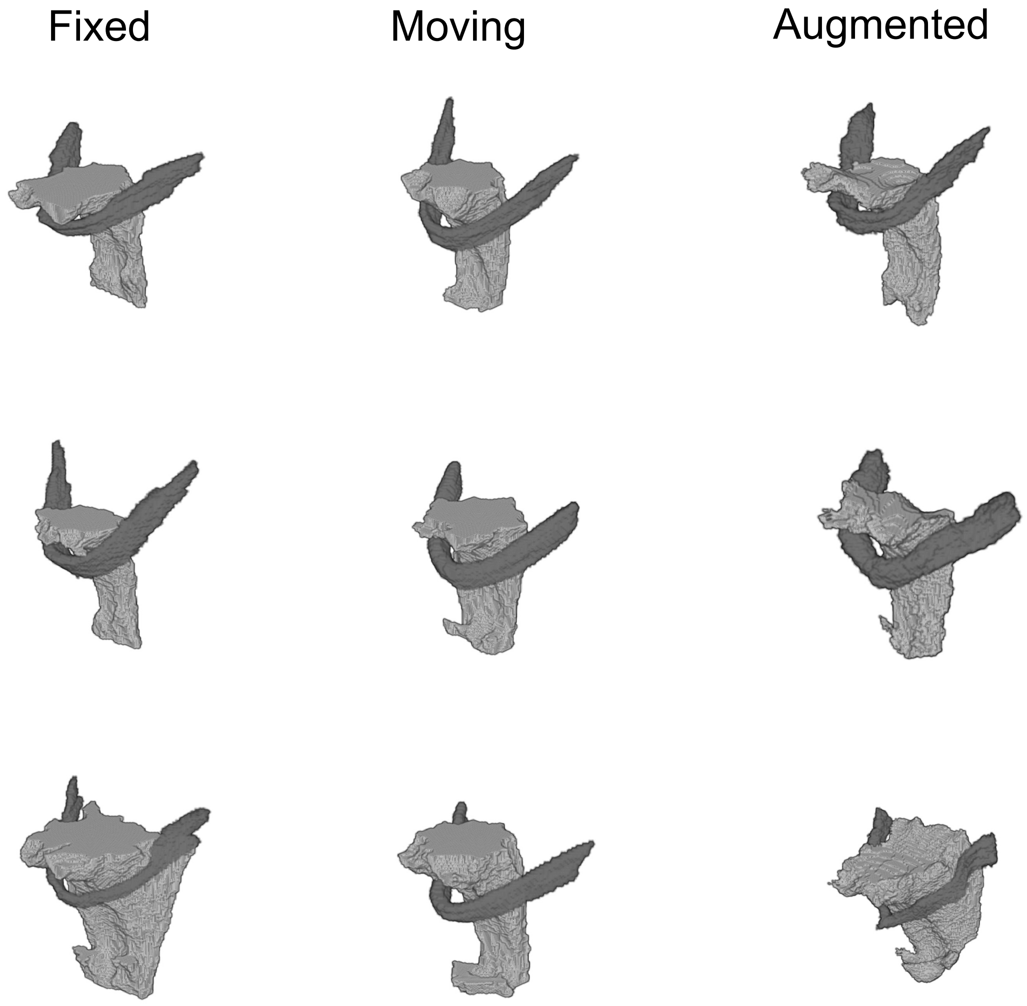

Figure 77: Examples of ground-truth segmentations created using the registration-based augmentation method. 'Fixed' and 'Moving' indicate segmentations of the fixed and moving images respectively, while 'Augmented' indicates a segmentation created using the augmentation method.

7.3.2.3 SDM-Based Augmentation

SDMs were created from the nine images in the original training dataset and then used to synthesise new and larger training datasets of 45 images and their corresponding GT segmentations. SDMs were created using the method developed by Rueckert et al. [303]. The full process for synthesising the images and corresponding segmentations was as follows:

1. A reference image, I_{ref}, was randomly chosen from the nine images in the original training dataset.
2. The other eight images were rigidly registered to I_{ref}.

3. Eight vector displacement fields, u, were created by first affinely and then non-linearly registering I_{ref} to each of the eight images created in step 2.
4. The mean displacement field, $\bar{u}$, and the first $n \in \{1, 2, \ldots, 7\}$ principal modes of variation of the fields, p_n, were determined using principal component analysis.
5. New displacement fields were created by adding principal modes of variation to the mean displacement field:

$$u_{\text{new}} = \bar{u} + \beta p_n \tag{32}$$

where β is a weighting term. As recommended by Rueckert et al. [303], values of β were randomly chosen within the range $\pm 3 \times \sqrt{\lambda_n}$ where λ_n is the eigenvalue of p_n.

6. Images and corresponding segmentations were synthesised by deforming I_{ref} and its corresponding segmentations according to the displacement fields created in step 5.

In one experiment, only one SDM was created from the nine images in the original dataset. Forty-five images and their corresponding segmentations were synthesised using this SDM and then used as the training dataset. In another experiment, nine SDMs were created, each using a different I_{ref} so that each image in the original training dataset was used as I_{ref}. Five images and their corresponding segmentations were synthesised using each SDM and the 45 resulting images were then used as the training dataset. Section 7.3.3 provides more details about the image registration and its optimisation. Examples of GT segmentations created using the SDM-based augmentation method are shown in Figure 78.

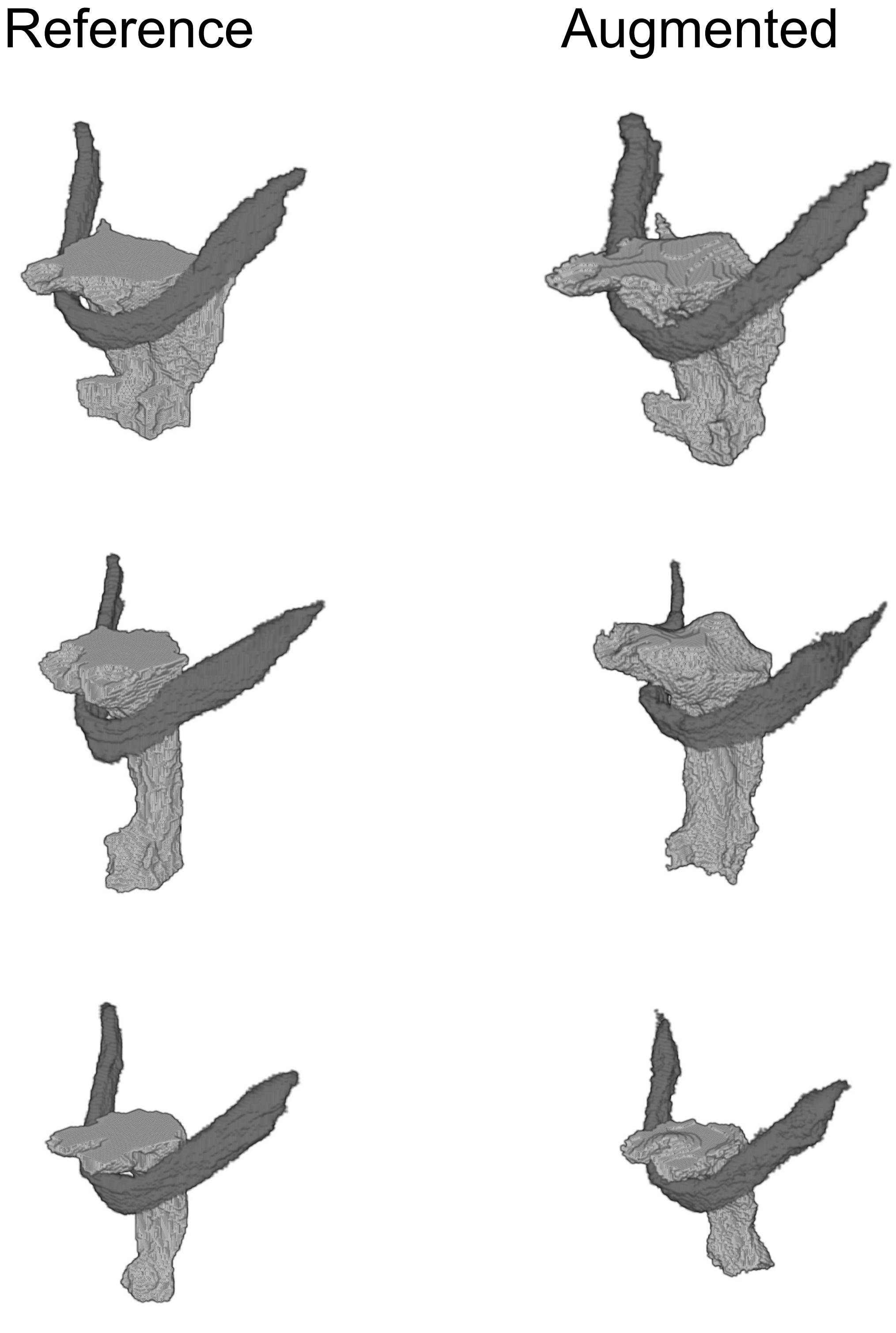

Figure 78: Examples of ground-truth segmentations created using the stastistical-deformation-model-based augmentation method. 'Reference' indicates the segmentation of the reference image used to create a model, while 'Augmented' indicates a segmentation created using the augmentation method.

7.3.3 Image Registration

The registration-based and SDM-based augmentation methods described in sections 7.3.2.2 and 7.3.2.3 respectively both require image registration. This section describes the registration that was used in these methods. NiftyReg version 1.5.39 [237,241] was used to perform rigid, affine and non-linear registration. NiftyReg performs rigid and affine registration using a block-matching method [238] and non-linear registration using a FFD method (introduced in section 3.4.2) [236]. Several registrations were performed per method using different combinations of parameter values (listed in Table 12), and then evaluated using the DSC (introduced in section 3.3.2), enabling identification of optimal values. Default values were used for all other parameters.

Table 12: Image registration parameter values. 'Rigid and affine' and 'Non-linear' indicate the registration method. $\boldsymbol{N_{levels}}$ and $\boldsymbol{N_{iterations}}$ are parameters in the block-matching method. $\boldsymbol{N_{levels}}$ indicates the number of levels to use to generate the pyramids for the coarse-to-fine approach, while $\boldsymbol{N_{iterations}}$ is the maximum number of iterations of the least trimmed squares method. Grid spacing, λ and γ are parameters in the free-form deformation method. Grid spacing is the spline grid spacing in voxels, λ is the weighting of the bending energy term, while γ is the weighting of the first order penalty term.

Parameter	**Rigid and affine**	**Non-linear**
$\boldsymbol{N_{\text{levels}}}$	{2, 3, 4}	N/A
$\boldsymbol{N_{\text{iterations}}}$	{4, 5, 6}	N/A
Grid spacing	N/A	{4, 5, 6}
$\boldsymbol{\lambda}$	N/A	{0, 0.001}
$\boldsymbol{\gamma}$	N/A	{0, 0.01}

7.3.4 Evaluation Metrics

At test time, the accuracy of the segmentations estimated by the segmentation method developed using each augmentation approach were evaluated using two metrics: the DSC and the HD (introduced in section 3.3.2).

7.3.5 Five-Fold Cross-Validation

A five-fold cross-validation was performed to evaluate the generalisability of the segmentation method. In each fold, the train/validation/test dataset split was 9/3/3 respectively. Test datasets were created by randomly splitting the 15 images into five groups of three images. Validation datasets were created in the same way, while ensuring that in each fold the images in the validation dataset were all different from the images in the test dataset. As a result, each image was included in the validation dataset of one fold, the testing dataset of another fold, and the training datasets of the remaining three folds. Table 13 lists the images in each dataset. Since hyperparameter optimisation was not performed as nnU-Net instead identifies suitable hyperparameter values using heuristic rules, a nested cross-validation was not required and the test dataset was held out until test time.

Table 13: The identifiers of the images in the validation and test datasets of each fold, and the identifiers of the images used as the reference images during statistical deformation model creation.

Fold	Reference	Validation dataset	Test dataset
1	1	2, 10, 12	6, 7, 9
2	4	1, 5, 13	3, 8, 15
3	13	4, 9, 11	5, 10, 14
4	14	6, 8, 15	1, 12, 13
5	15	3, 7, 14	2, 4, 11

7.3.6 Statistical Tests

Paired sample t-tests were performed to compare the DSCs of the segmentations estimated by different versions of the segmentation method. Groups of HDs were compared in the same way as groups of DSCs. Since each of the 15 image was included in one of the test datasets in the cross-validation, each group consisted of 15 values (one per image). The normality of a group was assessed using a Chi-squared goodness-of-fit test. All values were normally distributed. All statistical tests were performed using MATLAB 2019b (MathWorks, Natick, MA). A significance level of 5% was used, corrected using the Holm-Bonferroni method to compensate for multiple comparisons.

7.4 Results

Examples of segmentations estimated by different versions of the proposed method are shown in Figure 79. Columns (A), (B) and (C) in Figure 79 show examples with relatively low, average and high DSCs respectively.

The segmentation CNN training and validation losses are shown in Figure 80. The number of epochs required for validation loss stabilisation depended on the data augmentation method used during CNN training. On average (median), stabilisation required approximately 40 epochs. Stabilisation was fastest, requiring approximately 20 epochs, when only SDM-based data augmentation was used. Conversely, stabilisation was slowest, requiring approximately 60 epochs, when no data augmentation was used.

The effect of the post-processing step of the proposed method on the accuracy of the estimated segmentations is shown in Figure 81. In most cases, the step improved the accuracy of the estimated segmentations. On average (median), the step increased the DSCs of the estimated LVP and pharynx segmentations in 60% (9 of 15) and 80% (12 of 15) cases respectively, and decreased the HDs of the segmentations in 67% (10 of 15) cases. However, on average (median), the step also decreased the DSCs of the estimated LVP segmentations in 13% (2 of 15) cases. The only version of the proposed method where the step did not reduce the accuracy of any of the segmentations was the one where only the default nnU-Net data augmentations were used during training.

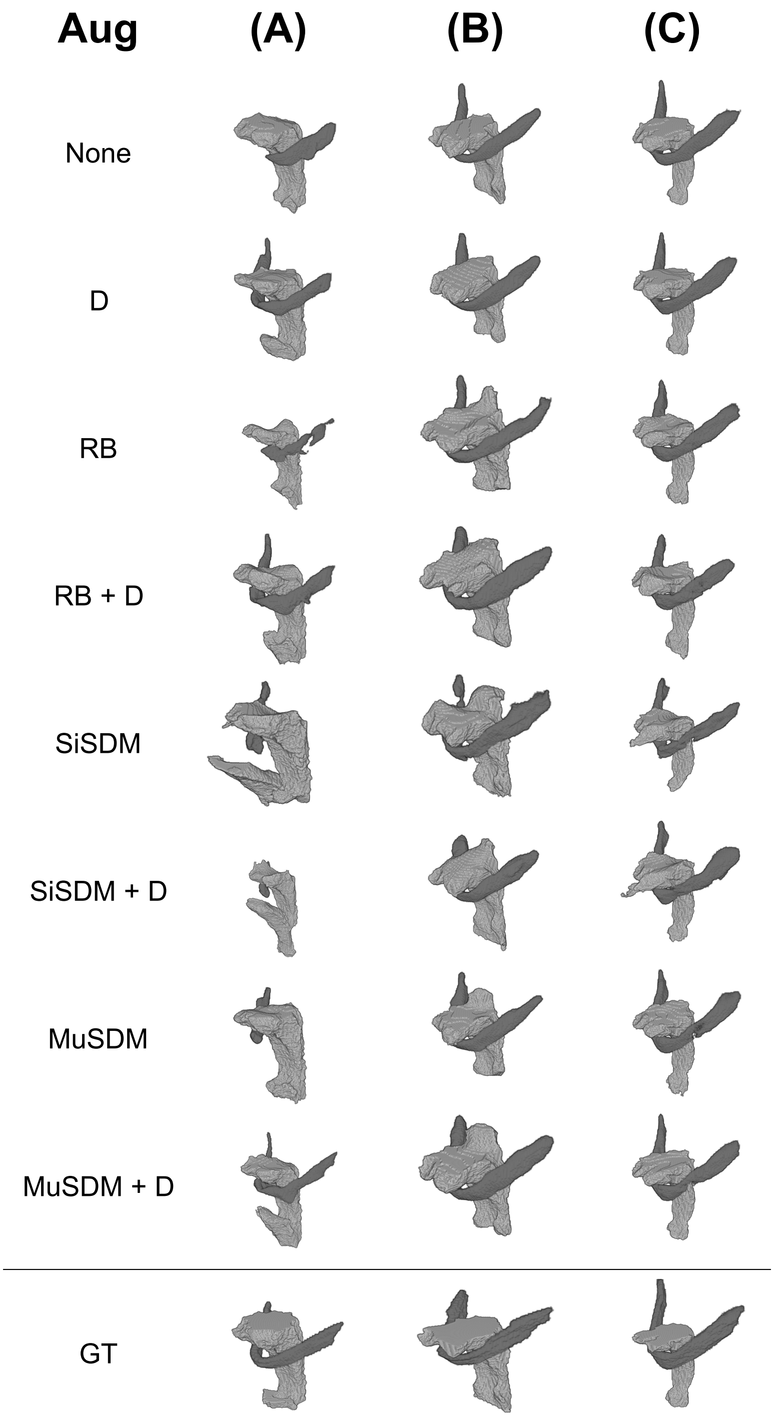

Figure 79: Segmentations estimated by different versions of the proposed method. The "Aug" column indicates the type of data augmentation used during segmentation convolutional neural network training: "None" indicates no augmentation; "D" indicates the default nnU-Net augmentations; "RB" indicates registration-based augmentation; "SiSDM" indicates single statistical deformation model (SDM) based augmentation; "MuSDM" indicates multiple SDM based augmentation; "+ D" indicates that the default nnU-Net augmentations were also used; "GT" indicates ground-truth segmentations. Each column shows segmentations of a different image. Columns (A), (B) and (C) show segmentations with relatively low, average and high Dice coefficients respectively. Dark and light grey indicate the levator veli palatini and pharynx respectively.

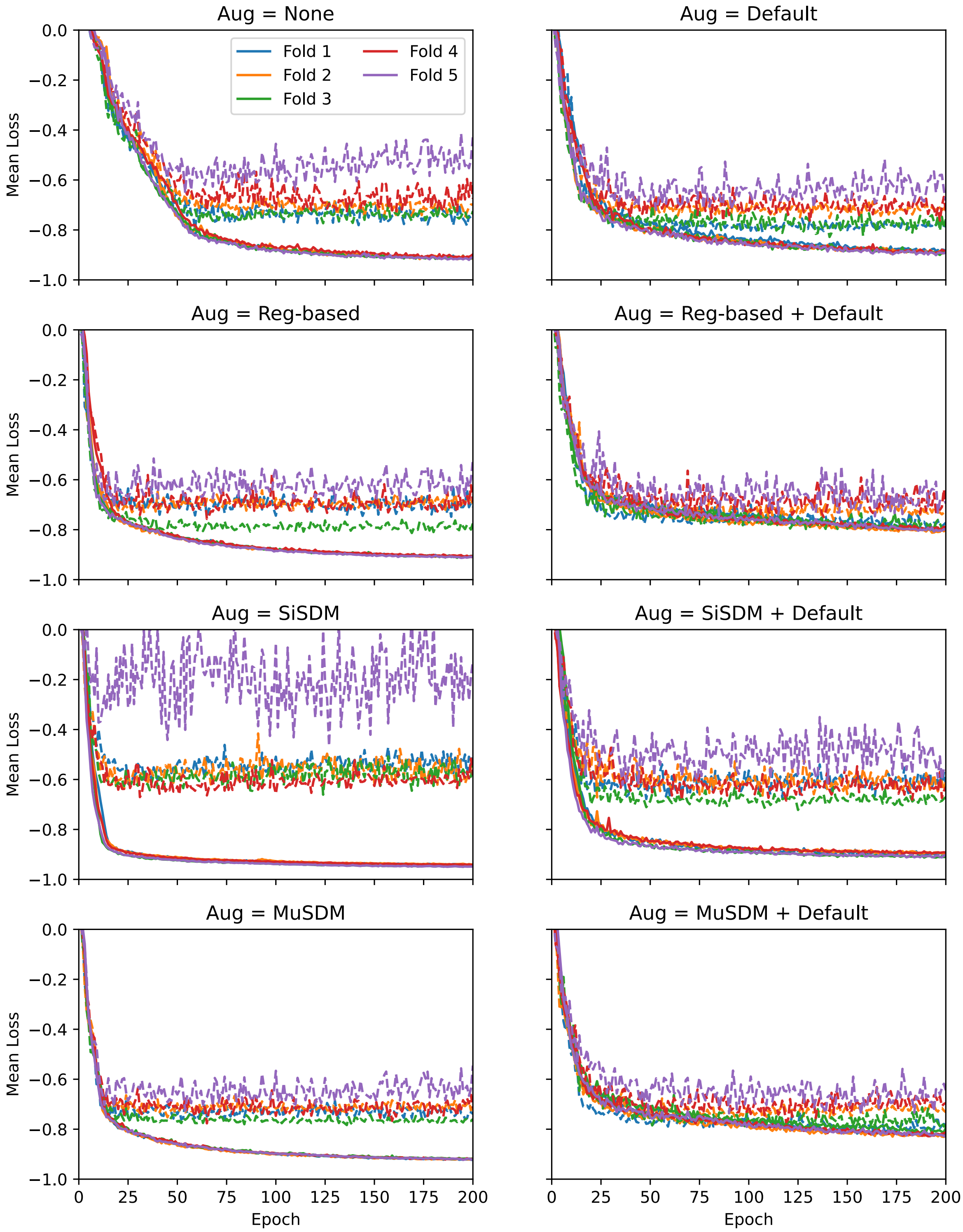

Figure 80: Training and validation losses of the segmentation convolutional neural network (CNN). "Aug" indicates the type of data augmentation used during segmentation convolutional neural network training: "None" indicates no augmentation; "Default" indicates the default nnU-Net augmentation; "Reg-based" indicates registration-based augmentation; "Single SDM" indicates single statistical deformation model (SDM) based augmentation; "Multiple SDM" indicates multiple SDM based augmentation; "+ Default" indicates that the default nnU-Net augmentations were also used. In the figure legend, "Fold" indicates the cross-validation fold. Solid lines indicate training losses, while dashed lines indicate validation losses.

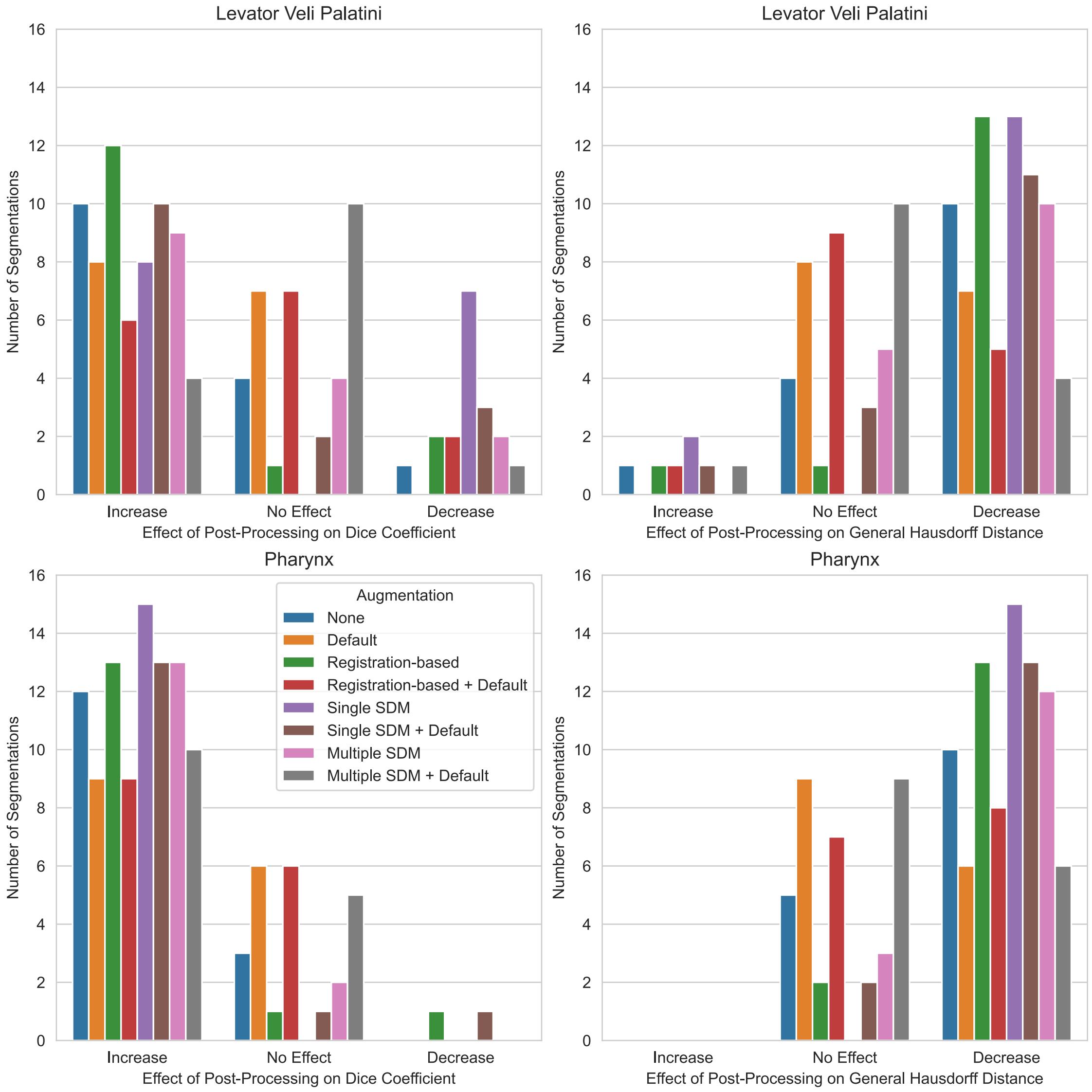

Figure 81: Effect of post-processing step of proposed method on segmentation estimation accuracy. The colour code indicates the type of data augmentation used during segmentation convolutional neural network training: "None" indicates no augmentation; "Default" indicates the default nnU-Net augmentation; "Registration-based" indicates registration-based augmentation; "Single SDM" indicates single statistical deformation model (SDM) based augmentation; "Multiple SDM" indicates multiple SDM based augmentation; "+ Default" indicates that the default nnU-Net augmentations were also used.

The accuracy of the segmentations estimated by different versions of the proposed method are shown in Figure 82. In all cases, the DSC of the LVP segmentation was lower than that of the pharynx segmentation. However, on average (median), in 47% (7 of 15) cases the HD of the LVP segmentation was lower than that of the pharynx segmentation. Two versions of the proposed method consistently segmented both the LVP and pharynx with a lower accuracy than the other methods: the versions where single SDM based data augmentation was used during segmentation CNN training. No version consistently segmented both the LVP and pharynx with a higher accuracy than all the other versions. In fact, there were no statistically significant differences between the accuracies of the segmentations (neither LVP nor pharynx) estimated by the other six versions of the method. On average (median), the DSC and HD of the LVP segmentation estimated by the other six versions of the method was approximately 0.70 and 6 mm respectively, while the DSC and HD of the pharynx segmentation was approximately 0.85 and 6 mm respectively.

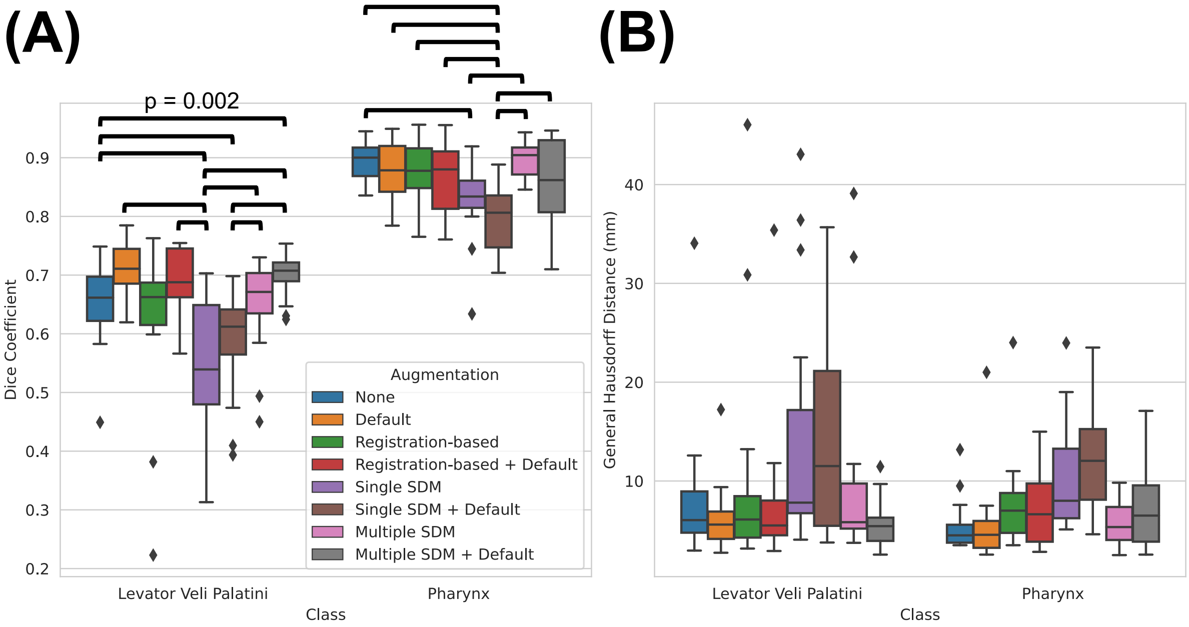

Figure 82: Dice coefficients and general Hausdorff distances of segmentations estimated by different versions of the proposed method. The colour code indicates the type of data augmentation used during segmentation convolutional neural network training: "None" indicates no augmentation; "Default" indicates the default nnU-Net augmentation; "Registration-based" indicates registration-based augmentation; "Single SDM" indicates single statistical deformation model (SDM) based augmentation; "Multiple SDM" indicates multiple SDM based augmentation; "+ Default" indicates that the default nnU-Net augmentations were also used. Black bars above box plots indicate statistically significant differences (5% significance level, p<0.001 unless indicated) between groups of Dice coefficients. There were no statistically significant differences between groups of general Hausdorff distances.

7.5 Discussion

A DL-based method to automatically segment the LVP and pharynx in 3D T_2-weighted MR images of the vocal tract was successfully developed using a state-of-the-art framework (nnU-Net). The method consists of three sequential steps: a pre-processing step to normalise the image voxel values, segmentation estimation using a CNN and then a post-processing step to ensure there is only a single region of voxels per class in the estimated segmentations.

The method segmented the pharynx more accurately than the LVP, with a median DSC of 0.85 and 0.70 respectively and a median HD of 6 mm for both classes. This result is unsurprising for two main reasons. First, the image contrast between the pharynx and the soft tissue that surrounds it is much greater than that of the LVP, thus facilitating pharynx boundary identification. Second, since the LVP is a smaller anatomical feature than the pharynx, its segmentations consist of a smaller number of voxels and segmentation errors therefore have a larger impact on the DSC.

As shown in Figure 80, in almost all cases during segmentation CNN training the validation loss stabilised and then stayed approximately constant in the later stages of training. This result suggests that 200 epochs was a suitable training duration that did not result in the segmentation CNN overfitting the training data, thus justifying the reduction in training duration from the default nnU-Net one of 1000 epochs.

Several different data augmentation methods were investigated to try to improve the generalisation of the proposed method. None of the methods caused a statistically significant improvement in performance compared with no augmentation, when evaluated using the DSC and HD. This result suggests that the methods did not sufficiently increase the anatomical variability in the training dataset images to cause an improvement in the generalisability of the segmentation CNN. This result is not surprising for the default nnU-Net data augmentations as these do not increase anatomical variability.

Registration-based augmentation methods have been shown to cause improvements in the performance of DL-based methods to segment the brain [297,298] and knee [298] in 3D MR images, even when the training dataset is created from a small number of images [298]. In [297,298], the size of each training dataset created using registration-based augmentation methods was at least 1500 images, while in this work the size was 45 images. The former datasets will contain more anatomical variability than the latter. It is therefore possible that a larger training dataset than 45 images is required before a segmentation method performance improvement occurs.

Single SDM based data augmentation caused statistically significant decreases in segmentation method performance. This result was most likely caused by insufficient anatomical variability in the augmented images, as a result of them being synthesised from a single image. In contrast, multiple SDM based data augmentation did not cause a decrease in segmentation method performance.

The LVP and pharynx GT segmentations each consist of a single region of voxels. However, most LVP and pharynx segmentations estimated by the CNN of the proposed method consisted of several unconnected voxel regions. The purpose of the post-processing step of the proposed method is to remove all connected components apart from the largest one. As shown in Figure 81, in most cases the post-processing step improved the accuracy of the estimated segmentations. This shows that most of the unconnected regions in the estimated segmentations were spurious. However, in several cases the post-processing step decreased the accuracy of the estimated LVP segmentations. In these cases, the estimated segmentations consisted of several relatively large unconnected regions as well as relatively small unconnected regions. The removal of these relatively large regions decreased the accuracy of the estimated LVP segmentations (see Figure 79 column (A) for example). A more sophisticated post-processing step which considers the size of each connected component may be able to avoid removing such regions, however, ideally there should only be a single region per class in the segmentations estimated by the CNN. The only version of the proposed method where the step did not reduce the accuracy of any of the segmentations was the one where only the default nnU-Net data augmentations were used during CNN training. However, it should be noted that the step only decreased the accuracy of a single LVP segmentation estimated by the version of the proposed method where no data augmentation was used during CNN training.

In very recent work [17], four state-of-the-art DL-based methods, including a 3D U-Net created using nnU-Net, were used to segment the LVP and five other anatomical features (adenoids, lateral pharyngeal wall, posterior pharyngeal wall, pterygoid raphe and soft palate) in 3D T_1-weighted MR images. The best performing method (3D UX-Net [219]) segmented the LVP with an average DSC of 0.56, while the proposed method, developed to segment 3D T_2-weighted rather than T_1-weighted MR images, segmented the LVP with an average DSC of 0.70. One likely reason for the difference in performance is that images cropped about the vocal tract were used to train the proposed method, while images of the entire head were used in [17]. Identification of the location of the LVP within the latter images is less challenging than in the former, thus making segmentation less challenging. Another likely reason is greater anatomical variability in the dataset used in the previous work, compared with the dataset used in this work. More specifically, the dataset used in previous work included 50 images while the dataset used in this work included 15 images.

An additional possible reason for the difference in performance is that LVP visibility is better in T_2-weighted images than T_1-weighted images, thus making segmentation less challenging.

This work is a step towards the ultimate goal of automatic LVP segmentation and measurement in clinical practice. However, a large amount of future work is required to develop a method suitable for use in clinical practice. More specifically, three major challenges must be overcome. One challenge concerns the dataset used to develop the method, while the other two are technical.

First, as explained in section 4.2.3.2, a larger and more diverse dataset, both in terms of subjects and image acquisition, must be created and used to develop the method. More specifically, a dataset more representative of the target patient population is required: since the target patient population primarily consists of children, the dataset must contain images of children. In addition, since LVP anomalies are prevalent in the target population, the dataset must contain images of LVPs with anomalies as well as LVPs without. In addition, a dataset with images acquired using many different MRI scanners and pulse sequences is required to ensure that methods developed using the dataset perform well on images from different sources. This generalisability is a key requirement for methods suitable for use in clinical practice. While there are publicly available 3D MR image sets of the vocal tract [18,19], these do not have the required image contrast to visualise the LVP.

Second, a method suitable for clinical practice must include any image cropping pre-processing steps so that these steps do not need to be performed separately. The inputs to the proposed method are cropped images centred on the LVP and pharynx. However, to be suitable for clinical practice, the input should be full images. Future work should therefore aim to add a pre-processing step to automatically crop images.

Third, a method suitable for use in clinical practice must automatically measure aspects of the LVP such as its length and thickness. Future work should therefore aim to develop such methods in partnership with clinical teams to ensure that the measured aspects are clinically relevant. The development of such methods could aid the development of segmentation methods by informing the segmentation accuracy required for reliable measurements.

7.6 Conclusions

For the first time, the feasibility of automatic segmentation of the LVP and pharynx in 3D T_2-weighted MR images of the vocal tract has been demonstrated. This work is a step towards the ultimate goal of automatic LVP segmentation and measurement in clinical practice. The effect of different data augmentation methods on the accuracy of the proposed segmentation method was investigated, but none of the methods was found to cause statistically significant improvements in segmentation method accuracy.

Regarding automatic LVP segmentation and measurement, several challenges must still be overcome to enable the development of a method suitable for use in clinical practice. In particular, a larger, more diverse and representative dataset of 3D MR images of the vocal tract must be created, methods to automatically crop such images must be developed, and methods to automatically measure aspects of the LVP must be developed.

Chapter 8: Conclusions

8.1 Summary

The main aim of the work presented in this thesis was to begin to address the unmet need for methods to perform automatic quantitative analysis of the vocal tract, soft palate and LVP in MR images, by developing methods to segment such images and developing a framework for motion quantification.

The first contribution of this work was the creation of GT segmentations of the entire vocal tract and soft palate in an existing speech MRI datasets. Creation of such segmentations was necessary to enable the development and evaluation of segmentation methods due to the lack of publicly available speech MRI datasets that include segmentations. A dataset acquired using a speech MRI technique that does not require specialised MRI equipment and software was deliberately chosen in order to facilitate acquisition of similar images in other centres and consequently the application of image analysis methods developed using the dataset. As described in section 4.1.4, a protocol for creating GT segmentations was devised and used to segment the vocal tract, soft palate and four other anatomical structures in 392 2D rtMR images of speech. Intra-rater agreement in the segmentations was found to be high, suggesting that the GT segmentation creation protocol enabled reproducible results.

The second contribution of this work was the development of a method to segment the entire vocal tract and soft palate and four other anatomical structures in 2D rtMR images of speech. The method, described in chapter 5, has been peer reviewed and published [205] and is a step towards enabling automatic measurement of vocal tract and soft palate size, shape and motion in 2D rtMR images of speech. It was developed and evaluated using the speech MRI dataset and GT segmentations described in section 4.1, and is DL-based, consisting of a CNN to segment images followed by a post-processing step to remove anatomically impossible regions in the images. At the time it was published [205], the method overcame the limitations of existing segmentation methods that either only segmented the air-tissue boundaries between the vocal tract and adjacent tissues or only fully segmented the vocal tract. Since then, three similar methods have been developed [206,207,218], however, the proposed method remains the method that achieved the highest accuracy. The method includes an extension to automatically calculate the minimum distance between the soft palate and the posterior pharyngeal wall, a measurement of particular interest to clinicians who perform speech assessments. Although primarily designed to enable automatic measurement of vocal tract and soft palate size, shape and motion in 2D rtMR images of speech, the 2D segmentation method was designed to also enable measurement of tongue size, shape and motion in order to broaden its potential applications and utility.

The third contribution of this work was the development of a novel metric based on velopharyngeal closure to enable more clinically relevant evaluation of the performance of the image analysis methods. To calculate the metric, GT segmentations are compared with the segmentations estimated by a method. The metric quantifies the number of velopharyngeal closures in the GT segmentations that also occur in the estimated segmentations. In chapter 5 and chapter 6, the metric was shown to be more sensitive to differences in method performance than standard evaluation metrics.

The fourth contribution of this work was the development of a framework for motion estimation in 2D rtMR images of speech. This deep learning framework for nonlinear registration of 2D MR images of speech, described in chapter 6, builds on the 2D segmentation method described in chapter 5 and estimates displacement fields between such images. The framework was developed using the speech MRI datasets and GT segmentations described in section 4.1 and has been peer reviewed and published [291]. It represents another step towards enabling automatic measurement of soft palate motion in 2D rtMR images of speech. The framework was compared with several state-of-the-art traditional registration methods and deep learning frameworks for nonlinear registration and found to estimate displacement fields that more accurately captured velopharyngeal closures. There are currently no other reports in the literature of the application of deep learning frameworks for nonlinear registration to 2D rtMR images of speech.

The fifth contribution of this work was the acquisition of a new MRI dataset and the creation of corresponding GT segmentations of the LVP and pharynx. Acquisition of such a dataset was necessary to enable the development and evaluation of segmentation methods due to the lack of publicly available MRI datasets in which the LVP can be adequately visualised. As described in section 3.1.9, due to the current lack of consensus on the optimal image contrast for visualising the LVP in 3D MR images, an image contrast investigation was performed prior to acquiring the dataset. In this investigation, MR images with different image contrasts were acquired and the visibility of the LVP in these images was compared. The results of this investigation suggested that the LVP was more visible in T_2-weighted images than in T_1-weighted and PD-weighted images. Consequently, a dataset of 15 3D T_2-weighted images of the vocal tract was acquired. The dataset was acquired using a 3D MRI technique that does not require specialised MRI equipment and software in order to facilitate acquisition of similar images in other centres and consequently the application of image analysis methods developed using the dataset. As described in section 4.2.3.1.2, a protocol for creating GT segmentations was devised and used to segment the LVP and pharynx in the images.

The sixth contribution of this work was the development of a method to segment the LVP and pharynx in 3D T_2-weighted MR images of the vocal tract. The method, described in chapter 7, is a step towards enabling automatic measurement of LVP size, shape and configuration in this type of image. It was developed and evaluated using the new dataset and GT segmentations described in section 4.2 and is DL-based, consisting of a CNN to segment images. Until very recently [17], there were no reports in the literature of any methods to segment the LVP. CNNs trained using small amounts of data typically do not generalise well to new data. As only a small amount of data was available to train the segmentation method, attempts were made to improve the generalisation of the method by using data augmentation to increase the size of the training dataset. As described in chapter 7, several different data augmentation methods were investigated, however, none of these were found to improve the generalisation of the method.

8.2 Future work

While the work presented in this thesis makes several contributions towards addressing the unmet need for methods to perform automatic quantitative analysis of the vocal tract, soft palate and LVP in MR images, much future work is required to develop methods suitable for use in clinical assessment of speech. The two main challenges to overcome are firstly to develop methods that automatically perform the specific quantitative analysis of interest to clinicians and secondly to thoroughly evaluate these methods in order for them to be accepted by clinicians and approved for use in clinical practice.

A key requirement to overcome these two main challenges is the creation of larger and more diverse datasets with GT segmentations. More specifically, datasets that are more diverse in terms of subjects and image contrasts are required. Future work should therefore aim to create such datasets. Creation of this type of dataset is essential for two main reasons. First, it would enable the development of DL-based methods that are more generalisable and consequently perform as intended on a wider range of data. Second, it would enable more thorough evaluation of the methods, providing the evidence required for approval of the methods for use in clinical practice and for promotion of trust in the methods by clinicians.

Regarding subjects, the datasets used in this work consisted of healthy adult volunteers. While using such datasets is appropriate for demonstrating the feasibility of developing specific image analysis methods, datasets that better represent the target patient population are required to develop methods suitable for use in clinical practice, especially given that DL-based methods typically perform poorly on data with different characteristics to the data used to train them. Since the target patient population primarily consists of children, datasets that includes images of children are required. In addition, since velopharyngeal closure does not always occur in the speech of patients with VPI, speech MRI datasets containing image series where velopharyngeal closure does not occur as well as image series where it does are required. Since some patients with VPI have an abnormal LVP, MRI datasets that include images of individuals with such an LVP as well as images of individuals with a normal LVP are required to enable the development of generalisable methods to segment the LVP in such images. Finally, creating datasets of images of individuals with a range of demographics is critical in order to enable the development of methods that are generalisable and fair.

Regarding image acquisition, each dataset used in this work consisted of images acquired using a single MRI scanner and pulse sequence. Again, while using such datasets is appropriate for demonstrating the feasibility of specific image analysis methods, datasets of images acquired using many different MRI scanners and pulse sequences are required to develop methods that are more generalisable and therefore more suitable for use in clinical practice.

As part of the work presented in this thesis, two segmentation methods and a motion quantification framework were developed. While the development of these methods and framework are steps towards automatic quantitative analysis of the vocal tract, soft palate and LVP in MR images, future work is required to extend these methods and framework so that they automatically perform the measurements of particular interest to clinicians. More specifically, the 2D image segmentation method should be extended so that it automatically measures the total length, effective length and thickness of the soft palate and the depth of the pharynx [8]. The 3D image segmentation method should be extended so that it automatically measures LVP length and thickness and the distance between origins of the muscle [8]. The motion quantification framework should be extended so that it automatically measures the direction in which the soft palate elevates during speech, the speed at which the soft palate elevates and the distance by which the soft palate elevates.

In conclusion, the work presented in this thesis makes several contributions towards addressing the unmet need for image analysis methods suitable for use in clinical speech assessment. While future work is required to extend the methods and framework presented in this thesis so that they are suitable for use in clinical practice, their development is nevertheless an achievement and has created new opportunities to contribute to the ultimate goal of improving the treatment outcomes of patients with VPI.

References

[1] A.W. Kummer, Speech/Resonance Disorders and Velopharyngeal Dysfunction, in: Cleft Palate Craniofacial Cond. A Compr. Guid. to Clin. Manag., 4th ed., Jones & Bartlett Learning, Burlington, Massachusetts, 2020: pp. 259–293.

[2] C.J. Johnson, J.H. Beitchman, E.B. Brownlie, Twenty-year follow-up of children with and without speech-language impairments: Family, educational, occupational, and quality of life outcomes, Am. J. Speech-Language Pathol. 19 (2010) 51–65. https://doi.org/10.1044/1058-0360(2009/08-0083).

[3] A.W. Kummer, S.L. Clark, E.E. Redle, L.L. Thomsen, D.A. Billmire, Current practice in assessing and reporting speech outcomes of cleft palate and velopharyngeal surgery: A survey of cleft palate/craniofacial professionals, Cleft Palate-Craniofacial J. 49 (2012) 146–152. https://doi.org/10.1597/10-285.

[4] N. Hodgins, C. Hoo, P. McGee, C. Hill, A survey of assessment and management of velopharyngeal incompetence (VPI) in the UK and Ireland, J. Plast. Reconstr. Aesthetic Surg. 68 (2015) 485–491. https://doi.org/10.1016/j.bjps.2014.12.011.

[5] C. de Blacam, S. Smith, D. Orr, Surgery for velopharyngeal dysfunction: A systematic review of interventions and outcomes, Cleft Palate-Craniofacial J. 55 (2018) 405–422. https://doi.org/10.1177/1055665617735102.

[6] N.M. Kurnik, E.M. Weidler, K.M. Lien, K.N. Cordero, J.L. Williams, M. Temkit, S.P. Beals, D.J. Singh, T.J. Sitzman, The Effectiveness of Palate Re-Repair for Treating Velopharyngeal Insufficiency: A Systematic Review and Meta-Analysis, Cleft Palate-Craniofacial J. 57 (2020) 860–871. https://doi.org/10.1177/1055665620902883.

[7] C.T. Arendt, K. Eichler, M.G. Mack, D. Leithner, S. Zhang, K.T. Block, Y. Berdan, R. Sader, J.L. Wichmann, T. Gruber-Rouh, T.J. Vogl, M.C. Hoelter, Comparison of contrast-enhanced videofluoroscopy to unenhanced dynamic MRI in minor patients following surgical correction of velopharyngeal dysfunction, Eur. Radiol. 31 (2021) 76–84. https://doi.org/10.1007/s00330-020-07098-9.

[8] J.L. Perry, T.D. Snodgrass, I.R. Gilbert, B.P. Sutton, A.L. Baylis, E.M. Weidler, R.W. Tse, S.L. Ishman, T.J. Sitzman, Establishing a Clinical Protocol for Velopharyngeal MRI and Interpreting Imaging Findings, Cleft Palate-Craniofacial J. (2022).

https://doi.org/10.1177/10556656221141188.

[9] M. Uecker, S. Zhang, D. Voit, A. Karaus, K.D. Merboldt, J. Frahm, Real-time MRI at a resolution of 20 ms, NMR Biomed. 23 (2010) 986–994. https://doi.org/10.1002/nbm.1585.

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

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

[12] S.G. Lingala, B.P. Sutton, M.E. Miquel, K.S. Nayak, Recommendations for real-time speech MRI, J. Magn. Reson. Imaging. 43 (2016) 28–44. https://doi.org/10.1002/jmri.24997.

[13] J.L. Perry, A.E. Haenssler, K.J. Kotlarek, J.Y. Chen, X. Fang, Y. Guo, K. Mason, M. Webb, Does the Type of MRI Sequence Influence Perceived Quality and Measurement Consistency in Investigations of the Anatomy of the Velopharynx?, Cleft Palate-Craniofacial J. 59 (2022) 741–750. https://doi.org/10.1177/10556656211025191.

[14] V. Ramanarayanan, S. Tilsen, M. Proctor, J. Töger, L. Goldstein, K.S. Nayak, S. Narayanan, Analysis of speech production real-time MRI, Comput. Speech Lang. 52 (2018) 1–22. https://doi.org/10.1016/j.csl.2018.04.002.

[15] M. Labrunie, P. Badin, D. Voit, A.A. Joseph, J. Frahm, L. Lamalle, C. Vilain, L.-J. Boë, Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning, Speech Commun. 99 (2018) 27–46. https://doi.org/10.1016/j.specom.2018.02.004.

[16] K. Somandepalli, A. Toutios, S.S. Narayanan, Semantic Edge Detection for Tracking Vocal Tract Air-tissue Boundaries in Real-time Magnetic Resonance Images, in: INTERSPEECH, 2017: pp. 631–635.

[17] J. Liu, D. Brown, S. Baek, K. Mason, Assessing Deep Learning Methodologies for Automatic Segmentation of the Velopharyngeal Mechanism, in: Proc. Mach. Learn. Res., 2023: pp. 1–5.

[18] Y. Lim, A. Toutios, Y. Bliesener, Y. Tian, S.G. Lingala, C. Vaz, T. Sorensen, M. Oh, S. Harper, W. Chen, Y. Lee, J. Töger, M.L. Monteserin, C. Smith, B. Godinez, L. Goldstein,

D. Byrd, K.S. Nayak, S.S. Narayanan, A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images, Sci. Data. 8 (2021) 1–14. https://doi.org/10.1038/s41597-021-00976-x.

[19] K. Isaieva, Y. Laprie, J. Leclère, I.K. Douros, J. Felblinger, P.A. Vuissoz, Multimodal dataset of real-time 2D and static 3D MRI of healthy French speakers, Sci. Data. 8 (2021) 1–9. https://doi.org/10.1038/s41597-021-01041-3.

[20] K. Isaieva, Y. Laprie, F. Odille, I.K. Douros, J. Felblinger, P.A. Vuissoz, Measurement of tongue tip velocity from real-time MRI and phase-contrast cine-MRI in consonant production, J. Imaging. 6 (2020). https://doi.org/10.3390/JIMAGING6050031.

[21] I.K. Douros, Y. Xie, C. Dourou, K. Isaieva, P.-A. Vuissoz, J. Felblinger, Y. Laprie, 3D Dynamic Spatiotemporal Atlas of the Vocal Tract during Consonant-Vowel Production from 2D Real Time MRI, J. Imaging. 8 (2022) 1–21. https://doi.org/10.3390/jimaging8090227.

[22] H.J. Giegerich, Speech sounds and their production, in: English Phonol. An Introd., Cambridge University Press, 1992: pp. 1–28. https://doi.org/10.1017/cbo9781139166126.002.

[23] Wikipedia, Head neck, (2007). https://commons.wikimedia.org/wiki/File:Illu01_head_neck.jpg (accessed September 21, 2022).

[24] C. Children's, Velopharyngeal Function and Disfunction, (2023). https://www.cincinnatichildrens.org/service/s/speech/specialty-clinics/vpi-clinic/velopharyngeal (accessed March 27, 2023).

[25] J.L. Perry, D.P. Kuehn, B.P. Sutton, Morphology of the Levator Veli Palatini Muscle Using Magnetic Resonance Imaging, Cleft Palate-Craniofacial J. 50 (2013) 64–75. https://doi.org/10.1597/11-125.

[26] B. J.G., S. B.C., Levator palati and palatal dimples: Their anatomy, relationship and clinical significance, Br. J. Plast. Surg. 38 (1985) 326–332. http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L15062370%5Cnhttp://dx.doi.org/10.1016/0007-1226(85)90236-X%5Cnhttp://sfx.library.uu.nl/utrecht?sid=EMBASE&issn=00071226&id=doi:10.1016/0007-1226(85)90236-X&atitle=Levator+palati+and.

[27] A.D. Scott, M. Wylezinska, M.J. Birch, M.E. Miquel, Speech MRI: Morphology and

function, Phys. Medica. 30 (2014) 604–618. https://doi.org/10.1016/j.ejmp.2014.05.001.

[28] E. Olszewska, B.T. Woodson, Palatal anatomy for sleep apnea surgery, Laryngoscope Investig. Otolaryngol. 4 (2019) 181–187. https://doi.org/10.1002/lio2.238.

[29] A.W. Kummer, J.L. Marshall, M.M. Wilson, Non-cleft causes of velopharyngeal dysfunction: Implications for treatment, Int. J. Pediatr. Otorhinolaryngol. 79 (2015) 286–295. https://doi.org/10.1016/j.ijporl.2014.12.036.

[30] J.H. Beitchman, R. Nair, M. Clegg, B. Ferguson, P.G. Patel, Prevalence of Psychiatric Disorders in Children with Speech and Language Disorders, J. Am. Acad. Child Psychiatry. 25 (1986) 528–535. https://doi.org/10.1016/S0002-7138(10)60013-1.

[31] J.H. Beitchman, B. Wilson, E.B. Brownlie, H. Walters, W. Lancee, Long-term consistency in speech/language profiles: I. Developmental and academic outcomes, J. Am. Acad. Child Adolesc. Psychiatry. 35 (1996) 804–814. https://doi.org/10.1097/00004583-199606000-00021.

[32] J.H. Beitchman, B. Wilson, E.B. Brownlie, H. Walters, A. Inglis, W. Lancee, Long-term consistency in speech/language profiles: II. Behavioral, emotional, and social outcomes, J. Am. Acad. Child Adolesc. Psychiatry. 35 (1996) 815–825. https://doi.org/10.1097/00004583-199606000-00022.

[33] J.H. Beitchman, B. Wilson, C.J. Johnson, L. Atkinson, A. Young, E. Adlaf, M. Escobar, L. Douglas, Fourteen-year follow-up of speech/language-impaired and control children: Psychiatric outcome, J. Am. Acad. Child Adolesc. Psychiatry. 40 (2001) 75–82. https://doi.org/10.1097/00004583-200101000-00019.

[34] D.S. Inman, P. Thomas, P.D. Hodgkinson, C.A. Reid, Oro-nasal fistula development and velopharyngeal insufficiency following primary cleft palate surgery - An audit of 148 children born between 1985 and 1997, Br. J. Plast. Surg. 58 (2005) 1051–1054. https://doi.org/10.1016/j.bjps.2005.05.019.

[35] Y.S. Phua, T. De Chalain, Incidence of oronasal fistulae and velopharyngeal insufficiency after cleft palate repair: An audit of 211 children born between 1990 and 2004, Cleft Palate-Craniofacial J. 45 (2008) 172–178. https://doi.org/10.1597/06-205.1.

[36] N. Yuan, A.H. Dorafshar, K.E. Follmar, C. Pendleton, K. Ferguson, R.J. Redett, Effects of cleft width and veau type on incidence of palatal fistula and velopharyngeal

insufficiency after cleft palate repair, Ann. Plast. Surg. 76 (2016) 406–410. https://doi.org/10.1097/SAP.0000000000000407.

[37] J. Rautio, M. Andersen, S. Bolund, J. Hukki, H. Vindenes, P. Davenport, K. Arctander, O. Larson, A. Berggren, F. Åbyholm, D. Whitby, A. Leonard, J. Lilja, E. Neovius, A. Elander, A. Heliövaara, P. Eyres, G. Semb, Scandcleft randomised trials of primary surgery for unilateral cleft lip and palate: 2. Surgical results, J. Plast. Surg. Hand Surg. 51 (2017) 14–20. https://doi.org/10.1080/2000656X.2016.1254646.

[38] L.J. Kobrynski, K.E. Sullivan, Velocardiofacial syndrome, DiGeorge syndrome: the chromosome 22q11.2 deletion syndromes, Lancet. 370 (2007) 1443–1452. https://doi.org/10.1016/S0140-6736(07)61601-8.

[39] Cleft Registry and Audit NEtwork, Annual Report 2021, (2021). https://www.crane-database.org.uk/content/uploads/2021/12/CRANE-2021-Annual-Report_23Dec21.pdf (accessed September 21, 2022).

[40] Cleft Care Scotland, Annual Report 2021/22, (n.d.). https://www.cleftcare.scot.nhs.uk/wp-content/uploads/2022/07/2021-22-Annual-Report_CCS-v1.0.pdf (accessed September 21, 2022).

[41] Wikipedia, Cleft palate 3, (2005). https://commons.wikimedia.org/wiki/File:Cleftpalate3.png (accessed September 21, 2022).

[42] Wikipedia, Cleft lip 2, (2005). https://commons.wikimedia.org/wiki/File:Cleftlip2.svg (accessed September 21, 2022).

[43] Wikipedia, Cleft lip 3, (2005). https://commons.wikimedia.org/wiki/File:CleftLip3.png (accessed September 21, 2022).

[44] Wikipedia, Cleft palate 1, (2005). https://commons.wikimedia.org/wiki/File:Cleftpalate1.png (accessed September 21, 2022).

[45] Wikipedia, Cleft palate 2, (2005). https://commons.wikimedia.org/wiki/File:Cleftpalate2.png (accessed September 21, 2022).

[46] J. Goodship, I. Cross, J. Liling, C. Wren, A population study of chromosome 22q11 deletions in infancy, Arch. Dis. Child. 79 (1998) 348–351. https://doi.org/10.1136/adc.79.4.348.

[47] A.W. Kummer, Surgical Management, in: Cleft Palate Craniofacial Cond. A Compr. Guid. to Clin. Manag., 4th ed., Jones & Bartlett Learning, Burlington, Massachusetts, 2020: pp. 453–494.

[48] F. Åbyholm, L. D'Antonio, S.L. Davidson Ward, L. Kjøll, Pharyngeal Flap and Sphincterplasty for Velopharyngeal Insufficiency ..., Cleft Palate-Craniofacial J. 42 (2005) 501–511.

[49] F. V. Mehendale, M.J. Birch, L. Birkett, D. Sell, B.C. Sommerlad, Surgical Management of Velopharyngeal Incompetence in Velocardiofacial Syndrome, Cleft Palate-Craniofacial J. 41 (2004) 124–135. https://doi.org/10.1597/01-110.

[50] J.A. Perkins, C.W. Lewis, J.S. Gruss, L.E. Eblen, K.C.Y. Sie, Furlow Palatoplasty for Management of Velopharyngeal Insufficiency: A Prospective Study of 148 Consecutive Patients, Plast. Reconstr. Surg. 116 (2005) 72–80. https://doi.org/10.1097/01.PRS.0000169694.29082.69.

[51] S.R. Sullivan, S. Vasudavan, E.M. Marrinan, J.B. Mulliken, Submucous cleft palate and velopharyngeal insufficiency: Comparison of speech outcomes using three operative techniques by one surgeon, Cleft Palate-Craniofacial J. 48 (2011) 561–570. https://doi.org/10.1597/09-127.

[52] J.C.C. Widdershoven, B.M. Stubenitsky, C.C. Breugem, A.B. MinkvanderMolen, Outcome of Velopharyngoplasty in Patients With Velocardiofacial Syndrome, Arch. Otolaryngol. Neck Surg. 134 (2008) 1159. https://doi.org/10.1001/archotol.134.11.1159.

[53] D. Sell, V. Pereira, Instrumentation in the Analysis of the Structure and Function of the Velopharyngeal Mechanism, in: S. Howard, A. Lohmander (Eds.), Cleft Palate Speech Assess. Interv., Wiley-Blackwell, 2011: p. 373.

[54] D.W. McRobbie, E.A. Moore, M.J. Graves, M.R. Prince, Early Daze: Your First Week in MR, in: MRI from Pict. to Prot., 3rd ed., Cambridge University Press, Cambridge, 2017: pp. 11–25. https://doi.org/10.1017/9781107706958.003.

[55] D.W. McRobbie, E.A. Moore, M.J. Graves, M.R. Prince, Getting in Tune: Resonance and Relaxation, in: MRI from Pict. to Prot., 3rd ed., Cambridge University Press, Cambridge, 2017: pp. 124–143. https://doi.org/10.1017/9781107706958.010.

[56] D.W. McRobbie, E.A. Moore, M.J. Graves, M.R. Prince, Spaced out: Spatial Encoding, in: MRI from Pict. to Prot., 3rd ed., Cambridge University Press, Cambridge, 2017: pp.

102–123. https://doi.org/10.1017/9781107706958.009.

[57] D.W. McRobbie, E.A. Moore, M.J. Graves, M.R. Prince, Acronyms Anonymous I: Spin Echo, in: MRI from Pict. to Prot., 3rd ed., Cambridge University Press, Cambridge, 2017: pp. 185–206. https://doi.org/10.1017/9781107706958.013.

[58] D.W. McRobbie, E.A. Moore, M.J. Graves, M.R. Prince, Acronyms Anonymous II: Gradient Echo, in: MRI from Pict. to Prot., 3rd ed., Cambridge University Press, Cambridge, 2017: pp. 207–224. https://doi.org/10.1017/9781107706958.014.

[59] M.A. Bernstein, K.F. King, X.J. Zhou, Introduction to pulse sequences, in: Handb. MRI Pulse Seq., 2004: pp. 573–577. https://doi.org/10.1016/B978-012092861-3/50020-0.

[60] D.W. McRobbie, E.A. Moore, M.J. Graves, M.R. Prince, Lost in the Pulse Sequence Jungle?, in: MRI from Pict. to Prot., 3rd ed., Cambridge University Press, Cambridge, 2017: pp. 41–54. https://doi.org/10.1017/9781107706958.005.

[61] K.S. Nayak, Y. Lim, A.E. Campbell-Washburn, J. Steeden, Real-Time Magnetic Resonance Imaging, J. Magn. Reson. Imaging. 55 (2022) 81–99. https://doi.org/10.1002/jmri.27411.

[62] A.D. Curtis, H.L.M. Cheng, Primer and Historical Review on Rapid Cardiac CINE MRI, J. Magn. Reson. Imaging. 55 (2022) 373–388. https://doi.org/10.1002/jmri.27436.

[63] P.S. Rajiah, C.J. François, T. Leiner, Cardiac MRI : State of the Art, Radiology. 307 (2023) e223008.

[64] M. Stone, A guide to analysing tongue motion from ultrasound images, Clin. Linguist. Phonetics. 19 (2005) 455–501. https://doi.org/10.1080/02699200500113558.

[65] B. Bernhardt, B. Gick, P. Bacsfalvi, M. Adler-Bock, Ultrasound in speech therapy with adolescents and adults, Clin. Linguist. Phonetics. 19 (2005) 605–617. https://doi.org/10.1080/02699200500114028.

[66] J. Cleland, J.M. Scobbie, A.A. Wrench, Using ultrasound visual biofeedback to treat persistent primary speech sound disorders, Clin. Linguist. Phonetics. 29 (2015) 575–597. https://doi.org/10.3109/02699206.2015.1016188.

[67] K. Al-hammuri, F. Gebali, I. Thirumarai Chelvan, A. Kanan, Tongue Contour Tracking and Segmentation in Lingual Ultrasound for Speech Recognition: A Review, Diagnostics. 12 (2022) 1–26. https://doi.org/10.3390/diagnostics12112811.

[68] C. Carignan, R.K. Shosted, M. Fu, Z.P. Liang, B.P. Sutton, A real-time MRI investigation of the role of lingual and pharyngeal articulation in the production of the nasal vowel

system of French, J. Phon. 50 (2015) 34–51. https://doi.org/10.1016/j.wocn.2015.01.001.

[69] D. Carey, M.E. Miquel, B.G. Evans, P. Adank, C. McGettigan, Vocal Tract Images Reveal Neural Representations of Sensorimotor Transformation During Speech Imitation, Cereb. Cortex. 33 (2017) 316–325. https://doi.org/10.1093/cercor/bhw393.

[70] M. Barlaz, R. Shosted, M. Fu, B. Sutton, Oropharygneal articulation of phonemic and phonetic nasalization in Brazilian Portuguese, J. Phon. 71 (2018) 81–97. https://doi.org/10.1016/j.wocn.2018.07.009.

[71] C. Scholes, J.I. Skipper, A. Johnston, The interrelationship between the face and vocal tract configuration during audiovisual speech, Proc. Natl. Acad. Sci. U. S. A. 117 (2020) 32791–32798. https://doi.org/10.1073/pnas.2006192117.

[72] C. Carignan, S. Coretta, J. Frahm, J. Harrington, P. Hoole, A. Joseph, E. Kunay, D. Voit, Planting the seed for sound change: Evidence from real-time MRI of velum kinematics in German, Language (Baltim). 97 (2021) 333–364. https://doi.org/10.1353/lan.2021.0020.

[73] M. Oh, D. Byrd, S.S. Narayanan, Leveraging real-time MRI for illuminating linguistic velum action, in: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, 2021: pp. 3964–3968. https://doi.org/10.21437/Interspeech.2021-1823.

[74] M. Fujimoto, S. Shinohara, D. Mochihashi, Articulation of geminate obstruents in the Ikema dialect of Miyako Ryukyuan: A real-time MRI analysis, J. Int. Phon. Assoc. (2021) 1–25. https://doi.org/10.1017/S0025100321000013.

[75] N. Almeida, S. Silva, C. Cunha, A. Teixeira, Data-Driven Analysis of European Portuguese Nasal Vowel Dynamics in Bilabial Contexts, Appl. Sci. 12 (2022). https://doi.org/10.3390/app12094601.

[76] M. Belyk, S. Waters, E. Kanber, M.E. Miquel, C. McGettigan, Individual differences in vocal size exaggeration, Sci. Rep. 12 (2022) 1–12. https://doi.org/10.1038/s41598-022-05170-6.

[77] A.J. Beer, P. Hellerhoff, A. Zimmermann, K. Mady, R. Sader, E.J. Rummeny, C. Hannig, Dynamic near-real-time magnetic resonance imaging for analyzing the velopharyngeal closure in comparison with videofluoroscopy, J. Magn. Reson. Imaging. 20 (2004) 791–797. https://doi.org/10.1002/jmri.20197.

[78] A.L. Silver, K. Nimkin, J.E. Ashland, S.S. Ghosh, A.J.W. van der Kouwe, M.T. Brigger, C.J.

Hartnick, Cine Magnetic Resonance Imaging With Simultaneous Audio to Evaluate Pediatric Velopharyngeal Insufficiency, Arch. Otolaryngol. Neck Surg. 137 (2011) 258–263. https://doi.org/10.1001/archoto.2011.11.

[79] C. Drissi, M. Mitrofanoff, C. Talandier, C. Falip, V. Le Couls, C. Adamsbaum, Feasibility of dynamic MRI for evaluating velopharyngeal insufficiency in children, Eur. Radiol. 21 (2011) 1462–1469. https://doi.org/10.1007/s00330-011-2069-7.

[80] P. Sagar, K. Nimkin, Feasibility study to assess clinical applications of 3-T cine MRI coupled with synchronous audio recording during speech in evaluation of velopharyngeal insufficiency in children, Pediatr. Radiol. 45 (2015) 217–227. https://doi.org/10.1007/s00247-014-3141-7.

[81] C. Kulinna-Cosentini, C. Czerny, A. Baumann, M. Weber, K. Sinko, TrueFisp versus HASTE sequences in 3T cine MRI: Evaluation of image quality during phonation in patients with velopharyngeal insufficiency, Eur. Radiol. 26 (2016) 2892–2898. https://doi.org/10.1007/s00330-015-4115-3.

[82] J.L. Perry, K. Mason, B.P. Sutton, D.P. Kuehn, Can dynamic MRI be used to accurately identify velopharyngeal closure patterns?, Cleft Palate-Craniofacial J. 55 (2018) 499–507. https://doi.org/10.1177/1055665617735998.

[83] C. Hagedorn, M. Proctor, L. Goldstein, S.M. Wilson, B. Miller, M.L. Gorno-Tempini, S.S. Narayanan, Characterizing articulation in apraxic speech using real-time magnetic resonance imaging, J. Speech, Lang. Hear. Res. 60 (2017) 877–891. https://doi.org/10.1044/2016_JSLHR-S-15-0112.

[84] C.E.E. Wiltshire, M. Chiew, J. Chesters, M.P. Healy, K.E. Watkins, Speech Movement Variability in People Who Stutter: A Vocal Tract Magnetic Resonance Imaging Study, J. Speech, Lang. Hear. Res. 64 (2021) 2438–2452. https://doi.org/10.1044/2021_jslhr-20-00507.

[85] Z. Wu, W. Chen, M.C.K. Khoo, S.L. Davidson Ward, K.S. Nayak, Evaluation of upper airway collapsibility using real-time MRI, J. Magn. Reson. Imaging. 44 (2016) 158–167. https://doi.org/10.1002/jmri.25133.

[86] W. Chen, E. Gillett, M.C.K. Khoo, S.L. Davidson Ward, K.S. Nayak, Real-time multislice MRI during continuous positive airway pressure reveals upper airway response to pressure change, J. Magn. Reson. Imaging. 46 (2017) 1400–1408. https://doi.org/10.1002/jmri.25675.

[87] J. Ha, I. Sung, J. Son, M. Stone, R. Ord, Y. Cho, Analysis of speech and tongue motion in normal and post-glossectomy speaker using cine MRI, J. Appl. Oral Sci. 24 (2016) 472–480. https://doi.org/10.1590/1678-775720150421.

[88] C. Hagedorn, J. Kim, U. Sinha, L. Goldstein, S.S. Narayanan, Complexity of vocal tract shaping in glossectomy patients and typical speakers: A principal component analysis, J. Acoust. Soc. Am. 149 (2021) 4437–4449. https://doi.org/10.1121/10.0004789.

[89] E. Bresch, S. Narayanan, Real-time magnetic resonance imaging investigation of resonance tuning in soprano singing, J. Acoust. Soc. Am. 128 (2010) EL335–EL341. https://doi.org/10.1121/1.3499700.

[90] M. Echternach, F. Burk, M. Burdumy, L. Traser, B. Richter, Morphometric differences of vocal tract articulators in different loudness conditions in singing, PLoS One. 11 (2016) 1–17. https://doi.org/10.1371/journal.pone.0153792.

[91] S. Nishimura, T. Tanaka, M. Oda, M. Habu, M. Kodama, D. Yoshiga, K. Osawa, S. Kokuryo, I. Miyamoto, S. Kito, N. Wakasugi-Sato, S. Matsumoto-Takeda, T. Joujima, Y. Miyamura, S. Hitomi, N. Yamamoto, M. Uehara, M. Sasaguri, K. Ono, I. Yoshioka, K. Tominaga, Y. Morimoto, Functional evaluation of swallowing in patients with tongue cancer before and after surgery using high-speed continuous magnetic resonance imaging based on T2-weighted sequences, Oral Surg. Oral Med. Oral Pathol. Oral Radiol. 125 (2018) 88–98. https://doi.org/10.1016/j.oooo.2017.09.012.

[92] T. Tanaka, R. Tanaka, A.W.K. Yeung, M.M. Bornstein, S. Nishimura, M. Oda, M. Habu, O. Takahashi, D. Yoshiga, T. Sago, I. Miyamoto, M. Kodama, N. Wakasugi-Sato, S. Matsumoto-Takeda, T. Joujima, Y. Miyamura, Y. Morimoto, Real-time evaluation of swallowing in patients with oral cancers by using cine-magnetic resonance imaging based on T2-weighted sequences, Oral Surg. Oral Med. Oral Pathol. Oral Radiol. 130 (2020) 583–592. https://doi.org/10.1016/j.oooo.2020.05.009.

[93] Y.C. Kim, S.J. Lee, H. Park, Y.J. Choi, W.S. Jeong, Y.S. Lee, K.H. Choi, T.S. Oh, J.W. Choi, Swallowing analysis in hemi-tongue reconstruction using motor-innervated free flaps: A cine-magnetic resonance imaging study, Head Neck. (2023) 1097–1112. https://doi.org/10.1002/hed.27309.

[94] M. Belyk, C. McGettigan, Real-time magnetic resonance imaging reveals distinct 21 vocal tract configurations during spontaneous and volitional laughter, Philos. 22 Trans. R. Soc. B Biol. Sci. 377 (2022) 20210511.

https://doi.org/10.1098/rstb.2021.0511.

[95] M. Proctor, E. Bresch, D. Byrd, K. Nayak, S. Narayanan, Paralinguistic mechanisms of production in human “beatboxing”: A real-time magnetic resonance imaging study, J. Acoust. Soc. Am. 133 (2013) 1043–1054. https://doi.org/10.1121/1.4773865.

[96] P. Nimisha, T. Greer, R. Blaylock, S. Narayanan, Comparison of Basic Beatboxing Articulations Between Expert and Novice Artists using Real-Time Magnetic Resonance Imaging Nimis h a P atil , Timot hy G reer , Reed Bla y loc k , Sh ri k ant h Nara y anan S ignal Analysis and Interpretation L aboratory, in: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, 2017: pp. 2277–2281.

[97] P.W. Iltis, E. Schoonderwaldt, S. Zhang, J. Frahm, E. Altenmüller, Real-time MRI comparisons of brass players: A methodological pilot study, Hum. Mov. Sci. 42 (2015) 132–145. https://doi.org/10.1016/j.humov.2015.04.013.

[98] P.W. Iltis, J. Frahm, D. Voit, A.A. Joseph, E. Schoonderwaldt, E. Altenmüller, High-speed real-time magnetic resonance imaging of fast tongue movements in elite horn players., Quant. Imaging Med. Surg. 5 (2015) 374–81. https://doi.org/10.3978/j.issn.2223-4292.2015.03.02.

[99] P.W. Iltis, J. Frahm, D. Voit, A. Joseph, E. Schoonderwaldt, E. Altenmüller, Divergent oral cavity motor strategies between healthy elite and dystonic horn players, J. Clin. Mov. Disord. 2 (2015) 1–9. https://doi.org/10.1186/s40734-015-0027-2.

[100] P.W. Iltis, J. Frahm, D. Voit, A. Joseph, R. Burke, E. Altenmüller, Inefficiencies in motor strategies of horn players with embouchure dystonia comparisons to elite performers, Med. Probl. Perform. Art. 31 (2016) 69–77. https://doi.org/10.21091/mppa.2016.2014.

[101] P.W. Iltis, S.L. Gillespie, J. Frahm, D. Voit, A. Joseph, E. Altenmüller, Movements of the glottis during horn performance: A pilot study, Med. Probl. Perform. Art. 32 (2017) 33–39. https://doi.org/10.21091/mppa.2017.1007.

[102] P.W. Iltis, J. Frahm, D. Voit, A. Joseph, E. Altenmüller, A. Miller, Movements of the tongue during lip trills in horn players: Real-time MRI insights, Med. Probl. Perform. Art. 32 (2017) 209–214. https://doi.org/10.21091/mppa.2017.4042.

[103] H. Furuhashi, T. Chikui, D. Inadomi, T. Shiraishi, K. Yoshiura, Fundamental tongue motions for trumpet playing a study using cine magnetic resonance imaging (Cine MRI), Med. Probl. Perform. Art. 32 (2017) 201–208.

https://doi.org/10.21091/mppa.2017.4038.

[104] P.W. Iltis, M. Heyne, J. Frahm, D. Voit, A. Joseph, L. Atlas, Simultaneous dual-plane, real-time magnetic resonance imaging of oral cavity movements in advanced trombone players, Quant. Imaging Med. Surg. 9 (2019) 976–984. https://doi.org/10.21037/qims.2019.05.14.

[105] S.J. Hellwig, P.W. Iltis, A.A. Joseph, D. Voit, J. Frahm, E. Schoonderwaldt, E. Altenmüller, Tongue involvement in embouchure dystonia: new piloting results using real-time MRI of trumpet players, J. Clin. Mov. Disord. 6 (2019) 1–8. https://doi.org/10.1186/s40734-019-0080-3.

[106] P.W. Iltis, J. Frahm, E. Altenmüller, D. Voit, A. Joseph, K. Kozakowski, Tongue position variability during sustained notes in healthy vs dystonic horn players using real-time MRI, Med. Probl. Perform. Art. 34 (2019) 33–38. https://doi.org/10.21091/mppa.2019.1007.

[107] P.W. Iltis, J. Frahm, D. Voit, D. Wood, S. Taylor, Oral Cavity Movements of the Tongue during Large Interval Slurs in High-Level Horn Players: A Descriptive Study, Med. Probl. Perform. Art. 37 (2022) 89–97. https://doi.org/10.21091/mppa.2022.2014.

[108] R. Nelkenstock, P.W. Iltis, D. Voit, J. Frahm, Movement patterns in tuba playing : comparison of an embouchure dystonia case with healthy professional tuba players using real-time MRI imaging, Front. Neurol. 14 (2023). https://doi.org/10.3389/fneur.2023.1106217.

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

[110] R. Jin, R.K. Shosted, F. Xing, I.R. Gilbert, J.L. Perry, J. Woo, Z.P. Liang, B.P. Sutton, Enhancing linguistic research through 2-mm isotropic 3D dynamic speech MRI optimized by sparse temporal sampling and low-rank reconstruction, Magn. Reson. Med. 89 (2023) 652–664. https://doi.org/10.1002/mrm.29486.

[111] M. Burdumy, L. Traser, F. Burk, B. Richter, M. Echternach, J.G. Korvink, J. Hennig, M. Zaitsev, One-second MRI of a three-dimensional vocal tract to measure dynamic articulator modifications, J. Magn. Reson. Imaging. 46 (2017) 94–101. https://doi.org/10.1002/jmri.25561.

[112] M. Fu, M.S. Barlaz, J.L. Holtrop, J.L. Perry, D.P. Kuehn, R.K. Shosted, Z.P. Liang, B.P. Sutton, High-frame-rate full-vocal-tract 3D dynamic speech imaging, Magn. Reson. Med. 77 (2017) 1619–1629. https://doi.org/10.1002/mrm.26248.

[113] Y. Lim, Y. Zhu, S.G. Lingala, D. Byrd, S. Narayanan, K.S. Nayak, 3D dynamic MRI of the vocal tract during natural speech, Magn. Reson. Med. (2018) 1–10. https://doi.org/10.1002/mrm.27570.

[114] A.C. Freitas, M. Wylezinska, M.J. Birch, S.E. Petersen, M.E. Miquel, Comparison of Cartesian and Non-Cartesian Real-Time MRI Sequences at 1.5T to Assess Velar Motion and Velopharyngeal Closure during Speech, PLoS One. 11 (2016) e0153322. https://doi.org/10.1371/journal.pone.0153322.

[115] A.C. Freitas, M. Ruthven, R. Boubertakh, M.E. Miquel, Real-time speech MRI: Commercial Cartesian and non-Cartesian sequences at 3T and feasibility of offline TGV reconstruction to visualise velopharyngeal motion, Phys. Medica. 46 (2018) 96–103. https://doi.org/10.1016/j.ejmp.2018.01.014.

[116] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, M. Proctor, Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC), J. Acoust. Soc. Am. 136 (2014) 1307–1311. https://doi.org/10.1121/1.4890284.

[117] T. Sorensen, Z. Skordilis, A. Toutios, Y.-C. Kim, Y. Zhu, J. Kim, A. Lammert, V. Ramanarayanan, L. Goldstein, D. Byrd, K. Nayak, S. Narayanan, Database of volumetric and real-time vocal tract MRI for speech science, in: INTERSPEECH, 2017: pp. 645–649. https://doi.org/10.21437/Interspeech.2017-608.

[118] I.K. Douros, J. Felblinger, J. Frahm, K. Isaieva, A.A. Joseph, Y. Laprie, F. Odille, A. Tsukanova, D. Voit, P.A. Vuissoz, A multimodal real-time MRI articulatory corpus of French for speech research, in: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, 2019: pp. 1556–1560. https://doi.org/10.21437/Interspeech.2019-1700.

[119] J. Kim, A. Toutios, Y.C. Kim, Y. Zhu, S. Lee, S. Narayanan, USC-EMO-MRI corpus: An emotional speech production database recorded by real-time magnetic resonance imaging, in: Proc. 10th Int. Semin. Speech Prod. 2014, 2014: pp. 226–229.

[120] J. Töger, T. Sorensen, K. Somandepalli, A. Toutios, S.G. Lingala, S. Narayanan, K.

Nayak, Test–retest repeatability of human speech biomarkers from static and real-time dynamic magnetic resonance imaging, J. Acoust. Soc. Am. 141 (2017) 3323–3336. https://doi.org/10.1121/1.4983081.

[121] C. McGettigan, M. Miquel, D. Carey, S. Waters, E. Kanber, Vocal Learning in Adulthood: Investigating the mechanisms of vocal imitation using MRI of the vocal tract and brain 2015-2018, UK Data Serv. (2018). https://doi.org/10.5255/UKDA-SN-853317.

[122] S.L. Ettema, D.P. Kuehn, A.L. Perlman, N. Alperin, Magnetic resonance imaging of the levator veli palatini muscle during speech, Cleft Palate-Craniofacial J. 39 (2002) 130–144. https://doi.org/10.1597/1545-1569(2002)039<0130:MRIOTL>2.0.CO;2.

[123] D.P. Kuehn, S.L. Ettema, M.S. Goldwasser, J.C. Barkmeier, Magnetic resonance imaging of the levator veli palatini muscle before and after primary palatoplasty, Cleft Palate-Craniofacial J. 41 (2004) 584–592. https://doi.org/10.1597/03-060.1.

[124] S. Ha, D.P. Kuehn, M. Cohen, N. Alperin, Magnetic resonance imaging of the levator veli palatini muscle in speakers with repaired cleft palate, Cleft Palate-Craniofacial J. 44 (2007) 494–505. https://doi.org/10.1597/06-220.1.

[125] J.L. Perry, Variations in velopharyngeal structures between upright and supine positions using upright magnetic resonance imaging, Cleft Palate-Craniofacial J. 48 (2011) 123–133. https://doi.org/10.1597/09-256.

[126] J.L. Perry, D.P. Kuehn, B.P. Sutton, M.S. Goldwasser, A.D. Jerez, Craniometric and velopharyngeal assessment of infants with and without cleft palate, J. Craniofac. Surg. 22 (2011) 499–503. https://doi.org/10.1097/SCS.0b013e3182087378.

[127] Y. Bae, D.P. Kuehn, B.P. Sutton, C.A. Conway, J.L. Perry, Three-dimensional magnetic resonance imaging of velopharyngeal structures, J. Speech, Lang. Hear. Res. 54 (2011) 1538–1545. https://doi.org/10.1044/1092-4388(2011/10-0021).

[128] J.L. Perry, B.P. Sutton, D.P. Kuehn, J.K. Gamage, Using MRI for assessing velopharyngeal structures and function, Cleft Palate-Craniofacial J. 51 (2014) 476–485. https://doi.org/10.1597/12-083.

[129] J.L. Perry, D.P. Kuehn, B.P. Sutton, J.K. Gamage, Sexual Dimorphism of the Levator Veli Palatini Muscle: An Imaging Study, Cleft Palate-Craniofacial J. 51 (2014) 544–552. https://doi.org/10.1597/12-128.

[130] L. Kollara, J.L. Perry, Effects of gravity on the velopharyngeal structures in children

using upright magnetic resonance imaging, Cleft Palate-Craniofacial J. 51 (2014) 669–676. https://doi.org/10.1597/13-107.

[131] J.L. Perry, D.P. Kuehn, B.P. Sutton, J.K. Gamage, X. Fang, Anthropometric analysis of the velopharynx and related craniometric dimensions in three adult populations using MRI, Cleft Palate-Craniofacial J. 53 (2016) e1–e13. https://doi.org/10.1597/14-015.

[132] L. Kollara, J.L. Perry, S. Hudson, Racial Variations in Velopharyngeal and Craniometric Morphology in Children: An Imaging Study, J. Speech, Lang. Hear. Res. 59 (2016) 27–38. https://doi.org/10.1044/2015_JSLHR-S-14-0236.

[133] G.C. Schenck, J.L. Perry, X. Fang, Normative velopharyngeal data in infants: Implications for treatment of cleft palate, J. Craniofac. Surg. 27 (2016) 1430–1439. https://doi.org/10.1097/SCS.0000000000002722.

[134] K.J. Kotlarek, J.L. Perry, X. Fang, Morphology of the levator veli palatini muscle in adults with repaired cleft palate, J. Craniofac. Surg. 28 (2017) 833–837. https://doi.org/10.1097/SCS.0000000000003373.

[135] W. Tian, H. Yin, R.J. Redett, B. Shi, J. Shi, R. Zhang, Q. Zheng, Magnetic resonance imaging assessment of the velopharyngeal mechanism at rest and during speech in Chinese adults and children, J. Speech, Lang. Hear. Res. 53 (2010) 1595–1615. https://doi.org/10.1044/1092-4388(2010/09-0105).

[136] J.L. Perry, D.P. Kuehn, B.P. Sutton, X. Fang, Velopharyngeal structural and functional assessment of speech in young children using dynamic magnetic resonance imaging, Cleft Palate-Craniofacial J. 54 (2017) 408–422. https://doi.org/10.1597/15-120.

[137] J.L. Perry, L. Kollara, D.P. Kuehn, B.P. Sutton, X. Fang, Examining age, sex, and race characteristics of velopharyngeal structures in 4- to 9-year-old children using magnetic resonance imaging, Cleft Palate-Craniofacial J. 55 (2018) 21–34. https://doi.org/10.1177/1055665617718549.

[138] J.L. Perry, K.J. Kotlarek, B.P. Sutton, D.P. Kuehn, M.S. Jaskolka, X. Fang, S.W. Point, F. Rauccio, Variations in velopharyngeal structure in adults with repaired cleft palate, Cleft Palate-Craniofacial J. 55 (2018) 1409–1418. https://doi.org/10.1177/1055665617752803.

[139] L. Kollara, A.L. Baylis, R.E. Kirschner, D.G. Bates, M. Smith, X. Fang, J.L. Perry, Velopharyngeal Structural and Muscle Variations in Children With 22q11.2 Deletion Syndrome: An Unsedated MRI Study, Cleft Palate-Craniofacial J. 56 (2019) 1139–1148.

https://doi.org/10.1177/1055665619851660.

[140] G.C. Schenck, J.L. Perry, M.M. O'Gara, A.M. Linde, M.F. Grasseschi, R.J. Wood, M.S. Lacey, X. Fang, Velopharyngeal Muscle Morphology in Children With Unrepaired Submucous Cleft Palate: An Imaging Study, Cleft Palate-Craniofacial J. 58 (2021) 313–323. https://doi.org/10.1177/1055665620954749.

[141] N. Tahmasebifard, C. Ellis, K. Rothermich, X. Fang, J.L. Perry, Evaluation of the Symmetry of the Levator Veli Palatini Muscle and Velopharyngeal Closure Among a Noncleft Adult Population, Cleft Palate-Craniofacial J. 58 (2021) 728–735. https://doi.org/10.1177/1055665620961269.

[142] M. Park, S.H. Ahn, J.H. Jeong, R.M. Baek, Evaluation of the levator veli palatini muscle thickness in patients with velocardiofacial syndrome using magnetic resonance imaging, J. Plast. Reconstr. Aesthetic Surg. 68 (2015) 1100–1105. https://doi.org/10.1016/j.bjps.2015.04.013.

[143] C. Filip, D. Impieri, I. Aagenæs, C. Breugem, H.E. Høgevold, T. Særvold, R. Aukner, K. Lima, K. Tønseth, T.G. Abrahamsen, Adults with 22q11.2 deletion syndrome have a different velopharyngeal anatomy with predisposition to velopharyngeal insufficiency, J. Plast. Reconstr. Aesthetic Surg. 71 (2018) 524–536. https://doi.org/10.1016/j.bjps.2017.09.006.

[144] T. Tran, J. Perry, S. Blemker, K. Mason, Simulation of Velopharyngeal Biomechanics Identifies Differences in Sphincter Pharyngoplasty Outcomes: A Matched Case–Control Study, Cleft Palate-Craniofacial J. (2022) 105566562211226. https://doi.org/10.1177/10556656221122634.

[145] K.N. Mason, Magnetic Resonance Imaging for Assessing Velopharyngeal Function: Current Applications, Barriers, and Potential for Future Clinical Translation in the United States, Cleft Palate-Craniofacial J. (2022). https://doi.org/10.1177/10556656221123916.

[146] D.J. Merlino, C.J. Vander Wert, A.B. Sauer, L.X. Yin, E.J. Moore, J.M. Morris, K.M. Van Abel, Detailed 3-dimensional surgical anatomy of the soft palate: a confluence of anatomy, radiology, and medical illustration, Oper. Tech. Otolaryngol. Neck Surg. 33 (2022) 272–280. https://doi.org/10.1016/j.otot.2022.10.007.

[147] D.W. McRobbie, E.A. Moore, M.J. Graves, M.R. Prince, Ghosts in the Machine: Quality Control, in: MRI from Pict. to Prot., 3rd ed., Cambridge University Press, Cambridge,

2017: pp. 166–182. https://doi.org/10.1017/9781107706958.012.

[148] I. Goodfellow, Y. Bengio, A. Courville, Machine Learning Basics, in: Deep Learn., MIT Press, 2016: pp. 96–161. https://www.deeplearningbook.org/contents/ml.html.

[149] T.M. Mitchell, Introduction, in: Mach. Learn., McGraw-Hill, 1997: pp. 1–19.

[150] I. Goodfellow, Y. Bengio, A. Courville, Deep Feedforward Networks, in: Deep Learn., MIT Press, 2016: pp. 164–223.

[151] D.P. Kingma, J.L. Ba, Adam: A method for stochastic optimization, in: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., 2015.

[152] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.

[153] C. Shorten, T.M. Khoshgoftaar, A survey on Image Data Augmentation for Deep Learning, J. Big Data. 6 (2019). https://doi.org/10.1186/s40537-019-0197-0.

[154] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J.N. Chiang, Z. Wu, X. Ding, Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation, Med. Image Anal. 63 (2020) 101693. https://doi.org/10.1016/j.media.2020.101693.

[155] I. Goodfellow, Y. Bengio, A. Courville, Convolutional Networks, in: Deep Learn., MIT Press, 2016: pp. 326–366.

[156] S. Ioffe, C. Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, in: Proc. 32nd Int. Conf. Mach. Learn., 2015: pp. 448–456. http://proceedings.mlr.press/v37/ioffe15.pdf.

[157] D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance Normalization: The Missing Ingredient for Fast Stylization, (2016). http://arxiv.org/abs/1607.08022.

[158] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, ICML Work. Deep Learn. Audio, Speech Lang. Process. 28 (2013).

[159] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image Segmentation Using Deep Learning: A Survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2022) 3523–3542. https://doi.org/10.1049/ipr2.12419.

[160] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: Proc. Conf. Neural Inf. Process. Syst., 2012: pp. 1106–1114. https://doi.org/10.1201/9781420010749.

[161] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image

Recognition, in: Int. Conf. Learn. Represent., 2015: pp. 1–14. http://arxiv.org/abs/1409.1556.

[162] J. Long, E. Shelhamer, T. Darrell, Fully Convolutional Networks for Semantic Segmentation, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., IEEE, 2015: pp. 3431–3440. https://doi.org/10.1109/CVPR.2018.00712.

[163] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2016: pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

[164] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is All you Need, in: Adv. Neural Inf. Process. Syst., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[165] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: Int. Conf. Learn. Represent., 2021. http://arxiv.org/abs/2010.11929.

[166] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, D. Shen, Transformers in medical image analysis, Intell. Med. 3 (2023) 59–78. https://doi.org/10.1016/j.imed.2022.07.002.

[167] F. Shamshad, S. Khan, S.W. Zamir, M.H. Khan, M. Hayat, F.S. Khan, H. Fu, Transformers in Medical Imaging: A Survey, Med. Image Anal. 88 (2023) 102802. https://doi.org/10.1016/j.media.2023.102802.

[168] P. Meyer, V. Noblet, C. Mazzara, A. Lallement, Survey on deep learning for radiotherapy, Comput. Biol. Med. 98 (2018) 126–146. https://doi.org/10.1016/j.compbiomed.2018.05.018.

[169] M. Field, N. Hardcastle, M. Jameson, N. Aherne, L. Holloway, Machine learning applications in radiation oncology, Phys. Imaging Radiat. Oncol. 19 (2021) 13–24. https://doi.org/10.1016/j.phro.2021.05.007.

[170] S.K. Zhou, H. Greenspan, C. Davatzikos, J.S. Duncan, B. Van Ginneken, A. Madabhushi, J.L. Prince, D. Rueckert, R.M. Summers, A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies with Progress Highlights, and Future Promises, Proc. IEEE. 109 (2021) 820–838.

https://doi.org/10.1109/JPROC.2021.3054390.

[171] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, D. Rueckert, Deep learning for cardiac image segmentation: A review, Front. Cardiovasc. Med. 7 (2020) 1–33. https://doi.org/10.3389/fcvm.2020.00025.

[172] M. Sermesant, H. Delingette, H. Cochet, P. Jaïs, N. Ayache, Applications of artificial intelligence in cardiovascular imaging, Nat. Rev. Cardiol. 18 (2021). https://doi.org/10.1038/s41569-021-00527-2.

[173] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88. https://doi.org/10.1016/j.media.2017.07.005.

[174] X. Chen, X. Wang, K. Zhang, K.M. Fung, T.C. Thai, K. Moore, R.S. Mannel, H. Liu, B. Zheng, Y. Qiu, Recent advances and clinical applications of deep learning in medical image analysis, Med. Image Anal. 79 (2022) 102444. https://doi.org/10.1016/j.media.2022.102444.

[175] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Med. Image Comput. Comput. Interv., Springer, 2015: pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

[176] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation., IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.

[177] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-net: Learning dense volumetric segmentation from sparse annotation, in: Med. Image Comput. Comput. Interv., 2016: pp. 424–432. https://doi.org/10.1007/978-3-319-46723-8_49.

[178] F. Milletari, N. Navab, S.A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: Proc. - 2016 4th Int. Conf. 3D Vision, 3DV 2016, 2016: pp. 565–571. https://doi.org/10.1109/3DV.2016.79.

[179] R.P.K. Poudel, P. Lamata, G. Montana, Recurrent Fully Convolutional Neural Networks for Multi-slice MRI Cardiac Segmentation, in: Reconstr. Segmentation, Anal. Med. Images, 2017: pp. 83–94. https://doi.org/10.1007/978-3-319-52280-7_8.

[180] J. Chen, L. Yang, Y. Zhang, M. Alber, D.Z. Chen, Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation, Adv. Neural Inf.

Process. Syst. (2016) 3044–3052.

[181] Q. Zheng, H. Delingette, N. Duchateau, N. Ayache, 3-D Consistent and Robust Segmentation of Cardiac Images by Deep Learning With Spatial Propagation, IEEE Trans. Med. Imaging. 37 (2018) 2137–2148. https://doi.org/10.1109/TMI.2018.2820742.

[182] D.M. Vigneault, W. Xie, C.Y. Ho, D.A. Bluemke, J.A. Noble, Ω-Net (Omega-Net): Fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks, Med. Image Anal. 48 (2018) 95–106. https://doi.org/10.1016/j.media.2018.05.008.

[183] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, (2021) 1–13. http://arxiv.org/abs/2102.04306.

[184] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H.R. Roth, D. Xu, UNETR: Transformers for 3D Medical Image Segmentation, Proc. - 2022 IEEE/CVF Winter Conf. Appl. Comput. Vision, WACV 2022. (2022) 1748–1758. https://doi.org/10.1109/WACV51458.2022.00181.

[185] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H.R. Roth, D. Xu, Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Springer International Publishing, 2022: pp. 272–284. https://doi.org/10.1007/978-3-031-08999-2_22.

[186] N. Siddique, S. Paheding, C.P. Elkin, V. Devabhaktuni, U-net and its variants for medical image segmentation: A review of theory and applications, IEEE Access. 9 (2021) 82031–82057. https://doi.org/10.1109/ACCESS.2021.3086020.

[187] F. Isensee, P.F. Jaeger, S.A.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods. 18 (2021) 203–211. https://doi.org/10.1038/s41592-020-01008-z.

[188] V.M. Campello, P. Gkontra, C. Izquierdo, C. Martin-Isla, A. Sojoudi, P.M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, M. Parreno, A. Albiol, F. Kong, S.C. Shadden, J.C. Acero, V. Sundaresan, M. Saber, M. Elattar, H. Li, B. Menze, F. Khader, C. Haarburger, C.M. Scannell, M. Veta, A. Carscadden, K. Punithakumar, X. Liu, S.A. Tsaftaris, X. Huang, X. Yang, L. Li, X. Zhuang, D. Vilades, M.L. Descalzo, A. Guala, L. La Mura, M.G. Friedrich,

R. Garg, J. Lebel, F. Henriques, M. Karakas, E. Cavus, S.E. Petersen, S. Escalera, S. Segui, J.F. Rodriguez-Palomares, K. Lekadir, Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The MMs Challenge, IEEE Trans. Med. Imaging. 40 (2021) 3543–3554. https://doi.org/10.1109/TMI.2021.3090082.

[189] N. Heller, F. Isensee, K.H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han, G. Yao, Y. Gao, Y. Zhang, Y. Wang, F. Hou, J. Yang, G. Xiong, J. Tian, C. Zhong, J. Ma, J. Rickman, J. Dean, B. Stai, R. Tejpaul, M. Oestreich, P. Blake, H. Kaluzniak, S. Raza, J. Rosenberg, K. Moore, E. Walczak, Z. Rengel, Z. Edgerton, R. Vasdev, M. Peterson, S. McSweeney, S. Peterson, A. Kalapara, N. Sathianathen, N. Papanikolopoulos, C. Weight, The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge, Med. Image Anal. 67 (2021) 101821. https://doi.org/10.1016/j.media.2020.101821.

[190] R. Dorent, A. Kujawa, M. Ivory, S. Bakas, N. Rieke, S. Joutard, B. Glocker, J. Cardoso, M. Modat, K. Batmanghelich, A. Belkov, M.B. Calisto, J.W. Choi, B.M. Dawant, H. Dong, S. Escalera, Y. Fan, L. Hansen, M.P. Heinrich, S. Joshi, V. Kashtanova, H.G. Kim, S. Kondo, C.N. Kruse, S.K. Lai-Yuen, H. Li, H. Liu, B. Ly, I. Oguz, H. Shin, B. Shirokikh, Z. Su, G. Wang, J. Wu, Y. Xu, K. Yao, L. Zhang, S. Ourselin, J. Shapey, T. Vercauteren, CrossMoDA 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation, Med. Image Anal. 83 (2023) 102628. https://doi.org/10.1016/j.media.2022.102628.

[191] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B.A. Landman, G. Litjens, B. Menze, O. Ronneberger, R.M. Summers, B. van Ginneken, M. Bilello, P. Bilic, P.F. Christ, R.K.G. Do, M.J. Gollub, S.H. Heckers, H. Huisman, W.R. Jarnagin, M.K. McHugo, S. Napel, J.S.G. Pernicka, K. Rhode, C. Tobon-Gomez, E. Vorontsov, J.A. Meakin, S. Ourselin, M. Wiesenfarth, P. Arbeláez, B. Bae, S. Chen, L. Daza, J. Feng, B. He, F. Isensee, Y. Ji, F. Jia, I. Kim, K. Maier-Hein, D. Merhof, A. Pai, B. Park, M. Perslev, R. Rezaiifar, O. Rippel, I. Sarasua, W. Shen, J. Son, C. Wachinger, L. Wang, Y. Wang, Y. Xia, D. Xu, Z. Xu, Y. Zheng, A.L. Simpson, L. Maier-Hein, M.J. Cardoso, The Medical Segmentation Decathlon, Nat. Commun. 13 (2022) 1–13. https://doi.org/10.1038/s41467-022-30695-9.

[192] H.R. Roth, Z. Xu, C. Tor-Díez, R. Sanchez Jacob, J. Zember, J. Molto, W. Li, S. Xu, B. Turkbey, E. Turkbey, D. Yang, A. Harouni, N. Rieke, S. Hu, F. Isensee, C. Tang, Q. Yu, J.

Sölter, T. Zheng, V. Liauchuk, Z. Zhou, J.H. Moltz, B. Oliveira, Y. Xia, K.H. Maier-Hein, Q. Li, A. Husch, L. Zhang, V. Kovalev, L. Kang, A. Hering, J.L. Vilaça, M. Flores, D. Xu, B. Wood, M.G. Linguraru, Rapid artificial intelligence solutions in a pandemic—The COVID-19-20 Lung CT Lesion Segmentation Challenge, Med. Image Anal. 82 (2022) 102605. https://doi.org/10.1016/j.media.2022.102605.

[193] A. Reinke, M.D. Tizabi, C.H. Sudre, M. Eisenmann, T. Rädsch, M. Baumgartner, L. Acion, M. Antonelli, T. Arbel, S. Bakas, P. Bankhead, A. Benis, M.J. Cardoso, V. Cheplygina, E. Christodoulou, B. Cimini, G.S. Collins, K. Farahani, B. van Ginneken, B. Glocker, P. Godau, F. Hamprecht, D.A. Hashimoto, D. Heckmann-Nötzel, M.M. Hoffman, M. Huisman, F. Isensee, P. Jannin, C.E. Kahn, A. Karargyris, A. Karthikesalingam, B. Kainz, E. Kavur, H. Kenngott, J. Kleesiek, T. Kooi, M. Kozubek, A. Kreshuk, T. Kurc, B.A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A.L. Martel, P. Mattson, E. Meijering, B. Menze, D. Moher, K.G.M. Moons, H. Müller, B. Nichyporuk, F. Nickel, M.A. Noyan, J. Petersen, G. Polat, N. Rajpoot, M. Reyes, N. Rieke, M. Riegler, H. Rivaz, J. Saez-Rodriguez, C.S. Gutierrez, J. Schroeter, A. Saha, S. Shetty, M. van Smeden, B. Stieltjes, R.M. Summers, A.A. Taha, S.A. Tsaftaris, B. Van Calster, G. Varoquaux, M. Wiesenfarth, Z.R. Yaniv, A. Kopp-Schneider, P. Jäger, L. Maier-Hein, Common Limitations of Image Processing Metrics: A Picture Story, ArXiv. (2022). http://arxiv.org/abs/2104.05642v6.

[194] L.R. Dice, Measures of the Amount of Ecologic Association Between Species, Ecology. 26 (1945) 297–302.

[195] S. Silva, A. Teixeira, Quantitative systematic analysis of vocal tract data, Comput. Speech Lang. 36 (2016) 307–329. https://doi.org/10.1016/j.csl.2015.05.004.

[196] C. Carignan, P. Hoole, E. Kunay, M. Pouplier, A. Joseph, D. Voit, J. Frahm, J. Harrington, Analyzing speech in both time and space: Generalized additive mixed models can uncover systematic patterns of variation in vocal tract shape in real-time MRI, Lab. Phonol. J. Assoc. Lab. Phonol. 11 (2020). https://doi.org/10.5334/labphon.214.

[197] J. Kim, A. Toutios, S. Lee, S.S. Narayanan, Vocal tract shaping of emotional speech, Comput. Speech Lang. 64 (2020). https://doi.org/10.1016/j.csl.2020.101100.

[198] R. Seselgyte, M.C. Swan, M.J. Birch, L. Kangesu, Velopharyngeal Incompetence in Children With 22q11.2 Deletion Syndrome: Velar and Pharyngeal Dimensions, J.

Craniofac. Surg. 32 (2021) 578–580. https://doi.org/10.1097/SCS.0000000000007202.

[199] J. Kim, N. Kumar, S. Lee, S. Narayanan, Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data, in: Proc. 10th Int. Semin. Speech Prod., 2014: pp. 222–225.

[200] C. Valliappan, R. Mannem, P.K. Ghosh, Air-tissue boundary segmentation in real-time magnetic resonance imaging video using semantic segmentation with fully convolutional networks, in: INTERSPEECH, 2018: pp. 3132–3136. https://doi.org/10.21437/Interspeech.2018-1939.

[201] C. Valliappan, A. Kumar, R. Mannem, G. Karthik, P.K. Ghosh, An improved air tissue boundary segmentation technique for real time magnetic resonance imaging video using SegNet, in: IEEE Int. Conf. Acoust. Speech Signal Process., 2019: pp. 5921–5925.

[202] R. Mannem, P.K. Ghosh, Air-tissue boundary segmentation in real time magnetic resonance imaging video using a convolutional encoder-decoder network, in: IEEE Int. Conf. Acoust. Speech Signal Process., 2019: pp. 5941–5945.

[203] E. Bresch, S. Narayanan, Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images, IEEE Trans. Med. Imaging. 28 (2009) 323–338. https://doi.org/10.1109/TMI.2008.928920.

[204] S. Silva, A. Teixeira, Unsupervised segmentation of the vocal tract from real-time MRI sequences, Comput. Speech Lang. 33 (2015) 25–46. https://doi.org/10.1016/j.csl.2014.12.003.

[205] M. Ruthven, M.E. Miquel, A.P. King, Deep-learning-based segmentation of the vocal tract and articulators in real-time magnetic resonance images of speech, Comput. Methods Programs Biomed. 198 (2021) 105814. https://doi.org/10.1016/j.cmpb.2020.105814.

[206] A. Bonà, M. Cavicchioli, Vocal tract segmentation of dynamic speech MRI images based on deep learning for neurodegenerative disease application, Politecnico di Milano, 2021.

[207] S. Erattakulangara, K. Kelat, D. Meyer, S. Priya, S.G. Lingala, Automatic Multiple Articulator Segmentation in Dynamic Speech MRI Using a Protocol Adaptive Stacked Transfer Learning U-NET Model, Bioengineering. 20 (2023) 623. https://doi.org/https://doi.org/10.3390/bioengineering10050623.

[208] S. Erattakulangara, S.G. Lingala, Airway segmentation in speech MRI using the U-net

architecture, in: IEEE Int. Symp. Biomed. Imaging, 2020: pp. 1887–1890.

[209] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 2481–2495.

[210] J. Yang, B. Price, S. Cohen, H. Lee, M.H. Yang, Object contour detection with a fully convolutional encoder-decoder network, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2016: pp. 193–202. https://doi.org/10.1109/CVPR.2016.28.

[211] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 681–685. https://doi.org/10.1007/BFb0054760.

[212] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models - their training and application, Comput. Vis. Image Underst. 61 (1995) 38–59. https://doi.org/10.1006/cviu.1995.1004.

[213] M. De Bruijne, M. Nielsen, Shape particle filtering for image segmentation, Lect. Notes Comput. Sci. 3216 (2004) 168–175. https://doi.org/10.1007/978-3-540-30135-6_21.

[214] L. Di Stefano, A. Bulgarelli, A simple and efficient connected components labeling algorithm, in: Proc. - Int. Conf. Image Anal. Process. ICIAP 1999, 1999: pp. 322–327. https://doi.org/10.1109/ICIAP.1999.797615.

[215] B. Maas, E. Zabeh, S. Arabshahi, QuickTumorNet: Fast automatic multi-class segmentation of brain tumors, Int. IEEE/EMBS Conf. Neural Eng. NER. 2021-May (2021) 81–85. https://doi.org/10.1109/NER49283.2021.9441286.

[216] S. Firuzinia, S.M. Afzali, F. Ghasemian, S.A. Mirroshandel, A robust deep learning-based multiclass segmentation method for analyzing human metaphase II oocyte images, Comput. Methods Programs Biomed. 201 (2021) 105946. https://doi.org/10.1016/j.cmpb.2021.105946.

[217] M. Rossi, L. Marsilio, L. Mainardi, A. Manzotti, P. Cerveri, CEL-Unet: Distance Weighted Maps and Multi-Scale Pyramidal Edge Extraction for Accurate Osteoarthritic Bone Segmentation in CT Scans, Front. Signal Process. 2 (2022) 1–16. https://doi.org/10.3389/frsip.2022.857313.

[218] T. Ivanovska, A. Daboul, O. Kalentev, N. Hosten, R. Biffar, H. Völzke, F. Wörgötter, A deep cascaded segmentation of obstructive sleep apnea-relevant organs from sagittal spine MRI, Int. J. Comput. Assist. Radiol. Surg. 16 (2021) 579–588.

https://doi.org/10.1007/s11548-021-02333-0.

[219] H.H. Lee, S. Bao, Y. Huo, B.A. Landman, 3D UX-Net: A Large Kernel Volumetric ConvNet Modernizing Hierarchical Transformer for Medical Image Segmentation, in: Int. Conf. Learn. Represent., 2023.

[220] D. Aalto, O. Aaltonen, R.P. Happonen, P. Jääsaari, A. Kivelä, J. Kuortti, J.M. Luukinen, J. Malinen, T. Murtola, R. Parkkola, J. Saunavaara, T. Soukka, M. Vainio, Large scale data acquisition of simultaneous MRI and speech, Appl. Acoust. 83 (2014) 64–75. https://doi.org/10.1016/j.apacoust.2014.03.003.

[221] P. Birkholz, S. Kürbis, S. Stone, P. Häsner, R. Blandin, M. Fleischer, Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties, Sci. Data. 7 (2020) 1–16. https://doi.org/10.1038/s41597-020-00597-w.

[222] K.K. Brock, S. Mutic, T.R. McNutt, H. Li, M.L. Kessler, Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132: Report, Med. Phys. 44 (2017) e43–e76. https://doi.org/10.1002/mp.12256.

[223] B.B. Avants, C.L. Epstein, M. Grossman, J.C. Gee, Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain, Med. Image Anal. 12 (2008) 26–41. https://doi.org/10.1016/j.media.2007.06.004.

[224] J.B.A. Maintz, M.A. Viergever, A survey of medical image registration, Med. Image Anal. 2 (1998) 1–36. https://doi.org/10.1016/S1361-8415(01)80026-8.

[225] A. Sotiras, C. Davatzikos, N. Paragios, Deformable medical image registration: A survey, IEEE Trans. Med. Imaging. 32 (2013) 1153–1190. https://doi.org/10.1109/TMI.2013.2265603.

[226] M.A. Viergever, J.B.A. Maintz, S. Klein, K. Murphy, M. Staring, J.P.W. Pluim, A survey of medical image registration – under review, Med. Image Anal. 33 (2016) 140–144. https://doi.org/10.1016/j.media.2016.06.030.

[227] E. Ferrante, N. Paragios, Slice-to-volume medical image registration: A survey, Med. Image Anal. 39 (2017) 101–123. https://doi.org/10.1016/j.media.2017.04.010.

[228] G. Haskins, U. Kruger, P. Yan, Deep learning in medical image registration: a survey, Mach. Vis. Appl. 31 (2020) 1–18. https://doi.org/10.1007/s00138-020-01060-x.

[229] Y. Fu, Y. Lei, T. Wang, W.J. Curran, T. Liu, X. Yang, Deep learning in medical image

registration: A review, Phys. Med. Biol. 65 (2020). https://doi.org/10.1088/1361-6560/ab843e.

[230] S. Uchida, Image processing and recognition for biological images, Dev. Growth Differ. 55 (2013) 523–549. https://doi.org/10.1111/dgd.12054.

[231] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C.M. Moore, M. Emberton, S. Ourselin, J.A. Noble, D.C. Barratt, T. Vercauteren, Weakly-supervised convolutional neural networks for multimodal image registration, Med. Image Anal. 49 (2018) 1–13. https://doi.org/10.1016/j.media.2018.07.002.

[232] M.A. Schmidt, G.S. Payne, Radiotherapy planning using MRI, Phys. Med. Biol. 60 (2015) R323–R361. https://doi.org/10.1088/0031-9155/60/22/R323.

[233] W. Bai, W. Shi, A. de Marvao, T.J.W. Dawes, D.P. O'Regan, S.A. Cook, D. Rueckert, A bi-ventricular cardiac atlas built from 1000+ high resolution MR images of healthy subjects and an analysis of shape and motion, Med. Image Anal. 26 (2015) 133–145. https://doi.org/10.1016/j.media.2015.08.009.

[234] J. Ehrhardt, R. Werner, A. Schmidt-Richberg, H. Handels, Statistical modeling of 4D respiratory lung motion using diffeomorphic image registration, IEEE Trans. Med. Imaging. 30 (2011) 251–265. https://doi.org/10.1109/TMI.2010.2076299.

[235] A. Schmidt-Richberg, R. Werner, H. Handels, J. Ehrhardt, Estimation of slipping organ motion by registration with direction-dependent regularization, Med. Image Anal. 16 (2012) 150–159. https://doi.org/10.1016/j.media.2011.06.007.

[236] D. Rueckert, L.I. Sonoda, C. Hayes, D.L.G. Hill, M.O. Leach, D.J. Hawkes, Nonrigid registration using free-form deformations: Application to breast mr images, IEEE Trans. Med. Imaging. 18 (1999) 712–721. https://doi.org/10.1109/42.796284.

[237] M. Modat, G.R. Ridgway, Z.A. Taylor, M. Lehmann, J. Barnes, D.J. Hawkes, N.C. Fox, S. Ourselin, Fast free-form deformation using graphics processing units, Comput. Methods Programs Biomed. 98 (2010) 278–284. https://doi.org/10.1016/j.cmpb.2009.09.002.

[238] S. Ourselin, A. Roche, G. Subsol, X. Pennec, N. Ayache, Reconstructing a 3D structure from serial histological sections, Image Vis. Comput. 19 (2001) 25–31. https://doi.org/10.1016/S0262-8856(00)00052-4.

[239] J.P. Thirion, Image matching as a diffusion process: An analogy with Maxwell's demons, Med. Image Anal. 2 (1998) 243–260. https://doi.org/10.1016/S1361-

8415(98)80022-4.

[240] T. Vercauteren, X. Pennec, A. Perchant, N. Ayache, Diffeomorphic demons: efficient non-parametric image registration., Neuroimage. 45 (2009) S61–S72. https://doi.org/10.1016/j.neuroimage.2008.10.040.

[241] M. Modat, D.M. Cash, P. Daga, G.P. Winston, J.S. Duncan, S. Ourselin, Global image registration using a symmetric block-matching approach, J. Med. Imaging. 1 (2014) 024003. https://doi.org/10.1117/1.jmi.1.2.024003.

[242] M. Mccormick, X. Liu, J. Jomier, C. Marion, L. Ibanez, ITK: enabling reproducible research and open science, Front. Neuroinform. 8 (2014) 1–11. https://doi.org/10.3389/fninf.2014.00013.

[243] B.D. de Vos, F.F. Berendsen, M.A. Viergever, H. Sokooti, M. Staring, I. Išgum, A deep learning framework for unsupervised affine and deformable image registration, Med. Image Anal. 52 (2019) 128–143. https://doi.org/10.1016/j.media.2018.11.010.

[244] J. Krebs, H. Delingette, B. Mailhe, N. Ayache, T. Mansi, Learning a Probabilistic Model for Diffeomorphic Registration, IEEE Trans. Med. Imaging. 38 (2019) 2165–2176. https://doi.org/10.1109/TMI.2019.2897112.

[245] G. Balakrishnan, A. Zhao, M.R. Sabuncu, J. Guttag, A. V. Dalca, VoxelMorph: A Learning Framework for Deformable Medical Image Registration, IEEE Trans. Med. Imaging. 38 (2019) 1788–1800. https://doi.org/10.1109/TMI.2019.2897538.

[246] A. V. Dalca, G. Balakrishnan, J. Guttag, M.R. Sabuncu, Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces, Med. Image Anal. 57 (2019) 226–236. https://doi.org/10.1016/j.media.2019.07.006.

[247] C. Qin, W. Bai, J. Schlemper, S.E. Petersen, S.K. Piechnik, S. Neubauer, D. Rueckert, Joint learning of motion estimation and segmentation for cardiac MR image sequences, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 11071 LNCS (2018) 472–480. https://doi.org/10.1007/978-3-030-00934-2_53.

[248] T. Estienne, M. Lerousseau, M. Vakalopoulou, E. Alvarez Andres, E. Battistella, A. Carré, S. Chandra, S. Christodoulidis, M. Sahasrabudhe, R. Sun, C. Robert, H. Talbot, N. Paragios, E. Deutsch, Deep Learning-Based Concurrent Brain Registration and Tumor Segmentation, Front. Comput. Neurosci. 14 (2020). https://doi.org/10.3389/fncom.2020.00017.

[249] B. Li, W.J. Niessen, S. Klein, M. de Groot, M.A. Ikram, M.W. Vernooij, E.E. Bron, Longitudinal diffusion MRI analysis using Segis-Net: A single-step deep-learning framework for simultaneous segmentation and registration, Neuroimage. 235 (2021) 118004. https://doi.org/10.1016/j.neuroimage.2021.118004.

[250] Z. Xu, M. Niethammer, DeepAtlas: Joint Semi-supervised Learning of Image Registration and Segmentation, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 11765 LNCS (2019) 420–429. https://doi.org/10.1007/978-3-030-32245-8_47.

[251] Y. He, T. Li, R. Ge, J. Yang, Y. Kong, J. Zhu, H. Shu, G. Yang, S. Li, Few-Shot Learning for Deformable Medical Image Registration with Perception-Correspondence Decoupling and Reverse Teaching, IEEE J. Biomed. Heal. Informatics. 26 (2022) 1177–1187. https://doi.org/10.1109/JBHI.2021.3095409.

[252] L. Qiu, H. Ren, RSegNet: A Joint Learning Framework for Deformable Registration and Segmentation, IEEE Trans. Autom. Sci. Eng. 19 (2021) 2499–2513. https://doi.org/10.1109/TASE.2021.3087868.

[253] X. Chen, Y. Xia, N. Ravikumar, A.F. Frangi, A Deep Discontinuity-Preserving Image Registration Network, Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-87202-1.

[254] Y. Chen, L. Xing, L. Yu, W. Liu, B. Pooya Fahimian, T. Niedermayr, H.P. Bagshaw, M. Buyyounouski, B. Han, MR to ultrasound image registration with segmentation-based learning for HDR prostate brachytherapy, Med. Phys. 48 (2021) 3074–3083. https://doi.org/10.1002/mp.14901.

[255] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial Transformer Networks, in: Adv. Neural Inf. Process. Syst., 2015: pp. 2017–2025. https://doi.org/10.1145/2948076.2948084.

[256] M.C.H. Lee, O. Oktay, A. Schuh, M. Schaap, B. Glocker, Image-and-Spatial Transformer Networks for Structure-Guided Image Registration, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2019: pp. 337–345. https://doi.org/10.1007/978-3-030-32245-8_38.

[257] X. Chen, Y. Xia, N. Ravikumar, A.F. Frangi, Joint segmentation and discontinuity-preserving deformable registration: Application to cardiac cine-MR images, (2022) 1–27. http://arxiv.org/abs/2211.13828.

[258] F. Zhao, Z. Wu, L. Wang, W. Lin, S. Xia, G. Li, A Deep Network for Joint Registration and Parcellation of Cortical Surfaces, Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-87202-1_17.

[259] Z. Wu, E. Rietzel, V. Boldea, D. Sarrut, G.C. Sharp, Evaluation of deformable registration of patient lung 4DCT with subanatomical region segmentations, Med. Phys. 35 (2008) 775–781. https://doi.org/10.1118/1.2828378.

[260] D.F. Pace, S.R. Aylward, M. Niethammer, A locally adaptive regularization based on anisotropic diffusion for deformable image registration of sliding organs, IEEE Trans. Med. Imaging. 32 (2013) 2114–2126. https://doi.org/10.1109/TMI.2013.2274777.

[261] R. Hua, J.M. Pozo, Z.A. Taylor, A.F. Frangi, Multiresolution eXtended Free-Form Deformations (XFFD) for non-rigid registration with discontinuous transforms, Med. Image Anal. 36 (2017) 113–122. https://doi.org/10.1016/j.media.2016.10.008.

[262] D. Li, W. Zhong, K.M. Deh, T.D. Nguyen, M.R. Prince, Y. Wang, P. Spincemaille, Discontinuity Preserving Liver MR Registration with Three-Dimensional Active Contour Motion Segmentation, IEEE Trans. Biomed. Eng. 66 (2019) 1884–1897. https://doi.org/10.1109/TBME.2018.2880733.

[263] F.F. Berendsen, A.N.T.J. Kotte, M.A. Viergever, J.P.W. Pluim, Registration of organs with sliding interfaces and changing topologies, Med. Imaging 2014 Image Process. 9034 (2014) 1–7. https://doi.org/10.1117/12.2043447.

[264] V. Delmon, S. Rit, R. Pinho, D. Sarrut, Registration of sliding objects using direction dependent B-splines decomposition., Phys. Med. Biol. 58 (2013) 1303–1314. https://doi.org/10.1088/0031-9155/58/5/1303.

[265] Y. Fu, S. Liu, H.H. Li, H. Li, D. Yang, An adaptive motion regularization technique to support sliding motion in deformable image registration:, Med. Phys. 45 (2018) 735–747. https://doi.org/10.1002/mp.12734.

[266] T.-P. Fries, T. Belytschko, The extended/generalized finite element method: An overview of the method and its applications, Int. J. Numer. Methods Eng. 84 (2010) 253–304. https://doi.org/10.1002/nme.

[267] E. Ng, M. Ebrahimi, An unsupervised learning approach to discontinuity-preserving image registration, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Springer International Publishing, 2020: pp. 153–162. https://doi.org/10.1007/978-3-030-50120-4_15.

[268] I.K. Douros, A. Tsukanova, K. Isaieva, P.A. Vuissoz, Y. Laprie, Towards a method of dynamic vocal tract shapes generation by combining static 3D and dynamic 2D MRI speech data, Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH. (2019) 879–883. https://doi.org/10.21437/Interspeech.2019-2880.

[269] I.K. Douros, A. Kulkarni, C. Dourou, Y. Xie, J. Felblinger, K. Isaieva, P.A. Vuissoz, Y. Laprie, Using silence MR image to synthesise dynamic MRI vocal tract data of CV, Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH. (2020) 3730–3734. https://doi.org/10.21437/Interspeech.2020-1173.

[270] I.K. Douros, A. Kulkarni, Y. Xie, C. Dourou, J. Felblinger, K. Isaieva, P.A. Vuissoz, Y. Laprie, MRI vocal tract sagittal slices estimation during speech production of CV, in: Eur. Signal Process. Conf., 2020: pp. 1115–1119. https://doi.org/10.23919/Eusipco47968.2020.9287834.

[271] F. Odille, J.A. Steeden, V. Muthurangu, D. Atkinson, Automatic segmentation propagation of the aorta in real-time phase contrast MRI using nonrigid registration, J. Magn. Reson. Imaging. 33 (2011) 232–238. https://doi.org/10.1002/jmri.22402.

[272] M. Ruthven, Real-time speech MRI: what is the minimum temporal resolution for clinical velopharyngeal closure assessment?, King's College London, 2016.

[273] D.W. McRobbie, E.A. Moore, M.J. Graves, M.R. Prince, Improving Your Image: How to Avoid Artefacts, in: MRI from Pict. to Prot., 3rd ed., Cambridge University Press, Cambridge, 2017: pp. 81–101. https://doi.org/10.1017/9781107706958.008.

[274] D.W. McRobbie, E.A. Moore, M.J. Graves, M.R. Prince, What You Set is What You Get: Basic Image Optimization, in: MRI from Pict. to Prot., 3rd ed., Cambridge University Press, Cambridge, 2017: pp. 67–80. https://doi.org/10.1017/9781107706958.007.

[275] P. Horos, Horos, (2023). www.horosproject.org.

[276] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J. V. Miller, S. Pieper, R. Kikinis, 3D Slicer as an image computing platform for the Quantitative Imaging Network, Magn. Reson. Imaging. 30 (2012) 1323–1341. https://doi.org/10.1016/j.mri.2012.05.001.

[277] M. Leppävuori, E. Lammentausta, A. Peuna, M.K. Bode, J. Jokelainen, J. Ojala, M.T. Nieminen, Characterizing Vocal Tract Dimensions in the Vocal Modes Using Magnetic Resonance Imaging, J. Voice. 35 (2021) 804.e27-804.e42.

https://doi.org/10.1016/j.jvoice.2020.01.015.

[278] T. Ikävalko, A.M. Laukkanen, A. McAllister, R. Eklund, E. Lammentausta, M. Leppävuori, M.T. Nieminen, Three Professional Singers' Vocal Tract Dimensions in Operatic Singing, Kulning, and Edge—A Multiple Case Study Examining Loud Singing, J. Voice. (2022). https://doi.org/10.1016/j.jvoice.2022.01.024.

[279] M. Belyk, C. McGettigan, Real-time magnetic resonance imaging reveals distinct vocal tract configurations during spontaneous and volitional laughter, Philos. Trans. R. Soc. B Biol. Sci. 377 (2022). https://doi.org/10.1098/rstb.2021.0511.

[280] K. Bettens, F.L. Wuyts, K.M. Van Lierde, Instrumental assessment of velopharyngeal function and resonance: A review, J. Commun. Disord. 52 (2014) 170–183. https://doi.org/10.1016/j.jcomdis.2014.05.004.

[281] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. D'Alché-Buc, E. Fox, R. Garnett (Eds.), Adv. Neural Inf. Process. Syst. 32, Curran Associates, Inc., 2019: pp. 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[282] J.M. Johnson, T.M. Khoshgoftaar, Survey on deep learning with class imbalance, J. Big Data. 6 (2019). https://doi.org/10.1186/s40537-019-0192-5.

[283] Z. Li, K. Kamnitsas, B. Glocker, Analyzing Overfitting under Class Imbalance in Neural Networks for Image Segmentation, IEEE Trans. Med. Imaging. 40 (2021) 1065–1077. https://doi.org/10.1109/TMI.2020.3046692.

[284] C.H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M. Jorge Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Lect. Notes Comput. Sci., 2017: pp. 240–248. https://doi.org/10.1007/978-3-319-67558-9_28.

[285] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, A.L. Martel, Loss odyssey in medical image segmentation, Med. Image Anal. 71 (2021). https://doi.org/10.1016/j.media.2021.102035.

[286] D. Krstajic, L.J. Buturovic, D.E. Leahy, S. Thomas, Cross-validation pitfalls when

selecting and assessing regression and classification models, J. Cheminform. 6 (2014). https://doi.org/10.1186/1758-2946-6-10.

[287] W. Tian, R.J. Redett, New velopharyngeal measurements at rest and during speech: Implications and applications, J. Craniofac. Surg. 20 (2009) 532–539. https://doi.org/10.1097/SCS.0b013e31819b9fbe.

[288] W. Tian, Y. Li, H. Yin, S.F. Zhao, S. Li, Y. Wang, B. Shi, Magnetic resonance imaging assessment of velopharyngeal motion in Chinese children after primary palatal repair, J. Craniofac. Surg. 21 (2010) 578–587. https://doi.org/10.1097/SCS.0b013e3181d08bee.

[289] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, N. Paragios, Dense image registration through MRFs and efficient linear programming, Med. Image Anal. 12 (2008) 731–741. https://doi.org/10.1016/j.media.2008.03.006.

[290] J. Chen, E.C. Frey, Y. He, W.P. Segars, Y. Li, Y. Du, TransMorph: Transformer for unsupervised medical image registration, Med. Image Anal. 82 (2022). https://doi.org/10.1016/j.media.2022.102615.

[291] M. Ruthven, M.E. Miquel, A.P. King, A segmentation-informed deep learning framework to register dynamic two-dimensional magnetic resonance images of the vocal tract during speech, Biomed. Signal Process. Control. 80 (2023) 104290. https://doi.org/10.1016/j.bspc.2022.104290.

[292] MONAI-Consortium, MONAI: Medical Open Network for AI, (2022). https://doi.org/10.5281/zenodo.6639453.

[293] M.B.M. Ranzini, J. Henckel, M. Ebner, M.J. Cardoso, A. Isaac, T. Vercauteren, S. Ourselin, A. Hart, M. Modat, Automated postoperative muscle assessment of hip arthroplasty patients using multimodal imaging joint segmentation, Comput. Methods Programs Biomed. 183 (2020). https://doi.org/10.1016/j.cmpb.2019.105062.

[294] J. Clough, N. Byrne, I. Oksuz, V.A. Zimmer, J.A. Schnabel, A. King, A Topological Loss Function for Deep-Learning based Image Segmentation using Persistent Homology, IEEE Trans. Pattern Anal. Mach. Intell. (2020) 1–14. https://doi.org/10.1109/TPAMI.2020.3013679.

[295] Y. Fu, Y. Lei, T. Wang, W.J. Curran, T. Liu, X. Yang, A review of deep learning based methods for medical image multi-organ segmentation, Phys. Medica. 85 (2021) 107–

122. https://doi.org/10.1016/j.ejmp.2021.05.003.

[296] M.H. Hesamian, W. Jia, X. He, P. Kennedy, Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges, J. Digit. Imaging. 32 (2019) 582–596. https://doi.org/10.1007/s10278-019-00227-x.

[297] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, A. V. Dalca, Data augmentation using learned transformations for one-shot medical image segmentation, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2019-June (2019) 8535–8545. https://doi.org/10.1109/CVPR.2019.00874.

[298] Z. Shen, Z. Xu, S. Olut, M. Niethammer, Anatomical Data Augmentation via Fluid-Based Image Registration, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2020: pp. 318–328. https://doi.org/10.1007/978-3-030-59716-0.

[299] J. Corral Acero, E. Zacur, H. Xu, R. Ariga, A. Bueno-Orovio, P. Lamata, V. Grau, SMOD - Data Augmentation Based on Statistical Models of Deformation to Enhance Segmentation in 2D Cine Cardiac MRI, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2019: pp. 361–369. https://doi.org/10.1007/978-3-030-21949-9_39.

[300] J. Corral Acero, H. Xu, E. Zacur, J.E. Schneider, P. Lamata, A. Bueno-Orovio, V. Grau, Left Ventricle Quantification with Cardiac MRI: Deep Learning Meets Statistical Models of Deformation, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 12009 LNCS (2020) 384–394. https://doi.org/10.1007/978-3-030-39074-7_40.

[301] J. Nalepa, G. Mrukwa, S. Piechaczek, P.R. Lorenzo, M. Marcinkiewicz, B. Bobek-billewicz, P. Wawrzyniak, P. Ulrych, J. Szymanek, M. Cwiek, W. Dudzik, M. Kawulok, M.P. Hayball, Data Augmentation Via Image Registration, 2019 IEEE Int. Conf. Image Process. (2019) 4250–4254.

[302] G.P. Penney, J.A. Schnabel, D. Rueckert, M.A. Viergever, W.J. Niessen, Registration-Based Interpolation, IEEE Trans. Med. Imaging. 23 (2004) 922–926. https://doi.org/10.1109/TMI.2004.828352.

[303] D. Rueckert, A.F. Frangi, J.A. Schnabel, Automatic Construction of 3-D Statistical Deformation Models of the Brain Using Nonrigid Registration, IEEE Trans. Med. Imaging. 22 (2003) 1014–1025. https://doi.org/10.1109/TMI.2003.815865.