Automatic Analysis of Magnetic Resonance Images of Speech Articulation

A Thesis submitted for the degree of

Doctor of Philosophy

Zeynab Raeesy

The Queen's College

University of Oxford

Michaelmas term 2013

To Milad,

to my family,

with love.

Abstract

Magnetic resonance imaging (MRI) technology has facilitated capturing the dynamics of speech production at fine temporal and spatial resolutions, thus generating substantial quantities of images to be analysed. Manual processing of large MRI databases is labour intensive and time consuming. Hence, to study articulation on large scale, techniques for automatic feature extraction are needed.

This thesis investigates approaches for automatic information extraction from an MRI database of dynamic articulation. We first study the articulation by observing the pixel intensity variations in image sequences. The correspondence between acoustic segments and images is established by forced alignment of speech signals recorded during the articulation. We obtain speaker-specific typical phoneme articulations that represent general articulatory configurations in running speech. Articulation dynamics are parametrised by measuring the magnitude of change in intensities over time. We demonstrate a direct correlation between the dynamics of articulation thus measured and the energy of the generated acoustic signals.

For more sophisticated applications, a parametric description of vocal tract shape is desired. We investigate different shape extraction techniques and present a framework that can automatically identify and extract the vocal tract shapes. The framework incorporates shape prior information and intensity features in recognising and delineating the shape. The new framework is a promising new tool for automatic identification of vocal tract boundaries in large MRI databases, as demonstrated through extensive assessments.

The segmentation framework proposed in this thesis is, to the best of our knowledge, novel in the field of speech production. The methods investigated in this thesis facilitate automatic information extraction from images, either for studying the dynamics of articulation or for vocal tract shape modelling. This thesis advances the state-of-the-art by bringing new perspectives to studying articulation, and introducing a segmentation framework that is automatic, does not require extensive initialisation, and reports a minimum number of failures.

Acknowledgements

This PhD has been a very rewarding and fulfilling experience for me. I would like to express my deepest appreciations to all those providing me with the possibility of this achievement; those who contributed by their valuable suggestions and encouragements, and those whose love and kindness made this happen.

I would like to first thank my supervisor Professor John Coleman, for his constant guidance, support, and useful remarks during my doctoral studies. He persuasively conveyed a spirit of adventure in research and scholarship and guided my research career with professionalism. I am specially thankful to him for introducing me to phonetics science, and grateful to him for his patience, insightful comments and advices on writing. This dissertation would have not been possible without his help and guidance.

I am more than thankful to Dr. Sylvia Rueda, who guided me through the image processing part of this research. I am grateful to her for devoting time and patience to engage with my research problems, and sharing her experience and ideas with me. She was also a good friend and I will not forget our long fruitful meetings at Old Road Campus in Oxford. I am grateful to her for allowing me to use some of her original codes for fuzzy connectedness, RBS, and evaluation in my research. I am thankful to Sergio Grau for making this collaboration happen by introducing me and my research to Sylvia.

I would like to express my appreciation to Professor Jayram K. Udupa from University of Pennsylvania for his guidance and enthusiasm. I am grateful to him for devoting his time, particularly in Sylvia's absence, to help me with technical or research issues that I encountered. Special thanks to him for providing me with the 3Dviewnix software and the original code for Oriented Active Shape models. I would like to thank Dewey Odhner for remotely assisting me to set up 3Dviewnix.

I would like to thank Dr. Ladan Baghai-Ravary for all her encouragement and technical advice. Collaborating with Ladan was a wonderful experience, full of excitement. Some of the materials in chapter 4, *parametrising degree of the articulatory*

movements, are the result of my collaboration with her, that was also published in proceedings of Interspeech 2011. Ladan has also been a good friend and colleague, and I always enjoyed our conversations in Persian.

I am grateful to Dr. Elinor Payne for being a great colleague and a skeptical assessor for my transfer of status and confirmation interviews. Attending the ICPhS congress in Hong Kong would have not been such an unforgettable memory if she was not there. Her stimulating company is among my best experiences at Oxford and Cambridge.

I would also like to thank Dr. Greg Kochanski and Dr. Anastassia Loukina for their contributions and help during my time at the Phonetics Laboratory. My computational tasks would have been impossible without the help of our IT department staff; special thanks to Jon Edwards and Amir Nettler for being very patient with me and my IT demands. Thanks to Sally James, phonetics lab secretary for her support.

I am grateful to my college, The Queen's College, for the support and for providing me with a safe and convenient environment to pursue my studies. Thanks to the amazing sponsors and funding bodies of the Clarendon scholarship for providing me with this unique opportunity to study at Oxford. Special thanks to my college advisor Dr. Charlie Louth for his support. My scholarship was partly funded by the governing body at The Queen's college, to whom I am thankful.

Thanks to many friends whose companionship made my PhD experience a more rewarding journey. Special thanks to my wonderful friend Monica, for her endless kindness and support, and for being a wonderful friend. Many thanks to my great friends Naimz, Tali and Ahora for the many amazing moments we spent together during my PhD years. I am thankful to my Cambridge friends for the heart-warming evenings and weekends that I shared with them. Special thanks to Sara for her kindness, and to Anoosheh and Kaveh for being great friends. I would like to thank Anita and Vinay for being there for me all along this journey, from the beginning to the end.

None of these would have been possible without the unconditional love and support of my family. Special thanks to my mother Zahra and my father Rahmat, for their constant encouragements and supports, and for their patience in not asking me “when are you going to finish?”. My entire academic journey, from pre-school to Oxford, was not possible without their sacrifices, kindness, encouragement and endless love. Special thanks to my brother Taha, for his love and support and for being so positive, also for his music selections that were my constant companions over the past few years. Thanks to my wonderfully kind and funny sister, Mehda, for her love and for the not-so-short happy and lively conversations during the dark winter evenings.

I do not know where to begin in thanking my husband Milad, the one and only. Words cannot express my appreciation; I would have not been where I am and would have not achieved this without his unconditional love. I am thankful to him for being my “patience stone” when I needed one, and for brightening my days with his exceptional sense of humour. I cannot thank him enough for his support and for believing in me. Thank you Milad for your endless love that made this happen.

Disclaimer

I Zeynab Raeesy of The Queen's College, being a DPhil student at the Phonetics Laboratory, hereby confirm that this dissertation and the work explained in it are my own work. Any work by others mentioned or utilised in this dissertation is explicitly clarified and referenced.

Zeynab Raeesy, December 2013

Publications

Portions of the research in this thesis have appeared in the following publications. I confirm that I am the main author of these publications.

- Z. Raeesy. Speaker-specific typical vocal tract shapes obtained using dynamic MRI. *In Proceedings of 17th International Congress of Phonetics Sciences (ICPhS)*, pages 1658–1661, Hong Kong, China, 2011.

- Z. Raeesy, L. Baghai-Ravary, and J. Coleman. Parametrising degree of articulator movement from dynamic MRI data. *In Proceedings of the 12th Interspeech Conference*, pages 2853–2856, Florence, Italy, 2011.

- Z. Raeesy, S. Rueda, J. K. Udupa, and J. Coleman. Automatic segmentation of vocal tract MR images. *In IEEE 10th International Symposium on Biomedical Imaging (ISBI)*, pages 1328–1331, San Francisco, CA, 2013.

- Z. Raeesy, S. Rueda, J. K. Udupa, and J. Coleman. A new technique for automatic shape extraction from vocal tract MRI. *In Proceedings of 10th International Seminar on Speech Production*, to appear, Cologne, Germany, 2014.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and rationale

Speech is the predominant method of communication between humans. Speech sound waves are produced through the mechanism of *articulation*, which involves coordination of different organs such as lungs, glottis, larynx, and articulators such as tongue, velum, and jaw. The air pushed from the lungs through the larynx to the mouth is *filtered* in the vocal tract to generate distinct sounds. The *shape* of the vocal tract, its deformations and the articulator movements that cause the deformations are thus very important in producing speech. Understanding how the articulatory system acts when producing speech is a key part of understanding how speech is transmitted between humans. Modelling and recognition of speech can be enhanced by considering information about the underlying speech production system.

Except for the lips, the rest of the articulators and their dynamics cannot be observed externally as speech is an internal, hidden process. Therefore, observing the shape of the vocal tract during speech production is not trivial. Among various methods for observing and measuring such internal movements (reviewed in section 2.2), magnetic resonance imaging (MRI) of static and dynamic articulation have been utilised to capture the underlying articulation.

Magnetic resonance imaging of human speech has improved substantially over

past decades. The long acquisition time required for capturing each image, ranging from seconds to several minutes (e.g. 3.2 seconds for each image in [Narayanan *et al.* 1995] and up to 3.4 minutes for each image in [Baer *et al.* 1987]), limited the preliminary studies to capturing *static* articulation, where the subject sustained the articulation of a sound during acquisition. However, a few protocols have been developed such as collecting data over several repetitions [Foldvik *et al.* 1995] or using zooming techniques [Demolin *et al.* 2000; 2002] to facilitate capturing *dynamic* speech at very fine temporal and spatial resolutions. For example, utilising these protocols allows capturing articulators' movements to reconstruct MRI movies of running speech from pseudo-images with a temporal resolution of 50 ms [Alvey *et al.* 2008], and capturing realtime speech at a rate of 166–250 ms/image [Demolin *et al.* 2000]. As a result the quantity of data to be processed has increased significantly, and the demand for automatic computerised methods for extracting information from images has grown rapidly. Manual and human supervised extraction of information from large databases is not only labour intensive and time consuming, but also can be subjective and prone to error.

Automatic identification and extraction of shape information from high quantities of continuous speech MRI data is not trivial, due to the complexity of the deformation space of the vocal tract and the high dimensionality of the data. For example, even in a short image sequence of a few seconds, the result of MRI acquisition can be 100 images in a sequence, where each image is a minimum of 128×128 pixels, hence several million data points to be analysed. Automatic feature extraction from MR images of speech articulation from large databases that capture running speech involves several challenges. These include (1) variance caused by phonemic variations, (2) variance caused by speakers' anatomical differences and habits of articulation, (3) extreme variation in shape, position and structure of articulators due to their specific articulations, (4) coarticulatory effects resulting from the context, and (5) blur and noise introduced into the images as a result of the fast motions of articulators. Any image processing method to be applied for analysis of

such data must therefore address these challenges.

Depending on the application and purpose of study, the dimensionality of vocal tract MRIs of running speech can be reduced via different approaches. An anatomically well-founded approach is to find the air-tissue boundaries and to trace outlines of the vocal tract and articulators. Another approach, that is more specific to MRIs of running rather than static speech, is to use image features in the spatiotemporal domain directly.

This thesis focuses on devising automatic methods for obtaining features from images to use them in the study of articulatory process. Most of the techniques used for investigating articulation using MR images of running speech in previous studies are either manual or semi-automatic. This thesis looks at articulation in a new way by looking at the images in the time domain and estimating the movements and amount of movements based on spatial image features. A new framework for automatic segmentation of the entire shape is introduced that is novel in the field of vocal tract shape extraction from images.

1.2 Methodology and objectives

Initially, we aim at utilising information that is not dependent on parametric shape or articulator, such as spatial and temporal information and the sequential order of images, to obtain relevant information regarding the articulation process. The grey-levels of image elements, *pixels* in this thesis, are used to observe the dynamics of articulators and the degree of movement in the articulation of different sounds. A novel approach for looking at articulatory configurations in running speech is proposed that provides typical or characteristic settings for the articulation of each phoneme. The configuration obtained for each phoneme is context-independent while reflecting the variance in the typical articulatory posture of the phoneme caused by coarticulation. The grey-level variance of pixels in the time domain is also used to parameterise the articulators' dynamics when generating different phonemes.

We demonstrate that investigating the motions and deformations of the articu-

lators using image-based information, i.e. without considering parametric features, can be used successfully in studying the dynamics of speech production. However, in applications where articulation is to be modelled or to be further used for synthesis purposes, a parametric description is required that can be quantified, measured and modelled. We thus explore the automatic feature extraction techniques for obtaining parametric description such as the vocal tract shape. Approaches for using pixel grey-levels for recognising the boundaries and retrieving the shape of the vocal tract are first examined. Preliminary results suggested that due to the nature of the images, image-based feature extraction methods without considering the shape priors are only applicable with manual supervision of experts. We build on the above observations to devise a new framework based on vocal tract shape models and pixel intensity properties to extract the shape of the vocal tract from images automatically. The performance of the proposed framework is extensively analysed from image processing and articulatory modelling perspectives to assess the applicability of the method in acoustic phonetics research applications.

The main objectives of this thesis are:

- to study the dynamics of articulation and general vocal tract shape using MR images of articulation;
- to devise methods to extract articulatory features from images with minimal or zero human supervision;
- to utilise spatial and temporal information of the image sequences to investigate articulatory movements;
- to devise automatic methods for segmenting images for parametric modelling of the vocal tract; and
- to assess the applicability and effectiveness of the proposed methods.

1.3 Thesis structure

The remainder of this thesis can be divided into four parts. The first part reviews the methods for observing and modelling human articulation (chapters 2 and 3). The second part describes a novel image-based analysis for studying articulation in sequences of images and investigates the correlations with generated acoustics (chapter 4). In the third part, a new approach for automatic segmentation of vocal tract images is presented (chapters 5 and 6). Finally, the last part includes an extensive evaluation of the proposed segmentation approach (chapter 7). A summary of the contents of the chapters is presented below.

Chapter 2 gives a general background for the thesis by reviewing the articulatory process and main methods devised for observing articulation. A review of techniques for modelling articulation is also presented in this chapter.

Chapter 3 explains the application of magnetic resonance imaging in capturing articulation and reviews earlier methods for processing and extracting features from MR images of the vocal tract. The description of the MRI database used in this thesis, and the preparation and image enhancement methods applied are included in this chapter. An introduction to pixel variation and hidden information is also presented.

Chapter 4 describes pixel-based approaches for studying articulation. An HMM forced-alignment method used for estimating phoneme duration and image sequence boundary identification is explained. A novel method is introduced for obtaining the typical articulatory configuration of different phonemes for each speaker. The degree of articulator movements is quantified and parameterised and its correlation with the generated acoustics in terms of dynamics of acoustics is investigated.

Chapter 5 starts by explaining the objectives for utilising a parametric shape descriptor for studying articulation. A review of feature extraction techniques applied

directly or indirectly in this chapter and chapter 6 is presented. A region-based approach for automatic segmentation of the vocal tract shape in MR images is applied and the results are analysed. The need for an approach that utilises the vocal tract shape model is discussed based on the outcomes of the region-based approach.

Chapter 6 describes a new framework for automatic boundary delineation and shape extraction, inspired by the results discussed in chapter 5. The application of the framework to the MRI database is explained and the results are analysed. Further adaptation of the framework for application to the MRI database is explained.

Chapter 7 presents a comprehensive evaluation of the proposed framework. The evaluations are divided into two parts: the first part investigates the performance in terms of image processing metrics, while the second part assesses the performance in articulation modelling and acoustic generation applications.

Chapter 8 highlights the conclusions and outcomes of this thesis. The future direction of this research and some open questions are also discussed in this chapter.

Chapter 2

Background

2.1 Mechanism of speech production

In human speech, messages in the brain are converted through speech movements into acoustic energy, which is transmitted to the outside air from the human vocal tract. On the listener's side, this acoustic energy, propagating through the air, is received by ears, and interpreted by the brain to comprehend the original message. Messages are encoded in different patterns of acoustic energy with regard to a common protocol (the language), which both the speaker and the listener must be familiar with to be able to encode and decode the message. The characteristics of speech vary depending on the individual's voice properties — determined by the physical shape of their vocal organs — and their manner of articulation. Figure 2.1 shows a schematic midsagittal view of vocal tract contours, with the articulators and regions specified.

The speech wave is the result of a set of consequent articulatory movements in the human vocal tract, where distinct movements lead to generation of sounds. To produce speech, source energy is generated by pushing the air through the vocal folds, or by frication resulting in a turbulent airflow. The articulators in the human vocal tract move and take different postures to form irregularly-shaped tubes in the vocal tract. The tube acts as a filter, and modify the source energy to generate

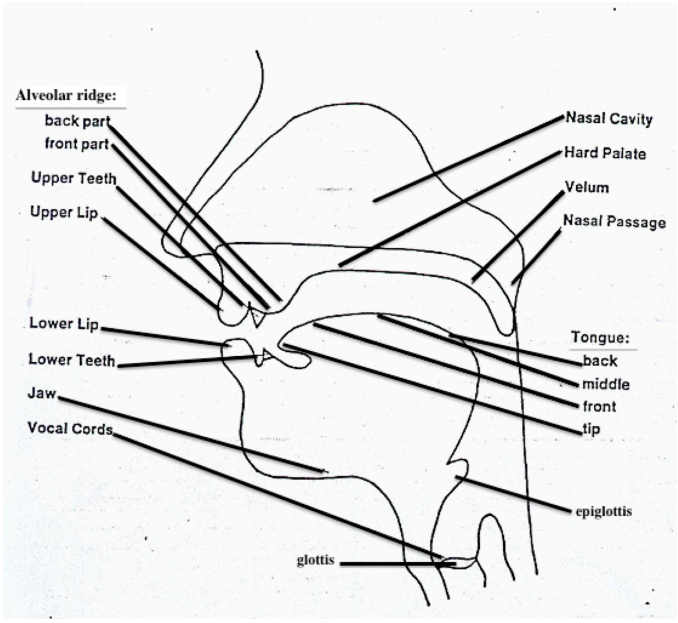


Figure 2.1: A schematic view of the vocal tract contour and articulators (original image from http://www.csulb.edu/∼phoneme/organs.html).

different sounds. Each of the resulting shapes of the vocal tract produces a particular pattern of acoustic energy, and subsequently a certain sound. Although details of the same sounds produced by different individuals are not identical, a common pattern can often be found in their spectral features and their articulation. Sounds can be categorised based on the movements of the vocal folds and places of articulation. For example, vibration of the vocal folds leads to voiced sounds. Other categories of sounds including stops, fricatives, liquids and nasals are defined based on the places of the articulation or positions of the articulators when producing the sounds. Since the movements of articulators — tongue, velum, lips and jaw — result in the formation of the different patterns of energy, consequently, the acoustic energy can be defined as a function of articulatory parameters. Manner of articulation is the way the constrictions in vocal tract impede the airflow and cause different resonant frequencies. The vowels generally are produced with an open vocal tract, while most of the consonants are produced with some sort of constriction or frication in the vocal tract.

2.2 Methods for observing human speech production

Articulation occurs inside a human's vocal tract and thus is very difficult to observe during speech production. Observing and recording the articulation process directly is possible either by physiological interventions or by invasive or non-invasive medical imaging techniques. Depending on the application, several tools and techniques have been used for scanning and measuring the articulation. Some techniques are designed to capture the movements and positions of articulators individually or together, while others capture the entire shape and dynamics of vocal tract from the glottis to lips. Observation tools can be categorised into *invasive* and *non-invasive* methods, where invasive techniques mostly capture movements and deformations by fixing coils or other sensors to the articulators or target places of articulation.

Electropalatography (EPG) [Hardcastle 1972] is a technique used for capturing the constrictions in the vocal tract, recorded through contact of the tongue and an artificial palate. A set of electrode sensors are fixed on an acrylic plate, fitted to the palate, inner and outer areas of the teeth. The tongue's contacts with the palate and teeth are detected by the electrodes and recorded along the time domain. EPG is an invasive technique hence it may affect the naturalness of speech. Also, its application is limited to recording constrictions resulted from tongue deformations and movements, in the oral cavity, anterior to the uvula.

Electromagnetic articulography (EMA), also known as electromagnetic midsagittal articulography (EMMA), is another invasive techniques applied to capture the articulators' motions [Schönle *et al.* 1987; Perkell *et al.* 1992]. Transducer coils are fitted on to the articulator along the midline of the vocal tract, and emit alternating currents when exposed to electromagnetic fields. Usually three electromagnetic sources of different frequencies are used. The system triangulates the distance measurements between the coils and the transmitters to record articulators' positions. Large data repositories of articulators' movements can be collected using EMMA; however, this method is limited to supra glottal midsagittal section of vocal tract anterior to the uvula, and cannot record the dynamics of all of the tract. In addition,

it is difficult to prepare and set up and is rather sensitive to alignments.

The above methods can capture the movements and dynamics of articulators. However, since they only capture the movements of some of the articulators, their application is limited when modelling speech production. For example, the pharynx and larynx cannot be captured with the above methods. In addition, they require devices to be fitted inside a human's vocal tract or attached to the articulators. The necessary interventions for direct observation often interfere with the subject's natural articulation process and consequently may affect the naturalness of speech.

Non-invasive techniques mostly include imaging methods such as X-ray, ultrasound, and magnetic resonance imaging (MRI). X-ray radiography has been widely used to capture images of entire vocal tract [Russell 1928; Fant 1960; Fujimura *et al.* 1973; Kiritani *et al.* 1977; Wood 1979; Johansson *et al.* 1983; Munhall *et al.* 1998] . The method uses high-frequency electromagnetic radiation, where X-rays can either be absorbed by the body organ or pass through. X-ray radiography is strong in capturing the bone structures, while its ability in capturing soft tissues is very limited. X-ray computed tomography (CT) can be used for digitisation of information collected by X-rays via computerised methods, rather than exposing them to photographic films. CT can be effectively used to capture 3D information that is needed about the tract shape, however it has been used to capture cross-sections at only few positions along the vocal tract length of a few vowels [Perrier *et al.* 1992]. The reason for this shortage in available data is that, in order to capture the entire length, for example at 5 mm intervals, the subject must be exposed to several times the X-ray dosages regarded as ethically acceptable in terms of risks. In addition, the design of X-ray systems limit the manoeuvrability of the subject, and any type of changes by repositioning the subject to capture images in another plane results in changing the head posture of the subject, hence a significant change in tract shape [Baer *et al.* 1987]. X-ray microbeam systems [Westbury *et al.* 1990] record the motions of articulators using pulses of narrow high-energy X-ray beam. The beam is rapidly directed to track motions of small gold pellets glued to tongue,

jaw, soft palate, and lips. X-ray microbeam systems are therefore categorised as invasive. The main drawback of X-ray imaging methods in capturing articulation is the restriction of its application due to the hazardous effects of being exposed to X-rays. Furthermore, soft tissues are captured at a very low resolution and therefore not all the articulators can be captured with high precision by this method.

Ultrasound is another imaging technique for capturing articulation dynamics [Kelsey *et al.* 1969; Sonies *et al.* 1981; Stone 1990; Gick 2002; Davidson 2012]. The ultrasound imaging technique uses high frequency sound waves to record tissue densities. The air–tissue boundaries are recognised based on the reflection behaviour of the sound waves when they hit boundaries, and the echoes returned. The ultrasound tool is portable, safe, non-invasive, and popular for modelling tongue dynamics. In addition, with high-speed ultrasound system, up to 60 frames per second can be captured [Wrench and Scobbie 2011]. Since the echoes at the air-soft tissue boundaries are very strong, the sound wave can hardly travel further through the airway, and thus areas such as hard palate cannot be captured. Thus, ultrasound techniques can only capture some articulators, and consequently are useful mostly for collecting information about the tongue shape and places of constrictions.

Magnetic resonance imaging is widely used for capturing images of the human body. The safe and non-invasive nature of MRI has made it a more popular choice compared to other biomedical imaging techniques [Baer *et al.* 1987; Narayanan *et al.* 1995; Foldvik *et al.* 1995; Story *et al.* 1996; Demolin *et al.* 1997; 2000; 2002]. MR image acquisition works based on the resonances of atomic hydrogen nuclei induced by radio frequency pulses applied to a particular slice of human body in a magnetic field. The contrast of soft tissues and quality of images have made MRI a popular choice in imaging speech articulation. In comparison to other safe and non-invasive techniques, MRI can capture the entire vocal tract area and is not limited to capturing some of the articulators. Several protocols have been developed to adapt MRI to capturing speech articulation. The MRI technique and its adaptation in speech research are thoroughly explained in chapter 3.

2.3 Modelling articulation

During the past decades, several approaches have been proposed for modelling articulation and the vocal tract. Most common techniques focus either on describing and modelling the entire shape of the tract, or modelling the articulators separately. More recently, a few approaches have been developed to estimate and study constriction locations and degrees. We review some of these approaches below.

2.3.1 Vocal tract modelling

In speech production studies, the vocal tract shape is traditionally modelled on the foundation of the acoustical tube model proposed by Fant [1960]. In his *Acoustic Theory of Speech Production*, speech is generated by a source-filter model. The vocal tract acts as a filter to modify the source energy released from the lungs through the larynx. His theory requires us to view the vocal tract as a tube with varying cross-sectional areas that can deform by moving the articulators to take irregular shapes for filtering the source differently. The tube can be described by a set of smaller tubes usually with equal length but different diameters, hence different cross-sectional areas. Fant's theory of an acoustic tube with varying cross-sectional areas is the basis for area function modelling of the vocal tract. The cross-sectional areas can be either measured directly from 3D models/images or estimated from 2D images.

Early studies of vocal tract shape modelling were based on lateral radiography, and the transverse areas were calculated from the airway widths [Chiba and Kajiyama 1941; Fant 1960; Heinz and Stevens 1964; Perkell 1969; Sundberg *et al.* 1987]. Since most of the data were available in 2D, rather than 3D, a main problem in articulation modelling based on area functions is estimating a generic area function formula that converts the midsagittal distances to area functions. Various methods such as using casts of the vocal tract [Ladefoged *et al.* 1971; Sundberg *et al.* 1987], measurements of cadavers [Heinz and Stevens 1964], and using callipers [Anthony 1964] have been used to derive an equation for area function.

Baer *et al.* [1988] were among the first researchers to generate 3D models of vocal tract from MRI data (although MRI had been used before by Rokkaku *et al.* [1986] for capturing articulation). They obtained morphological data for American English vowels, and proposed a linear relationship between midsagittal width and cross-sectional areas.

Story [2001] calculated the area functions for 12 vowels, 3 nasals and 3 plosives from MRI data of phoneme articulations in static postures. Narayanan *et al.* [1995] presented measurements and analysis of vocal tract length and area functions for fricatives using magnetic resonance imaging. In later work, Narayanan *et al.* [1997] investigated the geometry of the vocal tract for liquids using MRI and EPG, by direct measurements on coronal images. This work was further extended to the study of rhotics [Alwan *et al.* 1997].

Badin *et al.* [1998] developed a 3D model of articulation using *guided PCA* previously applied to sagittal (2D) data by Beautemps *et al.* [1996]. The articulatory parameters considered for generating the model included jaw height, tongue body, tongue dorsum, tongue tip, and tongue advancement. The model resolved the need for estimating the cross-sectional areas from midsagittal distances; however, it required 3D data to build the model. Their work was further extended to Swedish vowels by Engwall [2000], where tongue and lip models were estimated using the same approach.

Vocal tract area functions were calculated by Takemoto *et al.* [2006] using 3D cine-MRI data (multi-planar 2D cine-MRI). A synchronised sampling method with external trigger pulses has been developed by Masaki *et al.* [1997] to record movements of the speech organs as a set of sequential images. To calculate the area functions, first the vocal tract midline was decided semi-automatically, next the images perpendicular to the midline were re-sliced along the vocal tract at specific intervals. Finally, the sections for each plane were measured to obtain the area of the section orthogonal to the midline.

A hyperbolic grid-generation algorithm was devised by Milenkovic *et al.* [2010]

for investigating vocal tract curvatures. The idea is to use a 2D curve to identify the displacements of inner vocal tract outline (lower lips, tongue and anterior pharyngeal wall) from outer vocal tract outline (upper lips, palate and posterior pharyngeal wall). The area functions were then estimated from the articulatory parameters that describe these displacements.

In recent work, Vasconcelos *et al.* [2011] modelled the vocal tract shapes obtained from MRI data using a statistical deformable model (point distribution models), commonly used in the field of image processing. Unlike most of the statistical models used previously for modelling articulation, the vocal tract was treated as a uniform shape rather than a set of articulators. Deformable models were applied to describe the vocal tract shape of European Portuguese articulations in MR images, and to identify the shape in new images.

Midsagittal distance to cross-sectional areas. Since most of the available data for modelling articulation is in 2D, it is very challenging to obtain cross-sectional areas from midsagittal distances. Several approaches have been proposed for obtaining the area functions from midsagittal airway widths.

The most popular method for area function estimation is Heinz and Stevens [1964] classic (α, β) model. In this model, the relation between the midsagittal width and the area function all along the vocal tract length is defined in terms of a power function as

$$A(x) = \alpha d(x)^{\beta}, \tag{2.1}$$

where A is the cross-sectional area, x is the distance along the vocal tract midline, and d is the midsagittal width. The parameters α and β are the transformation parameters estimated from cross-sectional areas for each speaker specifically.

Similarly, Ladefoged *et al.* [1971] demonstrated an almost linear relationship between midsagittal width and area functions. Sundberg [1969] approached this problem by considering two different regions in estimations of area functions: the

buccal zone and pharynx. In the buccal zone, the values of 1.33 to 1.47 were obtained for β, where as α varied between 2.07 to 2.63. In pharynx, they considered an elliptic structure, with d_1 representing the minor axis of ellipse, i.e. midsagittal distance, and d_2 representing the major axis corresponding to the cross-dimension. At the pharynx, the area function A was obtained by

$$A = \frac{\pi}{4}\ d_1\ d_2. \tag{2.2}$$

Later, Johansson *et al.* [1983] used X-ray scans to study the relation between the midsagittal distance and area functions and proposed a linear relationship between the cross-sectional areas and square of midsagittal distances,

$$A = c_1\ d_1^2\ +\ c_2, \tag{2.3}$$

where c_1 and c_2 are coefficients that not only change along the regions of the vocal tract, but also take different values for different speakers [Johansson *et al.* 1983]. Thus, the parameters of the model must be adapted for different regions and different genders. Further studies [Perrier *et al.* 1992] confirmed a similar theory, in which the vocal tract length was divided into height regions and parameters of area functions were estimated for each region.

Maeda [1972] rejected the (α, β) model of Heinz and Stevens [1964] arguing that the hard palate area cannot be linearly modelled since tongue surface is not flat. Instead, he suggested that at the hard palate the area should be modelled by a hyperbola. He used the same elliptic-based approach of Sundberg [1969] at the pharyngeal region, but took into account the generally concave surface of the tongue. The area function is defined as

$$A_n = \frac{\pi}{4}\ s_n\ d_{1n}, \tag{2.4}$$

where A_n is the area function at those regions in which surface of the tongue is more concave, with s_n representing the major axis of the ellipse, and d_{1n} representing

midsagittal distance. At the hard palate, the area function was defined as

$$A_n = a_n \; {s_n}^{b_n} \left(s_n/(b_n + 1) \; - \; (s_n - d_{1n})/(k + 1) \right) \tag{2.5}$$

where a_n and b_n are coefficients for describing the palatal shape, and k describes the curvature of the tongue contour, and s_n is the distance between the top of the hard palate and the intersection of the tongue edges and the hard palate.

Soquet *et al.* [1998] carried out a comparative study between the dominant approaches in the literature and suggested that Heinz and Steven's (α, β) model is the most robust across speakers (i.e. it has the least variances). Their work further illustrated that the parameters of models are speaker-specific and must be adapted for each speaker.

Other methods have been proposed for obtaining cross-sectional areas such as Tiede and Yehia [1996]'s method for representing the cross-sectional areas using factor analysis of experimental data, or using control articulatory parameters in different planes to estimate the cross-sectional areas [Sorokin 1992] (see [Badin *et al.* 2005] for a comprehensive review).

More recently, Badin *et al.* [2005] extended the method in [Sorokin 1992] for obtaining cross-sectional areas based on sagittal, lateral and axial parameters of the vocal tract, using MRI data. The algorithm devised for modelling the articulation applies a set of parameters such as the height of the glottis, coordinates of the tongue root, coordinates of the tongue tip, etc. In addition to these control parameters, they include the anatomical parameters at the pharynx for achieving better accuracy in estimated cross-sectional areas.

In very recent work, Lammert *et al.* [2013] applied their previous method for recognising the regions-of-interest (ROIs) [Lammert *et al.* 2010] to estimate the cross-sectional areas of the vocal tract in real-time MR images. Circular regions were placed at the determined centres of regions-of-interests, orthogonal to the midline, with the radii chosen so that the circular region would just fill the cavity. The cross-distances were estimated, for each subject individually, based on the changes of the

intensity across the range of a sequence of real-time MR images (rtMRI movies).

2.3.2 Articulator modelling

Another approach for modelling articulation is to model articulators and their dynamics instead of the entire shape of the vocal tract. The proposed techniques mainly focus on statistical modelling of the tongue, lips and jaw. These can be thought of as control parameters, or as reduction of vocal tract shape to fewer dimensions.

In preliminary attempts to model articulators statistically, Harshman *et al.* [1977] applied component analysis to identify a set of features that can model the tongue shape for 10 English vowels. Maeda [1988] used a factor analysis in connection with principal component analysis to derive a linear model of lateral shapes of the vocal tract. The components in his linear model were defined in terms of articulatory features such as jaw height and tongue body position. Principal components of cross-sections of tongue shapes were determined by Stone *et al.* [1997].

Beautemps *et al.* [1996] were also among the first to develop statistical models for tongue shape. The model adopted guided PCA for defining factors of the tongue and jaw. They successfully predicted tongue shape with 88% variation. However, the tongue tip was excluded from their analysis as the method proved to be insufficient for describing the tongue tip. In later work, they improved the extent of tongue variation, covering up to 96% of variance. A larger set of parameters was used for defining articulators, namely: jaw height, tongue blade, tongue dorsum, tongue tip, tongue advancement, larynx height, lip protrusion, and lip vertical positions. Their technique was applied to 3D linear articulatory modelling of tongue, lips and face [Badin *et al.* 2002]. The data used for developing the model was 3D MR images and videos of a subject producing sustained articulation of French phonemes.

A three-dimensional model of the tongue was reported by Engwall [2003] based on MRI, EPG and EMA of static Swedish vowels. Similar to Badin *et al.* [2002]'s approach, the 3D images were collected in three stacks of parallel axial, coronal and sagittal frames. The articulation was modelled in terms of a set of parameters

including jaw height, tongue tip, tongue blade, tongue dorsum, tongue advance and tongue width. The articulatory model was a linear model controlled by the five introduced parameters. Their model covered 78% of variation of the 3D shapes. They suggested that the 3D tongue shape can be well defined and controlled by parameters defined in the sagittal plane.

Several studies have been carried out to model the articulators such as lips, jaw, cheeks and tongue for animation purposes. However, since in animation the visible parts only include lips, teeth and cheeks, and maybe the tongue tip or slightly the tongue dorsum, the main goal has been to model lips and teeth naturally, although the interaction of the tongue tip with teeth and lips is quite valuable. Examples of proposed techniques for modelling visible articulators can be found in [Badin *et al.* 2002; Ma and Cole 2004; King and Parent 2005].

Avila-García *et al.* [2005] developed a technique for modelling tongue and lips by combining active shape models (ASM) [Cootes *et al.* 1995] and Hough transforms [Hough 1962]: the active shape model dynamic Hough transform (ASDHT). They developed the model using images from a dynamic MRI database of articulation of a nonsense utterance. Their model was tested on three different corpora of original tongue shapes in the training set, unseen or test set and synthetic tongue image sequences.

In a relatively recent study, Bresch and Narayanan [2009] proposed an approach for automatic region-segmentation based on tract variables. Some of the tract variables of interest included lip aperture, velum aperture, and tongue tip constriction degree. The task was defined in terms of defining the air–tissue boundary and extraction of individual articulators for computing tract variables.

2.3.3 Modelling constriction

In addition to modelling the entire vocal tract in terms of tubes and area functions and modelling articulators individually, a few studies have focused on studying only the areas of constrictions in the vocal tract. For example, in [Bresch and Narayanan

2009], the ultimate goal is to trace the articulators contours automatically to estimate the desired areas of constriction.

Recently, Lammert *et al.* [2010] proposed an alternative approach for observing anatomically driven methods for studying articulation. They suggested investigating the co-variation of pixels across a sequence of images over time to discover the regions with high-correlating pixels. The variation of intensity in the observed regions illustrates the constrictions occurring over time in the vocal tract. They further investigated the consonantal speech kinematics by automatically estimating the articulatory constrictions in a series of real-time MR images [Proctor *et al.* 2011]. The maximally-dynamic correlated regions of pixel activity at the palate, the alveolar ridge and lips across the range of image were recognised automatically. This method was further used in [Lammert *et al.* 2013] to estimate cross-distances from 2D midsagittal images.

2.4 Discussion

As we reviewed above, various protocols have been proposed for modelling human articulation. Despite the variety and differences, a common factor that exists in modelling the vocal tract shape either statistically or in terms of area functions is that the models are mostly obtained from static images that are typically limited in quantity. The initial data to develop the model can therefore be manually traced, or semi-automatic methods with human supervision can be adopted. This is only feasible when using a small quantity of data. If the models are to be generated from a substantial volume of data, manual or semi-automatic methods for preparation and extracting information may not be feasible. Therefore, automatic methods for obtaining information are desirable.

In this thesis, we aim to study the articulation using minimal human supervision for extracting data. Our focus is to first examine the extent and nature of information that can be obtained without using shape priors or recognising the anatomical contours of articulators. We further extend our aim of automatic feature extraction

from the MR image to segmenting the vocal tract shapes in images.

2.5 Summary

In this chapter, we have reviewed some of the fundamentals of speech production research such as methods of acquiring data and modelling articulation. In the next chapter, we extensively review the application of MRI to speech production research, and various protocols devised to improve the ability to capture running speech using MRI. The information extraction and vocal tract modelling approaches used by other scholars are explained and discussed. Finally, we introduce the MRI database that is used in this research.

Chapter 3

Speech Production Research using MRI

3.1 Introduction

Magnetic resonance imaging (MRI) is a method of obtaining high quality images of the human body, and is frequently used in medical settings. The safe and non-invasive nature of MRI, together with its power to acquire high resolution multiplanar image slices of soft tissues, make it a very popular and promising tool for studies of human articulation.

Compared to other modalities utilised for capturing the articulation such as X-ray or EMA (reviewed in chapter 2), magnetic resonance imaging generates image data at a slower rate. A series of related techniques has been developed to address this issue, such as sustaining the articulation for a certain period of time [Baer *et al.* 1987; Narayanan *et al.* 1995], or obtaining high spatio-temporal sequence of images by collecting the data over several repetitions of articulation [Foldvik *et al.* 1995]. More recent advances in the magnetic resonance pulse sequence allows data to be obtained at a higher rate without much loss of accuracy in temporal or spatial domains [Demolin *et al.* 2000; 2002].

Improvements in the technology have facilitated collection of extensive MR image

databases of articulation. Thus, a new challenge emerged: to analyse and extract useful information from the images. Processing large sequences of images manually is very time consuming and labour intensive, hence more automatic and less supervised approaches are desirable for extracting useful features.

We begin this chapter by reviewing the various MRI protocols devised for capturing the human vocal tract and its kinematics (section 3.2). In section 3.3, we present an overview of the different approaches proposed and investigated in the literature for analysing articulation images. A description of Oxford University Phonetics Laboratory's MRI database of human articulation is provided in section 3.4, followed by a review of the image pre-processing we carried out to improve the quality of images, in section 3.5. We continue this chapter by illustrating the information that can be obtained by looking at the variation in images in a sequence (section 3.6). Finally, a conclusion on the general application of MR imaging in articulation and an introduction to our proposed approach are presented in section 3.7.

3.2 MRI techniques for capturing articulation

The foundation of MRI technology lies in the fact that proportions of fat and water in the human body behave differently when they are positioned in a strong magnetic field. MRI is based on proton nuclear magnetic resonance (NMR) that can detect presence of protons when subjected to a large magnetic field. Therefore, MRI is capable of capturing the concentration of protons. The density of protons is greater in water, and consequently soft tissues containing more water are better captured with MRI.

A sample MR image of the vocal tract in the sagittal plane is presented in Figure 3.1. [1] To capture the MR images, a subject usually lies supine inside the

[1]The image artifact at the top of the speaker's head is caused by scanning the area beyond the "field of view" of the image. These areas are aliased back to the image and result in the wrap-around. In this case, tissue from the subject's neck is wrapped around to appear over the top of the head. Normally, such artefacts are avoided by scanning a wider field of view, however this adds extra time to image acquisition for no improvement to the view of the region of interest – the vocal tract.

MRI scanner. Images of the desired part of the anatomy can be acquired in different orientations such as axial, coronal and sagittal.

Figure 3.1: An MR image of a subject articulating the /s/ in the middle of the word "saucer".

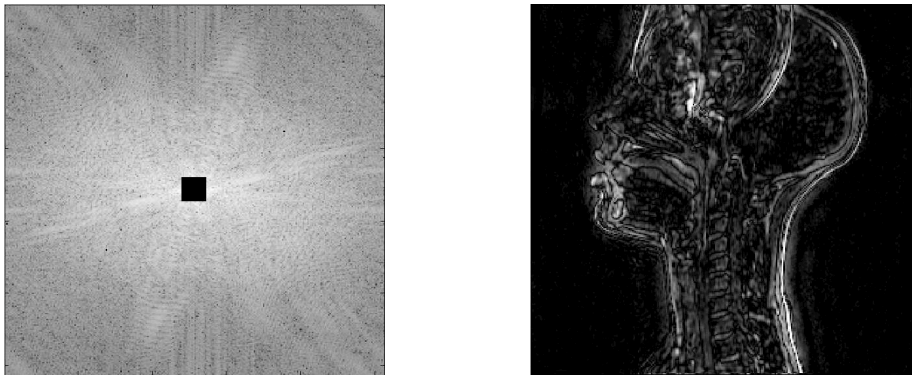
When a constant magnetic field is applied to tissues, the nuclei come to align with the magnetic field. A radiofrequency (RF) pulse is applied which disturbs the precession direction of the nuclei in the tissues, making the nuclei resonate at the same frequency as the hydrogen nuclei. The nuclei absorb the energy and come in phase with each other. As a result, the net magnetic vector (NMV) precesses in the transverse plane. The receiver coil is situated in the transverse plane, and when the NVM rotates around the transverse plane, it passes the receiver coil. Hence, a voltage is induced that is referred to as the *MR signal*. The MR signal is received as a function of time and is transformed into the frequency domain by Fourier transformation.

To capture a particular section of interest in the body, a gradient coil is used that generates a gradient field to excite the section of interest. The magnetic field is applied at a certain flip angle so that different intensities are applied to different sections in the body. The RF pulses are then selected according to the range of

magnetic fields that correspond to the section of interest. Three gradient magnetic fields are applied, one to determine the bandwidth of the frequency domain, one to modify the phase in each row of the selected slice (y direction) and one to modify the frequency on each column of the selected slice (x direction). The composite MR signal thus generated has specific frequency and phase characteristics that make recognising the tissue and the spatial locations in the tissue possible [Westbrook *et al.* 1998].

The composite signals are stored row by row in a matrix referred to as *K-space* until an entire slice is captured. K-space is in the frequency-phase domain, so an inverse Fourier transform is applied to generate the final spatial image from the K-space. The K-space matrix has two axes, both centred in the middle of the matrix space: the horizontal axis is the frequency axis and the vertical axis is the phase axis. The horizontal lines correspond to phases during the acquisition, while the data in the lines are the frequencies sampled during the scan. K-space matrix is symmetric around its axis. The centre data points in K-space contain data with low spatial resolution, but high signal amplitude, while the outer data points contain information of high spatial resolution and low signal amplitude. The strength of the received signal is reflected in computed images in terms of brightness. Figures 3.2 and 3.3 show images reconstructed by using only outer and inner K-space data points, respectively.

The duration of the scan in MRI depends on a few factors including the time it takes to fill each line of the K-space, number of lines in the K-space to be filled, and the number of times each row in K-space is sampled. Traversing K-space is therefore an important factor in scan time, and as we will explore in next sections, different K-space traversing approaches can be used to reduce the scan time in acquiring articulation images. In addition, the symmetric property of K-space can be used to reduce the dimensionality of data in large scale applications.



(a) K-space (b) reconstructed image

Figure 3.2: Image reconstructed using the outer K-space data points. The image has high resolution, but poor signal amplitude.

(a) K-space (b) reconstructed image

Figure 3.3: Image reconstructed using the central K-space data points. The image has strong signal amplitude, but poor resolution.

3.2.1 Static MRI

The relatively long acquisition time required to capture the entire anatomy of the vocal tract in contrast to the fast moving articulation mechanism limited early vocal tract MR imaging investigations to *static* articulations. For example, the maximum velocity of velar motion has been estimated to be around 14.1 cm/s and a relatively fast tongue tip motion has been estimated to have a velocity of 80 cm/s [Perkell *et al.* 1992; Engelke *et al.* 1996], while acquisition of the entire vocal tract shape, traditionally, could take from few seconds to minutes [Baer *et al.* 1987]. In static

MRI, the subject sustains a particular articulation long enough for a round of image acquisition to be complete. Consequently, the speech sounds to be studied are restricted to those with sustainable articulations.

Baer and colleagues were among the first scientists to investigate MR imaging of the vocal tract [Baer *et al.* 1987; Baer 1991]. In [Baer *et al.* 1987], the authors initially studied only the vowels /ɑ/ and /i/ articulated by two male speakers. Their research was extended later to the four point vowels by adding /æ/ and /u/ to their database [Baer 1991]. Their experiments comprised calculating the area functions of the vocal tract for the two male speakers and using the calculations to create synthetic speech, which was compared to the original speech both perceptually and acoustically. Further, they performed an analysis of the relation between the cross-sectional area and the midsagittal width for different vowel configurations.

Static MRI was applied in a few other studies focusing on evaluation of its precision and predicting the capability for quantitative measures obtained from MR images to estimate vocal tract resonances [Moore 1992; Greenwood *et al.* 1992]. These studies concluded that despite the relatively large errors, the obtained parameters can effectively predict the acoustic spectral measures and the resonance frequencies of the vocal tract.

Narayanan *et al.* [1995] pioneered the study of sustained consonant articulations with static MRI. The length of the vocal tract and the area functions were extracted from the MR articulation images of {/s/, /f/, /ʃ/, /θ/, /z/, /ʒ/, /v/ and /ð/}. The authors presented comprehensive morphological analyses of vocal tract and tongue shapes based on the MR images of four different speakers, but inter-speaker comparisons were not attempted.

Several studies of 2D and 3D vocal tract shape analysis based on static MRI have been carried out since these first applications of MRI to imaging the vocal tract, e.g. [Sulter *et al.* 1992; Dang *et al.* 1993; Story *et al.* 1996; Alwan *et al.* 1997; Narayanan *et al.* 1997; Ong and Stone 1998; Narayanan *et al.* 1999]. However, the nature of the static MR imaging protocol yields a few disadvantages in its applica-

tion. The long acquisition time restricts the method to the sounds with sustainable and prolongable articulations. Prolonging the articulation in turn introduces subject fatigue, which leads to inconsistencies caused by variation in the vocal tract shape. In addition, it is very unnatural and is not representative of real speech. Hence, inaccuracies in the succeeding measurements become inevitable. The most challenging issue with static MRI is the lack of ability to capture the *dynamics* of speech production. Speech production is a continuous and running activity, and measurements obtained from sustained production may not be representative of natural speech production, particularly in terms of jaw position, lip protrusion and tongue contours [Engwall 2000].

3.2.2 Dynamic MRI

A few attempts have been made for capturing the movements of the vocal tract, for example by using faster MRI devices [Mády *et al.* 2001; Hoole *et al.* 2000] or tagging the MRI [Stone and Lundberg 1996], but still the long acquisition time and the poor signal-to-noise ratio do not allow the dynamics of speech production to be captured.

The idea of the *dynamic* MRI protocol comes from the application of retrospective gating in cardiac MR imaging that can effectively capture motions of a beating heart. In the retrospective gating technique, data are acquired from an imaging scanner and an electrocardiogram (ECG) in an uninterrupted sequence. The acquisition of images is triggered by the ECG. The resulting image sequence represents the heartbeat at different phases.

If a motion is repetitive and the collection of the images is synchronised with the help of an external "clock", the movements can be captured through sequences of images. Dynamic MRI of speech was developed based on the same idea of using repetitions and capturing images at different phases over the course of repetitions. Token repetitions result in generating periodic movement of articulators similar to the heart cycle. The image acquisition in each repetition is triggered by an external stimulus. During each repetition, a "pseudo-frame" (K-space) is filled with rows of

data acquired, each row corresponding to a different phase during token articulation. The rows to be filled are selected by randomising the start of the acquisition relative to the utterance, but the number of rows to be collected is pre-determined. The number of pseudo-frames to be collected, N, depends on the repetition time and duration of each token. Once the acquisition is over, synchronised K-spaces are generated, offline, by accumulating the corresponding pseudo-frames that were captured at different times (i.e. different tokens) but correspond to the same phase of the articulation of the token. That is, synchronised K-space matrices are generated by including rows in the pseudo-frames that correspond to the same phase. If a row is sampled more than once, the new row in K-space is an average of the sampled rows, and if a row was missed in sampling, the nearest rows' information is used. The final images are reconstructed from the synchronised K-spaces using the inverse Fourier transform.

Foldvik *et al.* [1995] proposed a new technique for capturing the kinematics of articulation in MR images by using the *dynamic MRI* protocol. Numerous repetitions of the speech sequence are used to capture partial MR images, with the synchronisation method built-in in the scanner. Whole images are then reconstructed offline from the images collected during the several repetitions of the speech. Despite the relative success of this method compared to previous attempts for capturing articulation movement, the method is very sensitive to precise repetitions of the movements and is very vulnerable to variation. The technique is not robust to any irregularity caused by the subject making a mistake or other changes in articulation, or being short of breath and performs poorly in such cases.

Mohammad *et al.* [1997] addressed this lack of flexibility in dynamic MRI by recording the audio simultaneously inside the scanner, and performing the synchronisation using the recorded acoustics. In this group's method, the scanning start time at each repetition was randomised relative to the beginning of the utterance. During each round of acquisition pseudo-frames are filled sequentially row-by-row, where each row corresponds to a different phase. These pseudo-frames are not very

precise due to the short exposure time and the movement. However, owing to the randomisation, many pseudo-frames are collected with their rows corresponding to different time points in the utterance. In post-processing, the K-space pseudo-frames are synchronised using the recorded audio, and the pseudo-frames corresponding to the same phase in the utterance are accumulated to generate a synchronised K-space. They later extended their work to a multiplanar system [Shadle *et al.* 1999].

Dynamic MRI proves to be successful in imaging the articulation at rather fine temporal and spatial resolutions [Shadle *et al.* 1999; Alvey *et al.* 2008]. The analysis of images obtained by dynamic MRI suggests that the technique enables accurate volumetric measurements of the vocal tract to be obtained [Shadle *et al.* 1999]. The database we use in this research, described in detail in section 3.4, was collected using the dynamic MRI technique.

3.2.3 Real-time MRI

Real-time MR imaging of the vocal tract refers to imaging the articulatory movements directly at the time of scanning. Demolin *et al.* [2000; 2002] made significant improvements in addressing the challenge of real time acquisition of dynamic speech, by implementing an ultra-fast MR imaging method called turbo spin echo (TSE) zoom. The articulatory movements were visualised well as a direct result of clear air–tissue delineation. To prove the validity of their method, the authors compared real-time and static images of the same vowel articulation by the same speaker. The measurement metric was the midsagittal distance measured on a grid on both types of image. The analysis suggested that real-time MRI can in fact provide precise and reliable information about the position of the articulators involved in speech production. However, the limited imaging rate of 4–6 frames per second (166–250 milliseconds per frame) is considered a relatively low temporal resolution compared to the fast rate of articulation itself (e.g. from tens of milliseconds to a few hundredths of a second for vowels, tens of ms for an aspiration phonation, and milliseconds for a short burst). Attempts to improve the temporal resolution result

in obtaining less information for each image taken, hence the lower spatial resolution of the articulation.

Narayanan *et al.* [2004] successfully improved the temporal resolution of real-time MR imaging of the upper airway using gradient echo imaging with a fast interleaved *spiral* acquisition strategy. In their method, instead of fully sampled read-out in the K-space domain, the data were read in a spiralling pattern, which is faster than sequential read-out. The time efficiency of spirals allowed for an image acquisition rate of 110 ms (9 images per second) and an image reconstruction rate of up to 24 images per second (using a sliding window that made a more frequent reconstruction of the images possible). However, as with TSE zoom [Demolin *et al.* 2002], spatial resolution and image quality remain open challenges with this technique. Figure 3.4 (a) and (b) show two MRI images captured with TSE zoom and spiral acquisitions respectively, and (c) an MR image obtained with dynamic MRI. The spatial resolution is clearly better in image (c) compared to the image acquired by real-time MRI.

3.3 Analysing images of the vocal tract

There is a common step in studying speech production with MRI, independent of the MRI protocol chosen for capturing the images: the image analysis step. Although MR images of the vocal tract can visually be very informative to a human researcher, further processing is necessary to understand the mechanism of articulation, based on the images. The task of analysing and extracting relevant information from MR images of vocal tract is very challenging, in part due to the nature of the MR image acquisition technique. The similarity of the tissue textures of the vocal tract and the nature of articulation play another part in making information extraction complex. The task of segmentation becomes more difficult when dealing with large databases, particularly if the design is aimed at capturing the speech dynamics. A few of the challenges encountered in the vocal tract image information extraction are listed next.

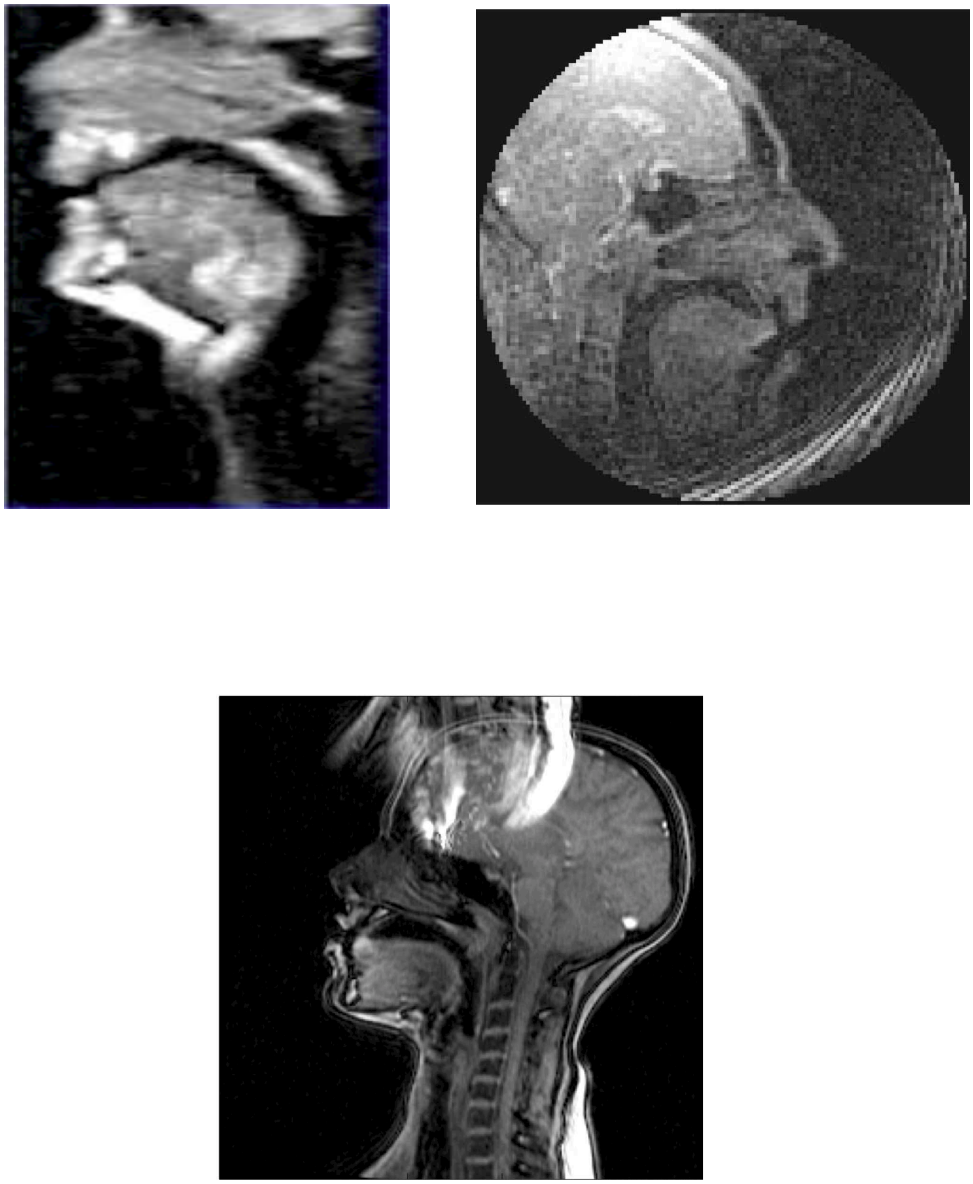


Figure 3.4: MR images captured using TSE zoom [Demolin *et al.* 2000] (top left), and spiral acquisition [Narayanan *et al.* 2004] (top right), and dynamic MRI [Alvey *et al.* 2008] (bottom). The acquisition rate is faster using spiral acquisition (8–9 frames per second), while TSE zoom can capture 5–6 images per second. There is a trade-off between the spatial and temporal resolutions in real-time MRI acquisition. The pixel size of the image in TSE zoom is 3.9 × 1.95 mm, and is 2.7 × 2.7 mm in spiral acquisition, while this number can be as small as 1.79 × 1.79 mm in dynamic MRI [Alvey *et al.* 2008].

Noise and blur in dynamic and real-time MRIs. The images collected with the purpose of capturing speech dynamics typically include a substantial amount of noise and blur, particularly around the edges (air–tissue boundaries). Using noise-reduction algorithms can help to some extent, but some useful information about the articulators might also be removed by noise-reduction algorithms.

Non-identification of the teeth. Calcified structures such as teeth and bones have small concentration of hydrogen, hence emit a very weak signal that is similar to imaging air, which is a problem in speech applications. To address this, some researchers use teeth indentations obtained by other media such as X-ray images or dental impressions and impose them on the images [Baer 1991], or they estimate them by using the information available on other images where there is tissue in direct contact with teeth, e.g. articulation of dental phonemes [Ventura *et al.* 2009; Proctor *et al.* 2010].

Lack of well-defined boundaries along the vocal tract. The back of the vocal tract can be open to the nasal cavity through the velum and open to trachea through the glottis. It is therefore very difficult to specify where exactly the nasal cavity or trachea start and where to put boundaries to separate them from the oral cavity in vocal tract segmentation. This is also the case at the mouth opening, i.e. lips, where no boundaries are defined to close the front of the vocal tract.

Surrounding tissue similarities. The intensity gradient of the tissues of the articulators and the surrounding tissues are all very similar in MR images, making it more difficult to separate the relevant vocal tract areas from non-relevant tissues (Figure 3.1). "Breaks" in the palate, linking oral and nasal cavities, can be one such consequence.

Depending on the application, a variety of approaches have been developed to extract the necessary and prominent information from the images. In this section, we review some of these techniques, and explain how our work relates to them.

3.3.1 Automatic, semi-automatic and manual methods

Several factors influence the decision about what *kind* of method to apply for extracting the relevant information from the MR images. These factors include the volume of data to be processed, the expected precision required in the analysis, the type of protocol utilised for collecting the images (e.g. static, dynamic, real-time), and last but not least, the application.

Manual segmentation is usually performed by an expert with use of digital aids. Manual analysis is generally more precise compared to semi-automatic or automatic segmentation as it employs human understanding. However, manual segmentation is rather labour intensive and time consuming, which restricts the method to small databases. As the analysis is based on someone's visual perception, the outcome might be biased.

Semi-automatic methods require supervision to some extent. That is, one or a few parts of the process must be performed/corrected by a human expert, but the procedure overall is almost human-independent. An example of this category is when a human observer is required to label different sections of the vocal tract for the initial input to the system. Depending on the amount of human supervision required, the application of these methods can be extended to larger, more comprehensive databases of MR images. Semi-automatic segmentation methods may also become labour intensive and biased if too much human supervision is needed.

Automatic analysis and processing methods, which require minimum human supervision, are more appropriate when dealing with extensive amounts of data, as processing large quantities of images can be extremely time consuming and labour intensive. In fact, the extensiveness of data facilitates the utilisation of computer-based automatic methods as sufficient data is available to train the system on how to find and extract prominent features from the images. However, where too much variation exists in the data, the automatic methods may result in losing accuracy compared to the manual segmentations.

From the above descriptions, it can be concluded that the MRI protocol used in

collecting the data is an influential factor in choosing the analysis method. Static databases are often very limited due to their nature, and manual or semi-automatic methods can readily be utilised. In contrast, the real-time databases are typically very large and are better analysed using automatic approaches with minimum human supervision. Since dynamic MRI protocol is adopted for gathering extensive as well as limited amounts of data, all three categories of the approaches might be suitable according to the size.

3.3.2 Extracting the shape of the vocal tract

A significant portion of vocal tract MRI studies are devoted to reconstructing the vocal tract shape model for a better understanding of speech production. The main goal is to obtain area functions (see chapter 2) to construct the tube-shape of the vocal tract that can effectively simulate the tract resonances.

The preliminary segmentation methods for extracting the vocal tract shape from MR images were limited to manual or semi-automatic (with considerable manual initiation) approaches. In the initial investigations, a number of researchers used a semi-automatic *thresholding* technique for segmenting the vocal tract shape [Baer 1991; Moore 1992; Badin *et al.* 1998]. The threshold-based approach relied entirely on the image properties for recognising the shape. A calibration method based on water-filled or oil-filled tubes was usually employed in this group of techniques. Crary *et al.* [1996] reported achieving a variation of less than 5% with this method.

Demolin *et al.* [1996] used a very different approach for segmentation and shape modelling by first specifying the outline of the cross section manually with the help of a digital pad, and then computing the area by digital integration, i.e. a polygon surface computation algorithm.

Soquet *et al.* [1998] carried out a comparative study of the two methods mentioned above and a proposed a third method. The third method, the *elastic* method, is again purely intensity-based. A small free-form curve is placed on the region of the interest (i.e. the vocal tract airway), and a computer algorithm makes the curve

expand until it hits the air–tissue boundary and stops. The manual initiation with the digital pad [Demolin *et al.* 1996] was used as the reference contour owing to its accuracy and reproducibility features and also for allowing human expertise on vocal tract geometry to be included. Based on the quantitative results, the authors concluded that all the methods perform very similarly, with threshold methods slightly outperforming the other two. However, they suggested that a drawback in applying these methods for segmentation is sensitivity to the parameter initialisation, such as the threshold level or the free-form curve definition.

The threshold segmentation approach was also adopted to extract the vocal tract contours from a dynamic MRI database of European Portuguese articulations [Ventura *et al.* 2009]. Although the segmentation was performed in a semi-automatic manner using a histogram-based thresholding technique, the overall procedure involved manual manipulations such as closure of mouth opening, larynx and velum.

Kröger *et al.* [2000] adopted a gridline approach to obtain the vocal tract shape parameter information from static MRI images. An initial speaker-specific but articulation-independent grid system was fitted to each image, and the centre line was estimated as the middle points of each gridline. Then, a new gridline system was created by taking into account the obtained centre line of the tract airway and the air–tissue boundary points of the initial grid. The cross-sectional vocal tract areas were estimated from the latter set of gridlines. Although previous methods in the literature had used the gridline approach for calculating the cross-sectional areas, the majority of them applied the gridline to the segmented contour and no method used gridline for extracting the contour directly.

Seeded region-growing [Adams and Bischof 1994] is another technique used for detecting the air–tissue boundaries of the vocal tract [Carbone *et al.* 2013]. In this method, a "seed" pixel inside the vocal tract area is selected and the region expands by comparing the grey values of the neighbouring pixels of the area already delimited to the mean value of grey-levels of the pixels already denoted as pixels inside the vocal tract region. Thresholding is again applied to stop the expansion and the

algorithm will terminate when reaching the maximum difference threshold between the new pixels being tested and the mean grey-level value of the pixels inside the region. Theoretically, the surrounding air has very similar grey-level value to the pixels within the airway, and thus it is expected for the region, unless a blocking method is applied to grow to the surrounding air. However, the authors did not acknowledge having encountered such an issue and did not discuss how it might be addressed.[2] In addition, the "seeds" were placed manually for each image, which makes the segmentation method semi-automatic.

Proctor *et al.* [2010] proposed an intensity-based approach for segmenting the vocal tract shape. A composite analysis grid was superimposed on every image frame to be segmented. The air–tissue boundaries were decided according the pixel intensity threshold along each gridline, and the vocal tract contour was constructed. The centre line of the tract airway was estimated by optimising the path from the glottis to the lips. The contours were constrained by the estimated centre lines. The boundaries were extracted from existing data and imposed on the image. An advantage of this method is that the contours can be corrected where lack of tissues leads to inaccurate segmentation such as the palate or dentition. However, the contour corrections make the method a semi-automatic approach which requires supervision. In addition, the problem of noise and blur, although partially addressed, was not entirely resolved with this approach.

In recent work, Vasconcelos *et al.* [2010; 2011] applied deformable shape models to both modelling and segmenting the entire vocal tract shape in new images. They used a dynamic MRI database of European Portuguese articulations for both training and testing purposes. Initially, landmarks were labelled and placed on the training set images. The decision on the landmark points to include was made considering the important features of the boundary or inner regions of the shape. The landmarks were subsequently used to create a point distribution model (PDM). The statistics of the positions and intensity properties from PDM helped developing a

[2] Although they mentioned that up to 18% of contour were discarded and were marked as outliers to the rest of the contours.

shape model using active shape models (ASMs) [Cootes *et al.* 1995] and active appearance models (AAMs) [Cootes *et al.* 2001]. The vocal tract shapes in the new unobserved (test) images were segmented using the constructed shape models. Vasconcelos *et al.*'s approach to segmentation opens a new door in the field of vocal tract shape extraction. However, a major drawback of their approach is the initial manual labelling and landmark tagging, which is a very time consuming task. Another critical issue associated to their work is the fact that the vocal tract boundary is not smooth and cannot simply be delineated by fitting a straight line between the two landmarks to represent the actual boundary. An extension to their work is also presented in Ventura *et al.* [2012], where the generated models were also employed for predicting the articulatory model for a few consonants.

3.3.3 Tracking articulatory movements or areas of constriction

A second family of vocal tract MR image segmentation approaches focuses on the dynamics of articulation. The main goal is to track the movements and extract the time-varying features of the vocal tract shape. The nature of these methods makes them appropriate choices for dynamic or real-time MRI databases only. The algorithms developed for tracking articulators are typically used for modelling the articulators as well, and thus many of them have been reviewed in the background chapter. Examples of such methods are Avila-García *et al.* [2005]'s tongue shape model extracted and modelled through active shape dynamic Hough transform (ASMDH), Bresch and Narayanan [2009]'s articulator-based contour tracking in the frequency domain applied to region segmentation, or Lammert *et al.* [2010]'s correlated regions of interest.

In [Bresch *et al.* 2006], a semi-automatic procedure was introduced that used the *snakes* algorithm [Kass *et al.* 1988] to track the contours in a series of successive MR images. The snakes algorithm is based on the pixel intensity gradient of the image and optical flow. The procedure starts by manually tracking the contour in the first image in a series of images, and then applying the snake optimisation to find the

contours in the rest of the images. Several challenges remain to be addressed in application of this method. The manual initialisation and determining the weights for different parameters of the snakes algorithm are among the shortcomings. Another limitation is inability to build the contours when there is occlusion in the image, e.g. when the tongue is too close to the palate.

Discussion. In previous sections we reviewed a variety of protocols devised for capturing vocal tract using MR image acquisition techniques. We also discussed how the acquisition protocol, amount of data and application affect the choice of information extraction method to be used. In addition, we reviewed some of the methods used in the literature to extract features from articulation images.

For this thesis, we use a dynamic MRI database of articulation that provides information about the deformations and movements of the articulators. The database, described in the next section, was collected as the output of a previous research project [Kochanski and Coleman 2008]. The original aim was to study articulation and coarticulation in the lower vocal tract. Since the laryngeal region is a small area, having sufficient spatial resolution was a necessity in collecting the data. Furthermore, studying the motions in the vocal tract requires an appropriate temporal resolution that can reflect the rate of articulation. Thus, a dynamic MRI acquisition protocol was employed to obtain fine temporal and spatial resolutions in images. The database contains several thousand images in form of image sequences (i.e. movies). Thus, analysis and information extraction from the database using "automatic" methods is inevitable.

We intend to investigate image analysis methods that can be applied to automatically and successfully extract useful articulatory information from this database (or others like it). The scale of the database restricts our choice of information extraction methods to those that can operate with minimum human supervision, hence excluding manual or semi-automatic approaches. We first focus on extracting features by considering the spatial information and sequential order of the images in the

database. The variation in image-wide spatial information provides us with essential data regarding the deformations in vocal tract shape and articulators' dynamics during speech production. We further develop data-driven approaches that can automatically extract the shape of the vocal tract for purposes of more traditional articulation modelling approaches, such as area function estimation.

In the next section, we describe the MRI database used in this work and present some statistics about the data.

3.4 Oxford University Phonetics Laboratory's MRI data

Oxford University Phonetics Laboratory possesses a database of dynamic MRI image sequences (i.e. movies) and their corresponding acoustic data (i.e. audio recordings), collected as the output of a previous research project [Kochanski and Coleman 2008]. These images were collected using the "gated" acquisition method [Alvey *et al.* 2008], a technique similar to that utilised for imaging a beating heart: for acquiring each image, several pulses (partial images) are collected from the anatomy, and the collected signals are put together to construct a complete image. Normally, the pulses are triggered by an echocardiogram and the images are collected over a course of many heart beats. Each of the pulses captures part of the image, and the image is not complete until all the repetitions are collected. Table 3.1 displays the parameters used for image acquisition of this database, as reported in [Alvey *et al.* 2008]. The imaging machine used was a 1.5 Tesla MRI unit (Signa HDx, GE Medical Systems, Milwaukee, WI, USA).

When this technique is applied to vocal tract imaging, subjects repeat the speech for a sufficient (pre-determined) number of times for the images to be completed. But instead of being triggered by a heart beat, the image acquisitions were synchronised to the articulation by using a metronome. The utterances were chosen to be phrases with four syllables or less to make the repetitions by subject more stable, and similar to a heart beat. The phrases were selected to contain a particular set of phones that are "prolongable", such as nasals, liquids and fricatives, to compensate for the long

Table 3.1: Parameters selected for image acquisition

parameters	values
TR	minimum
TE	minimum
trigger delay	minimum 10 ms
Gated phases to reconstruct	20
matrix size	256 × 256
slice thickness	10 mm
FOV	36 cm
scanning time	18–20 s
number of pseudo-images per phase	20
image temporal resolution	0.05 s

exposure time required for taking whole images. Table 3.2 presents a list of the database utterances and the number of subjects speaking them.

The images were collected from 20 British English native speakers (10 male and 10 female). To acquire the images, the subjects lay in the MRI scanner and articulated an utterance repeatedly, 20 times. The speakers were cued with each utterance prior to each scan by either an audio intercom, or on a 25 inch monitor in the control room that was reflected in a mirror fixed to the head coil. The timing of the repetitions of the utterance was governed by a metronome, and the speakers had been trained to speak in time to the metronome beats (the pace of the metronome had been set to a comfortable speech rate by the subjects themselves in a pre-test). For each utterance, 68 images of 256×256 pixels were captured at the intervals set according to the metronome rate. The images are mid-sagittal views from the bottom of the subjects' neck up to the top of the head. Figure 3.5 shows sample MRI images from the database. The images provide a clear view of the subject's entire vocal tract. The audio signal was simultaneously recorded during the repetitions using a non-magnetic microphone that was fitted inside the scanner approximately 5 cm from the subject's mouth. MRI scanners are generally very noisy and loud, and consequently the recorded audio has a low signal-to-noise ratio. To achieve a better signal-to-noise ratio, a narrow bandwidth microphone was used to minimise the scanner noise. The scanner noise was further reduced in the recorded audio by

Table 3.2: The utterances in the database and the number of subjects speaking them.

utterance	subjects	utterance	subjects
dance ordinance	9	answered	3
a far synod	8	answer enough	3
infancy fun	8	darn sunny	3
under done	8	enough soreness	3
undid a door	8	funded dinner	3
a disorder	7	funded order	3
endorse fun	7	in infancy	3
enforce	7	in ordinance	3
saucy dancer	7	needy suffer	3
unadorned	7	saucer	3
afforded	6	sinned inner	3
answered a door	6	sooner	3
cinder	6	under none	3
dancer four	6	under synod	3
enough fussiness	6	undone	3
sword sinner	6	unseen afar	3
under a door	6	awed fussiness	2
undid enough	6	duffer needed	2
dud ordinance	5	enough order	2
endorse force	5	enough suffered	2
enough dinner	5	incidence	2
enough sauce	5	indeed answered	2
indecency	5	sundered saucer	2
none sooner	5	unseen indeed	2
ordinance	5	unseen synod	2
enough ordered	5	adored enough	1
answer needed	4	adorn enough	1
asunder	4	answer none	1
enforce order	4	answered enough	1
fussiness	4	dancer	1
indeed adored	4	fun saucer	1
indeed undone	4	indeed differed	1
order sonny	4	inner soreness	1
sauce saucer	4	needy dinner	1
since infancy	4	none suffer	1
source sooner	4	undone dinner	1
answer fuss	3	unseen answer	1

signal processing techniques. The repetitions are more or less similar and share the same rhythm.

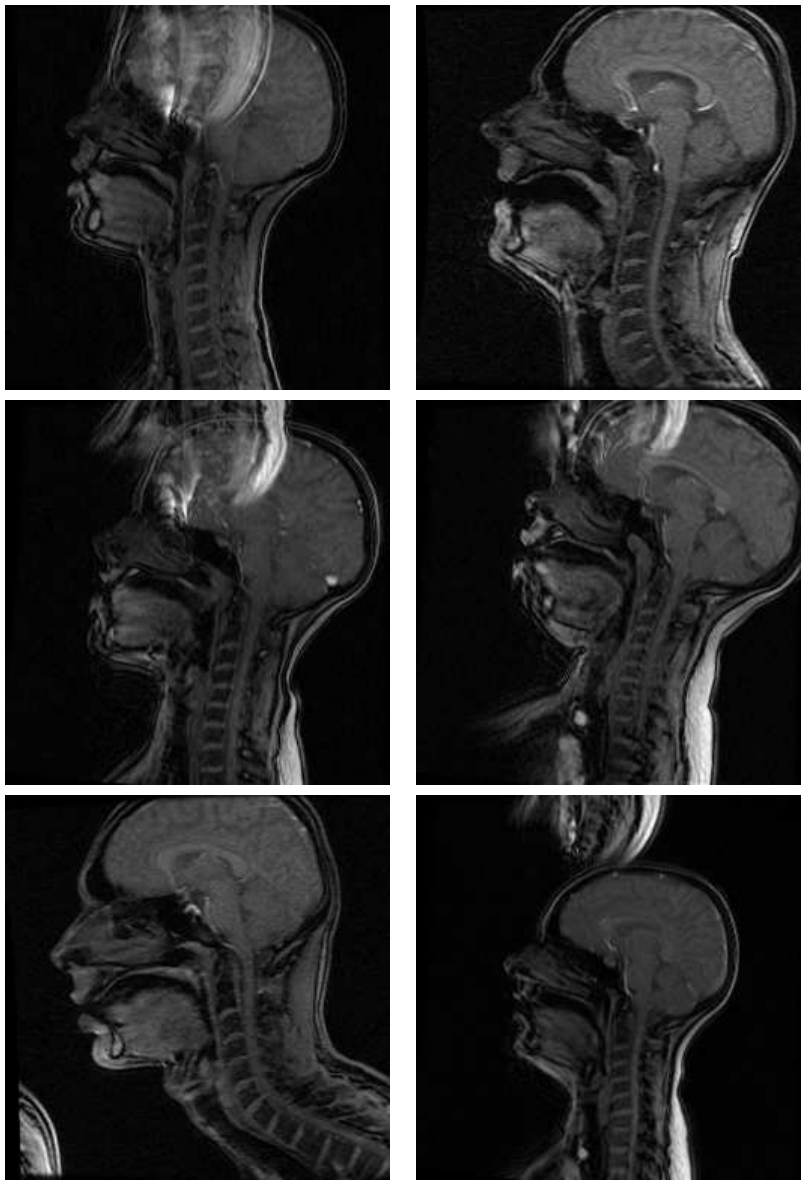


Figure 3.5: Arbitrarily chosen MR images of 6 different subjects articulating different phrases. From the Oxford University Phonetics Laboratory's MRI database.

3.5 Image enhancement and standardisation

A major challenge in MR image analysis is the inconsistency in what intensities represent. That is, the intensities of the same region in different images, even for the same person, do not imply the same meaning. In particular, in our MRI database, there is an extra dimension added by having multiple speakers. The inconsistency poses problems in image analysis and segmentation. Furthermore, many different image processing methods developed for large scale databases are based on a particular set of parameters that would lose their robustness if there are inconsistencies in what the intensities represent.

MRI standardisation [Nyúl and Udupa 1999] is a post-processing method widely applied in the field of biomedical image processing for achieving consistency in the intensity scales among a set of MR images. In this approach, the intensity histogram in all images is rescaled and transformed to a *standard* histogram that is calculated based on the properties of the images in the set. The standardisation of images is carried out in two steps. First, a number of tissue specific landmarks $LM_1, LM_2, ..., LM_N$ are chosen to represent the histogram properties, e.g. mean, median, deciles and so on (*training*). For every image frame, each of the landmarks are mapped to the pre-determined standard scale. Finally, the average location of each landmark LM_i^s on the standard scale is calculated, based on the location of the mapped landmarks. During the second step (*transformation*), the landmarks are once more identified in the image histogram. The landmarks are then mapped to the corresponding average landmark points on the standard scales. The intensities of each image are next mapped as per this transformation to the standard scale. Figure 3.6 shows the mapping function from the image's intensity scale to the standard scale.

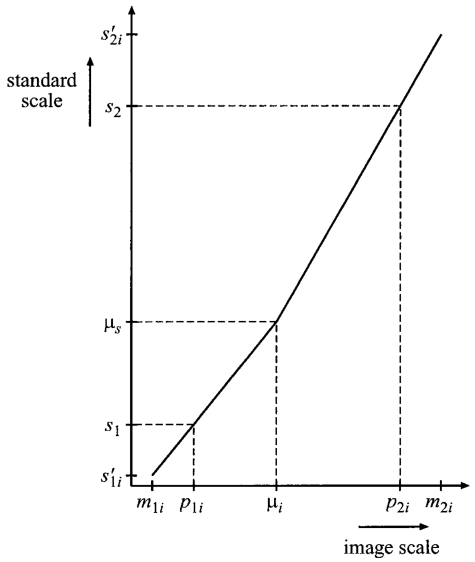
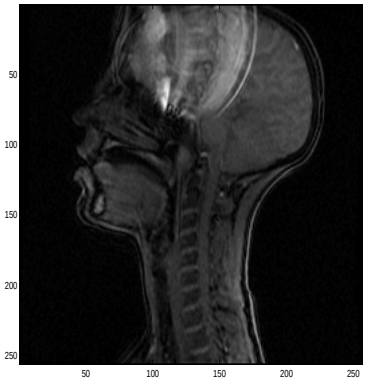
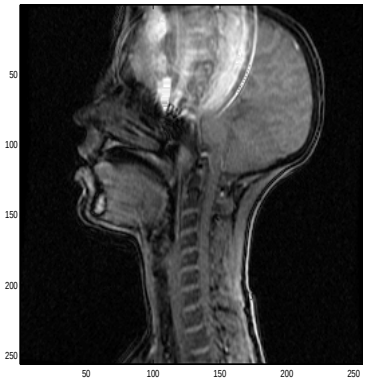


Figure 3.6: The transformation stage: intensity mapping function from [Nyúl and Udupa 1999]. Parameters p and s represent landmark points on the original image and the standard scale respectively, m_{1i} and m_{2i} values represent the minimum and maximum values of the histogram respectively, and μ_i and μ_s represent the mean on the original image histogram and standard histogram respectively. Values of s' are obtained after the mapping from minimum and maximum values on the histogram of the image to the standard scale.

In our implementation of the algorithm, the standard scales $[s1, s2]$ were set to [0, 400] respectively. The histogram landmark points were chosen to be the mode and percentiles of the histogram. A selection of original MR images and the standardised versions for a number of speakers are presented in Figures 3.7 and 3.8. A qualitative analysis of the images on the right hand column and left hand column confirms that, in general, the images in the right column are more similar in intensity profile compared to the images in the left column.



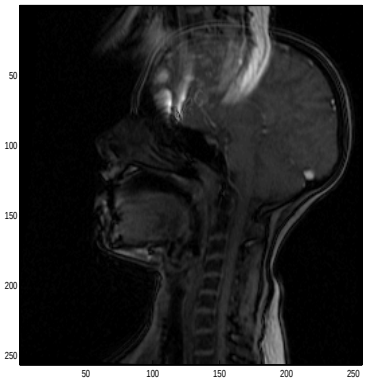
(a) original image (speaker A)



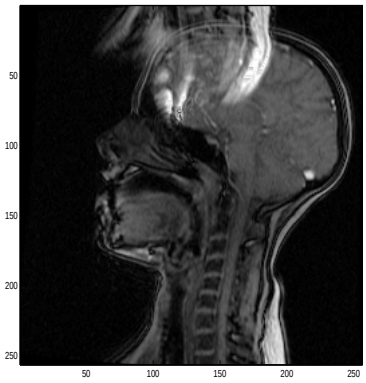
(b) standardised image (speaker A)

(c) original image (speaker C)

(d) standardised image (speaker C)

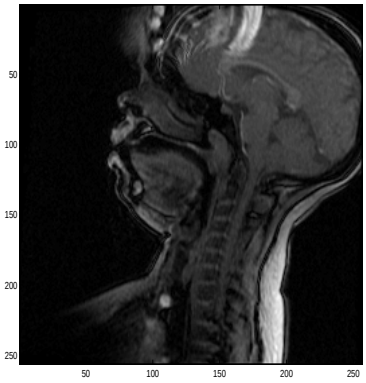


(e) original image (speaker L)

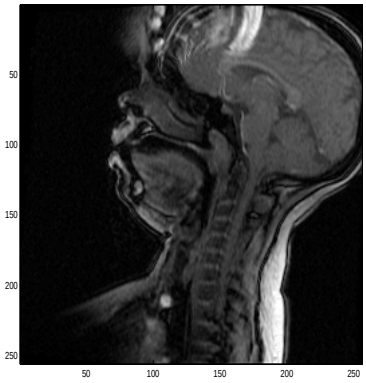


(f) standardised image (speaker L)

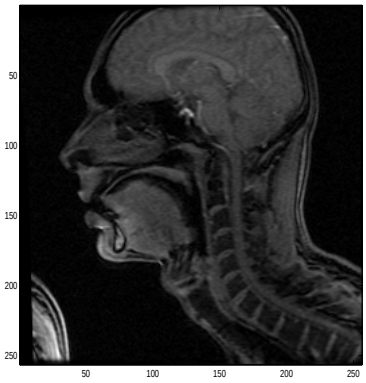
Figure 3.7: Original MR images (left) and the corresponding standardised images (right).



(a) original image (speaker M)

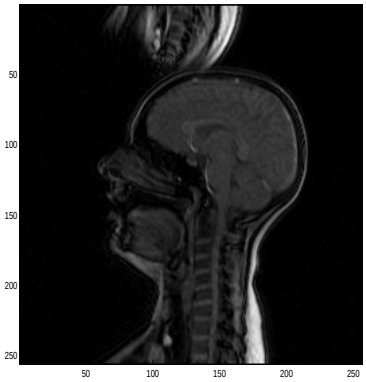


(b) standardised image (speaker M)

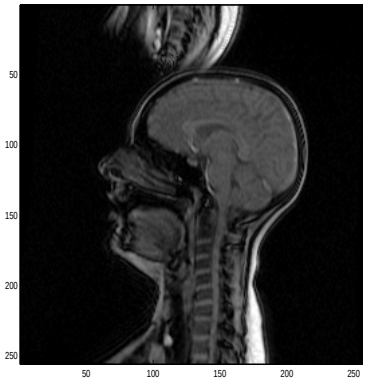


(c) original image (speaker P)

(d) standardised image (speaker P)



(e) original image (speaker R)



(f) standardised image (speaker R)

Figure 3.8: Original MR images (left) and the corresponding standardised images (right).

3.6 Pixel intensity variation and hidden information

The movement of the articulators is reflected in the images as changes in the intensity values of the pixels. The variation in the tissue densities results in receiving signals with variable strengths for each spatial point on the images, hence the intensity values of the spatial elements alter over time.

Figure 3.9 shows an example of tongue movement over a sequence of 66 images captured during articulation of the phrase "answer needed". To facilitate visually following the tongue movement in the images, the tongue contours were semi-automatically tracked using live wire approach [Hamarneh 2005] and superimposed on the image. The tongue contours in the images help define both the boundary between the tongue tissue and the air, and the shape and positioning of the tongue over time. A closer look at the images confirms that the intensity gradient is almost monotonic in the area of tongue body, below the contour, consisting of close shades that are also consistent across the image. On the contrary, the spatial attributes are not fixed as the tongue is constantly deforming to a different posture for a new articulatory setting.

The deformation in tongue shape can be observed in Figure 3.10 where the extracted tongue contours from the sequence are plotted in one figure. Therefore, the intensity of individual pixels in the vocal tract airway is constantly changing relevant to the new position of the tongue. The intensity variation of the pixels over time (across 68 images) within the area defined in Figure 3.11 is illustrated in Figure 3.12. We chose this area for demonstration particularly because it corresponds to the alveolar ridge place of articulation, where the tongue tip and front is expected to approach a few times during the phrase "answer needed".

As can be noticed in the sequence of sub-frames, the individual pixels' intensities vary considerably over the range of 68 images. Looking at the trend of these variations can provide us with useful information regarding the articulation and articulators' re-positionings occurring during the articulation. Intuitively, the pixels with more stable intensity values across the range of images are the areas where less

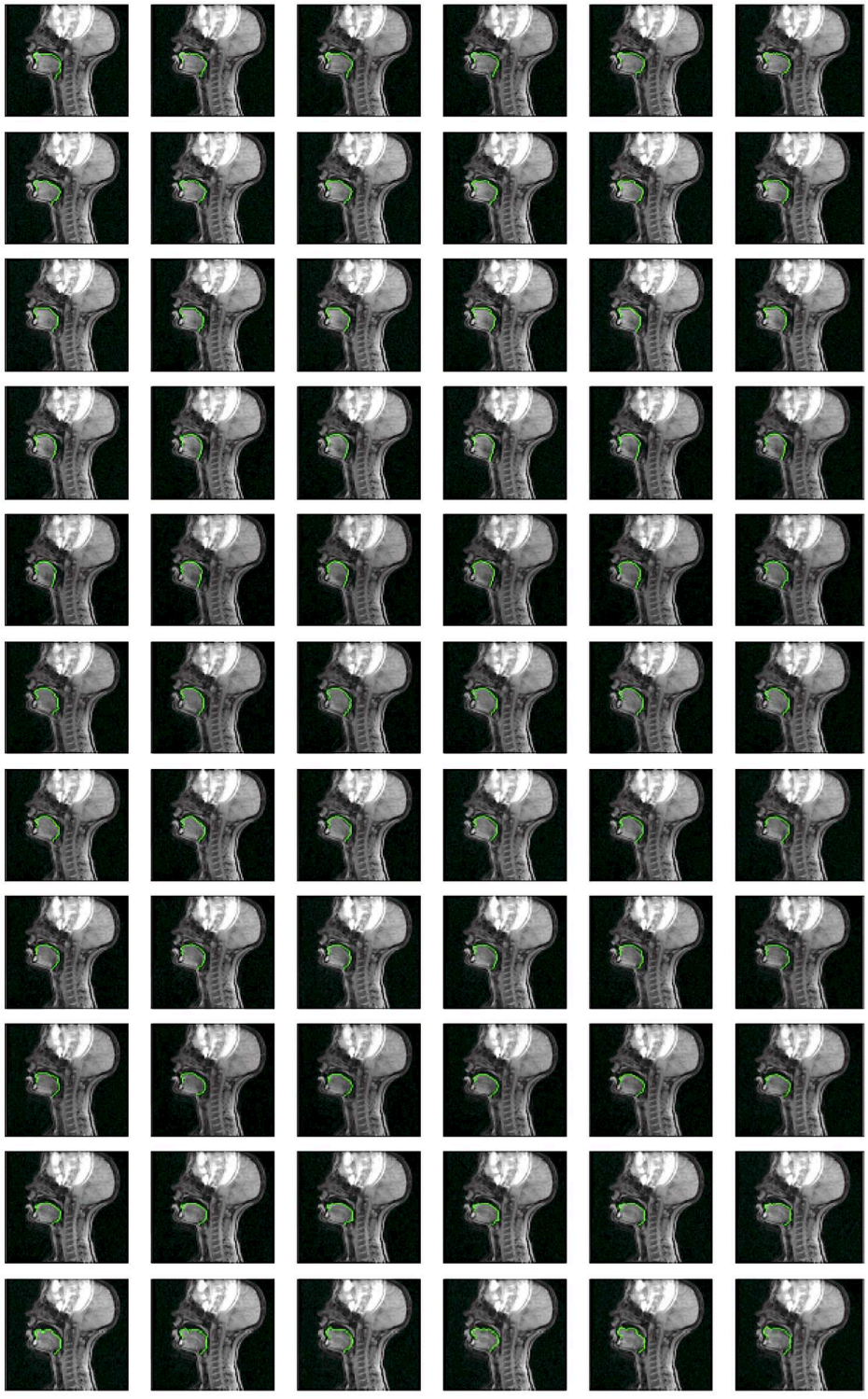


Figure 3.9: 66 successive images during an articulation of “answer needed”. Tongue positions are highlighted in green.

movement took place. On the other hand, where the intensity variation trend shows more dramatic changes, greater articulator movements are anticipated. This information can be applied to study the areas of stability or variation along the vocal tract in sequences of images.

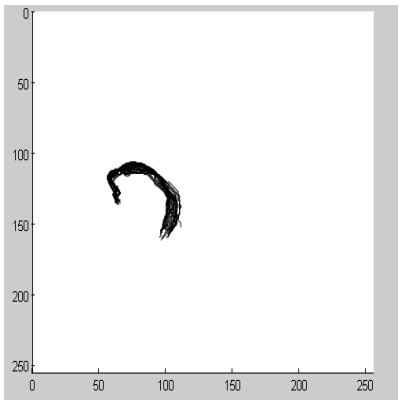


Figure 3.10: The extracted tongue contours from the 66 images in an overlaid plot.

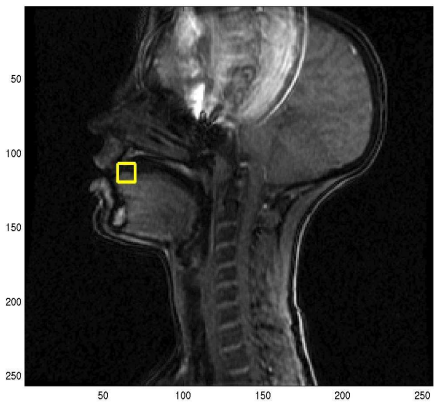


Figure 3.11: The region selected for studying pixel intensity variation.

3.7 Conclusion

In this chapter, we reviewed the basics of using magnetic resonance imaging technology for studying articulation. MRI is safe and non-invasive, and therefore is a popular and effective technology for collecting visual information about articula-

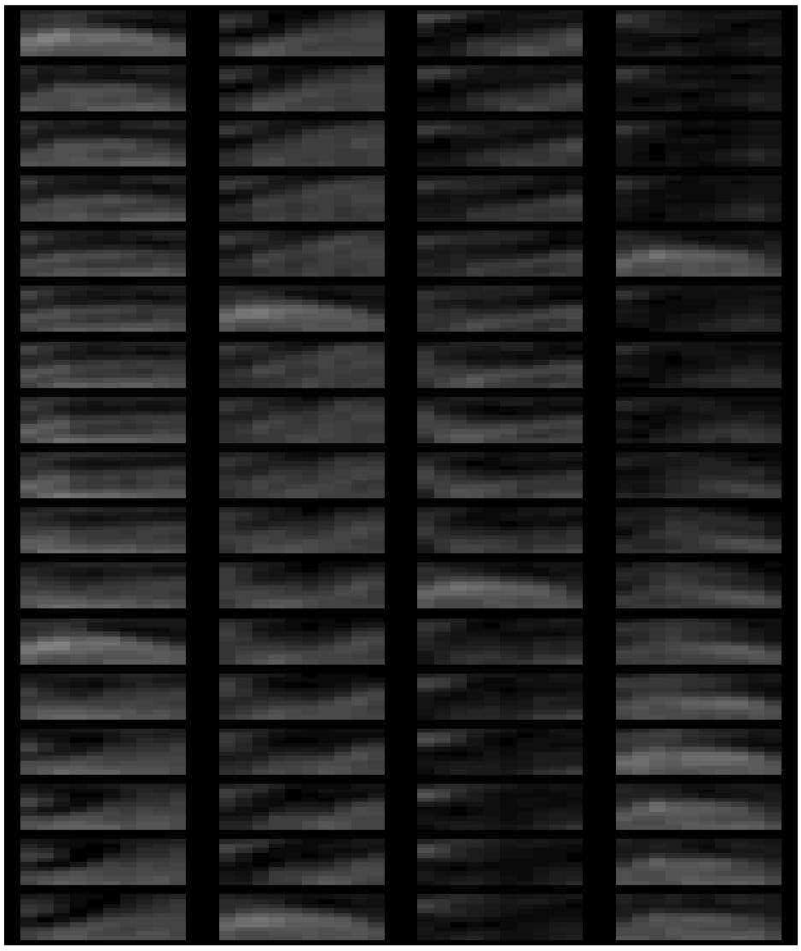


Figure 3.12: The pixel intensity variation in a region of 10×10 pixels beneath the alveolar ridge, in a sequence of 68 images corresponding to an articulation of "answer needed". The image sequence starts at the top left corner of the box. The direction of the sequence is down and then right.

tion. Depending on the MR imaging protocol used in capturing the articulation process, various approaches have been proposed to extract the useful information from images of vocal tract. We reviewed some of these approaches in section 3.3.

For larger databases, such as Oxford University Phonetics Laboratory's MRI database, methods for obtaining useful information from the images automatically

would be essential. The pixel intensity variation of successive images reveals significant information about the movements of articulators over time. In the next chapter, we utilise the pixels' grey-levels and their variation over time to study phoneme articulation in MR images of dynamic speech.

Chapter 4

Correlating Speech Acoustics and Articulation

4.1 Introduction

Study of the relationship between speech acoustics and articulation is not possible without specifying the correspondence between acoustic and articulation units. In this thesis, phoneme-sized units (phones) are chosen as the discrete speech units to be studied. Matching acoustic and articulation units of the MRI database described in the previous chapter involves an initial segmentation of the audio into appropriate phoneme-sized units, followed by locating the MR images corresponding to the articulation of individual phones (per speaker and per utterance).

The smallest meaningful unit that can be specified in the visual component of the database is a single MR image. However, due to the nature of a dynamic MRI database, a sequence of images (rather than a single image) are aligned with each acoustic unit. An important question therefore is how to summarise multiple visual units into a single unit that symbolises an individual phoneme. Note that multiplicity of visual units for each phone is a positive attribute of the dynamic MRI database; since details of phoneme articulation as the articulators move towards the target, reach the target and move away from the target are recorded. In this chapter,

we investigate the possibility of summarising the visual MRI data without losing the useful articulation details of a sound.

Moreover, the acoustic-visual pairs obtained are phrase-based and speaker-based, making observation of a general pattern of correlation between the acoustic and visual units challenging. A method for information extraction is devised to normalise the data and compensate for inter-speaker and intra-speaker variability. Finally, to investigate the correlation, we look at the acoustic and articulator dynamics, and investigate the effects of the vocal tract movements on the generated acoustics.

Section 4.2 first details the alignment technique used in this work for determining the corresponding acoustic and articulatory units. In section 4.3, the alignment results are used for finding "typical" vocal tract shapes for different phonemes for different speakers, and the results are analysed to examine the validity of the approach. Finally, in section 4.4 a new technique is described that parametrises the correlation between the dynamics of acoustics and articulation.

4.2 Alignment of the data components

The MRI database introduced in chapter 3 comprises three major components: the noise-reduced recorded audio, the acquired images and the transcriptions. For each sequence of images, there is a corresponding audio signal and a transcription (label) file indicating the spoken phrase. However, the trio of image sequence, audio signal and transcription were not originally aligned together in time. It is therefore not practical to specify the image frames an audio segment maps to, the speech unit it represents, and where the unit boundaries stand in both audio and visual data. Figure 4.1 illustrates a schematic representation of the alignment of audio signal, images and transcription together.

The alignment of audio and transcriptions could be achieved by manual human-labelling of the sequences, where a professional listens to the audio and follows the transcription simultaneously and decides where to place the boundary between two consecutive acoustic units. Similarly, for images and audio the same approach can

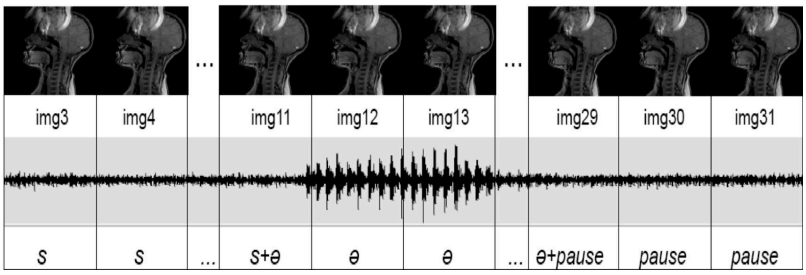


Figure 4.1: Schematic view of an alignment between the images, audio and the transcriptions for the word "(sau)cer".

be pursued, where a professional listens to the audio and tracks the articulation images visually, and based on position of articulators determines the approximate boundaries. Nevertheless, for visual data this method would be very error prone and difficult due to the inevitable co-articulatory effects in running speech. The manual method is not feasible for our MRI database due to the volume of data, and an automatic or semi-automatic alignment method is needed. Moreover, simultaneous alignment of the audio, images and transcriptions is a complex task owing to the number of parameters involved. The majority of these parameters, such as multiplicity of acoustic units for each transcription unit or the arbitrary starting frame of the image sequence, are introduced as a result of the acquisition technique adopted for collecting the data. An alternative approach to avoid these complications would be to adjust two of the three components (e.g. audio and transcriptions) first and use the outcome to align the former two with the third component (images) temporally. This latter approach is the technique employed here.

In this work, first the conventional forced alignment technique is used to coordinate the acoustics and transcriptions. The resulted timing and duration information is subsequently used to align the sequence of images with the acoustics and transcriptions.

4.2.1 Forced alignment of audio data to image sequences

Forced alignment is a special case of speech recognition. In automatic speech recognition (ASR), a system is typically designed to receive a speech audio signal and convert it to text. An ASR system is provided with a training set of possible transcription units such as words for the speech data, and during the recognition it searches through all possible word sequences to find the most likely match between the input speech data and the text units.

In forced alignment, the speech decoder engine is provided with the *exact* transcription of the spoken words, instead of a set of words. The decoder in this mode adjusts the speech signal and transcription in time, i.e. identifies the correspondence between the segments of speech signal and the speech units, and determines the timing and boundaries of each unit in the signal (beginning and ending points of each speech unit). In other words, the decoder time-stamps the transcriptions by matching the acoustic features present in the audio signal to the various learned speech unit features. Figure 4.2 illustrates the difference between an ASR system in recognition and alignment modes.

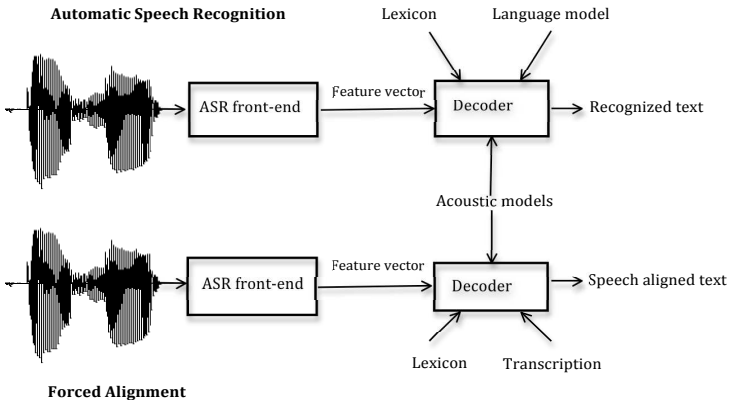


Figure 4.2: Diagram showing consecutive stages of an ASR system in recognition (top) and forced alignment (bottom) modes. In the forced alignment mode, the decoder is provided with the exact transcription of the signal instead of a set of the possible words and the language model.

A primary step in building an ASR system is to define and train a set of acoustic models corresponding to the speech units. In the next section, the training mechanism and the decoding algorithm utilised in this work are explained.

Hidden Markov Models

A hidden Markov model (HMM) is a machine learning technique that is used for determining the label or class that a unit in a sequence of units belongs to. In speech recognition applications, the sequence of units is an input speech signal and the classes can be sentences, words, morphemes, letters and so on. HMMs are probabilistic models that classify a sequence of speech units by finding the most likely match of audio frames to the possible labels.

The origin of HMM comes from Markov chain: a weighted automaton in which an input sequence can uniquely determine the sequence of states that the automaton will go through [Jurafsky and Martin 2008]. The main factor differentiating between a Markov chain and a hidden Markov model is that unlike Markov chain, in HMM a sequence of input symbols cannot uniquely predict the succeeding state sequence, i.e. the states are unobserved or *hidden*. When applying HMMs to speech recognition, the input speech is what can be observed and the underlying labels that it represents are the hidden parts that must be inferred. The majority of modern speech recognisers are developed on the basis of hidden Markov models.

An HMM comprises the following component (presented in Figure 4.3)

- A set of states $S = \{s_1, s_2, .., s_N\}$
- Transition probabilities $A = \{\mathrm{a}_{ij} | i, j = 1, ..., N\}$ probability of going from state i to state j
- Observation vector $O = \{o_1, o_2, ..., o_M\}$
- Observation likelihood $B = \{\mathrm{b}_1, \mathrm{b}_2, ..., \mathrm{b}_N\}$
- Start and end states s_0 and s_F respectively

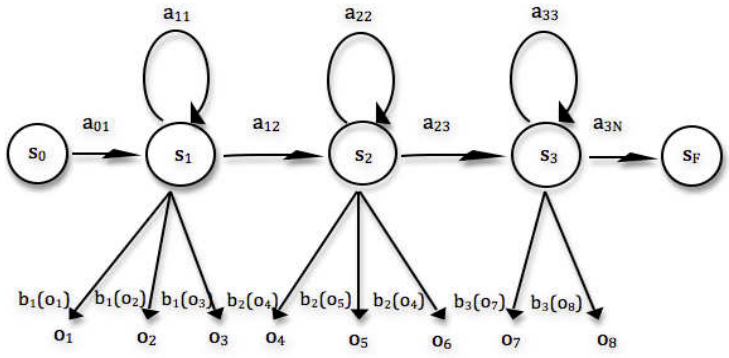


Figure 4.3: A three-state left-to-right HMM topology. The circles represent the different states of the HMM, a_{ij} is the transition probability from state i to state j, $O = o_1, ..., o_n$ is the observation sequence, and $b_i(o_k)$ is the probability density of state i emitting observation vector o_k.

Two assumptions make the first-order hidden Markov models suitable for tasks such as speech recognition. First the probability of each state in a Markov chain only depends on the previous state, rather than all the previous states. Second, the observation probability for each observation only depends on the state that produced the observation, not the previous or following states, or other observations.

An HMM λ is notated with its sets of parameters A (transition probabilities) and B (observation probabilities), $\lambda = \{A, B\}$. Developing an HMM involves different tasks such as training, likelihood computation and decoding.

Training. This is a fundamental step in developing an HMM model. The HMM parameters such as the transition probabilities and the observation probabilities are learned given a sequence of observations O and the set of possible HMM states S. The most popular algorithm adopted for training HMM transition and observation probabilities is the Baum-Welch algorithm [Baum 1972].

Likelihood Computation. This task consists of calculating the probability of a particular observation given a defined HMM, i.e. given an HMM λ and its parameters $\lambda = \{A, B\}$, and an observation sequence O, the likelihood $P(O|\lambda)$ is computed.

Typically, in calculating the likelihood, a dynamic programming approach such as forward algorithm is used, in which the observation probability is calculated by summing up the probabilities from all the hidden state paths that could generate the observation sequence.

Decoding. In the decoding step, the underlying variables that led to generation of a particular observation are determined. In more detail, during decoding, the most likely set of states that may have generated a particular observation are estimated. The hidden set of states are discovered by using the information from the observation vector O and the HMM parameters $\lambda = \{A, B\}$. The most conventional algorithm for decoding is the *Viterbi* algorithm. The Viterbi algorithm works exactly like the forward algorithm except that instead of summing the probabilities along the previous paths, it merely selects the maximum value.

In this research, the Hidden Markov Models Toolkit (HTK) [Young *et al.* 2006, accessed April 2010] was utilised that by default uses the Baum-Welch algorithm for learning, the forward algorithm for calculating the likelihoods, and the Viterbi algorithm for decoding. A typical set of HMM acoustic models can be developed using HTK through the following steps:

- Converting the transcriptions from word/sentence level to phoneme level (HTK HLEd function).
- Defining the HMM protocols for the acoustic models
- Initialisation of the acoustic models, i.e. generating one model for each present phoneme (HTK HCompV function)
- Performing embedded re-estimation on the HMM parameters (four rounds of training with HTK HERest function)
- Forced aligning the transcriptions with the acoustics (HTK HVite function in alignment mode)

- Improving the system parameters, e.g. increasing the number of mixtures, tying the models, adapting the models etc.

- Performing embedded re-estimation with the new introduced parameters (multiple iterations of HVite and HERest functions)

When the training is complete, the acoustic models are ready to be used for decoding (recognition) or forced alignment of the transcriptions. During the recognition, the system is provided with a lexicon, a language model and a set of unobserved test acoustics. The speech is converted to text and the generated text is compared with the original transcriptions, and the performance is evaluated based on the number of errors in the generated text such as deletions, insertions and substitutes. In the forced alignment mode, the system is provided with a set of unobserved speech signals and their exact transcriptions, and generates a new set of transcriptions for the audio with boundary and time information. The new transcriptions can be evaluated either with automatic measurements based on the statistics from the signal and the labels [Baghai-Ravary *et al.* 2011], or can be compared with a human labelled gold-standard.

Implementation of forced alignment

Two sets of data were used to train the ASR system. One was the MRI database acoustics introduced in chapter 3. However, due to the high level of noise inside the scanner, the collected speech data is relatively poor in quality even after the noise cancellation. In addition, neither the quantity nor the diversity of the acoustic MRI data were sufficient to train an HMM-based alignment system. Therefore, the models were trained on a second separate clean corpus, in addition to the speech from the MRI database. The clean corpus contains speech recordings collected for training British English acoustic models in a previous project [Loukina *et al.* 2009]. The MRI audio recordings were also included to provide data that is more representative of the characteristics of the target signal.

The dictionary used was Carnegie-Mellon University Pronouncing Dictionary (CMUdict)[1], extended with some out of vocabulary (OOV) words from the British National Corpus (BNC) database[2] in addition to British pronunciations of the words that were added by semi-automatic methods and evaluated by experienced phoneticians[3]. All the label files were converted to phoneme level labels using the lexicon from the dictionary.

In total, 75 models were initialised and trained, 72 phones plus silence (sil), short pause (sp) and noise silence (ns) models, including phones that might have been only available in the clean corpus and not the MRI database. The noise model was introduced and trained later when the models were re-trained with the MRI database acoustics. The reason was that the pauses and silences in the MRI audio were mostly the MRI scanner noise recorded with the audio, rather than a human silence or pause. The list of English phonemes in IPA and in the Arpabet dictionary notation is presented in Table 4.1 and their presence in the MRI data is marked with a dagger †.

Evaluation

To evaluate the performance of the aligner with different HMM settings, the results were compared with gold-standards. A set of 14 gold-standards were prepared by a professional phonetician from the data held out for testing purposes. The system was then provided with the original transcriptions and signals of these test utterances and automatic alignment results were compared with the manually-aligned gold-standards. The comparison of the obtained labels from different HMMs structures, with different number of Gaussian mixtures and number of states, and the gold-standards are presented in Table 4.2. The best alignment results are achieved from the HMMs with 5 states and 48 Gaussian mixtures, where 93% of the automatic labels agreed to within 50 ms of the manual ones; (87% agreed to better than

[1][CMU accessed November 2011].
[2][BNC accessed November 2011].
[3][Coleman *et al.* 2011, accessed March 2012].

Table 4.1: The list of English phones in the clean corpus in Arpabet and IPA notation. The numbers following the letters in Arpabet notation correspond to the stress on the phone: 0 means unstressed, 1 means primary stress, 2 means secondary stress. The † sign indicates the phones' presence in the MRI database.

AA0|ɑ	AA1|ˈɑ†	AA2|ˌɑ	AE0|æ	AE1|ˈæ	AE2|ˌæ
AH0| ə†	AH1|ˈʌ†	AH2|ʌ	AO0|ɔ	AO1|ˈɔ†	AO2|ˌɔ
AW0|aʊ	AW1|ˈaʊ	AW2|ˌaʊ	AY0|ɑɪ	AY1|ˈɑɪ	AY2|ˌɑɪ
EH0|ɛ†	EH1|ˈɛ	EH2|ˌɛ	ER0|ɜ	ER1|ˈɜ	ER2|ˌɜ
EY0|eɪ	EY1|ˈeɪ	EY2|ˌeɪ	IH0|ɪ†	IH1|ˈɪ†	IH2|ˌɪ
IY0|i	IY1|ˈiː†	IY2|ˌiː	OW0|əʊ	OW1|ˈəʊ	OW2|ˌəʊ
OY0|ɔɪ	OY1|ˈɔɪ	OY2|ˌɔɪ	UH0|ʊ	UH1|ˈʊ	UH2|ˌʊ
UW0|u	UW1|ˈuː†	UW2|ˌuː	OH1|ˈɒ	OH2|ˌɒ	B|b
CH|ʧ	D|d†	DH| ð	F|f†	G|g	HH|wh
JH|ʤ	K|k	L|l	M|m	N|n†	NG|ŋ
P|p	R|ɹ†	S|s†	SH|ʃ	T|t	TH|θ
V|v	W|w	Y|j	Z|z	ZH|ʒ	

30 ms). Therefore, we used the labels achieved from HMM models with 5 states and 48 mixtures for further steps of this research, i.e. the alignment of the audio, transcriptions and images explained in the following sections.

4.2.2 Alignment of images, audio and transcriptions

Each sequence of images, despite being acquired over multiple repetitions, only depicts one articulation instance of the phrase, and thus needs to be interpreted carefully.

To start with the image alignment, sub-sequences of images from the whole sequence are mapped to the segments of audio signal. In other words, windows of sequential frames representing the articulation of each phone are specified. The number of sequential images, i.e. window length, is calculated from the duration

Table 4.2: Evaluation of forced alignment results. The numbers represent the percentage of automatic labels that agreed to within 30 milliseconds or 50 milliseconds of the manual labels.

HMM structure	within 30 ms (%)	within 50 ms (%)
MFCC & 32 GMMs & 3-state	61.72	68.94
MFCC & 48 GMMs & 3-state	65.93	73.83
MFCC & 32 GMMs & 5-state	87.53	92.76
MFCC & 48 GMMs & 5-state	87.27	92.96

information of the transcription-aligned audio data

$$winlen(s_p) = K \frac{dur(s_p)}{\sum_{i=1}^{P} dur(s_i)}, \tag{4.1}$$

where P is the number of phones in a phrase, s_p is the target phone instance, $dur(s_i)$ is the duration of the phone s_i, and K is the number of frames in the phrase movie. The transcription-aligned speech signals are in fact the audio data recorded over the entire duration of image acquisition, consisting of 18–20 repetitions of the same phrase. Consequently, there are multiple instances of the phone repeated over the entire utterance. To achieve a single value for each phone duration, the average duration of different occurrences of the phone in the utterance is considered. Note that if a phone occurs in two different positions in one phrase, the durations at each position are calculated independently. A statistical analysis carried out on the duration variance of all of the phones calculated with the above technique found that standard deviation from the mean duration for each instance is on average roughly 16 milliseconds.

In addition, to check the validity of this approach further, the median and standard deviation of phone durations were calculated for some of the utterances. The results suggested that maximum deviation from median duration for each phone in each utterance is mostly on the order of milliseconds. The boxplot in Figure 4.4 illustrates the results for a randomly selected utterance, “enforce”, spoken by speaker L (just one utterance selected for illustration purposes). In this example, the maxi-

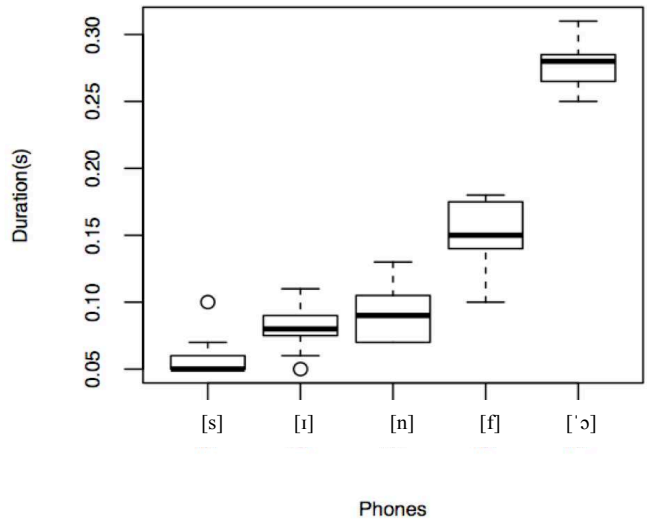


Figure 4.4: Variation of phone durations in 15 repetitions of "enforce" in a single utterance by a single speaker. The transcription of "enforce" has been reordered to display the durations in ascending order.

mum deviation from the median in almost 50 ms (in /f/).

For each speaker, all the obtained image sub-sequences for a particular phone across different utterances are combined into one set of articulation frames for that phone. An alternative approach would be to select only the middle frame in the image window of each phone in each utterance, to be added to the set of articulation frames of that phone. However, if only the middle frame is considered rather than the entire sequence of images in the segment, the details of the articulatory movements distributed across the sequence of frames in each phone's set are simply ignored.

In the next section, we observe how these generated sets for each acoustic unit are used to find an average image that represents the "typical" articulation of a phone (unit). The speaker-specific typical articulation of a phone is instance-independent, but all the surrounding contexts occurring in different instances (utterances) of the phone are taken into account when producing the typical image.

4.3 Speaker-specific typical vocal tract shapes

A "typical articulation" of a unit should exhibit the articulatory characteristics of that unit in different contexts and within different rhythmic patterns. That is, representation of an articulation would only be typical if it were context-independent, while representing context-dependent variations in the unit's articulation. The articulation representation therefore must generalise over all the attested articulatory configurations in producing that particular unit, with the more common configurations being more prominent than the less common configurations. In other words, if a probabilistic model were to be created to model the typical articulation of a unit, higher probabilities must be assigned to its more common articulatory configurations and vice versa.

In this section, a new data-oriented approach is introduced for obtaining the typical articulation of phonemes, using the image pixel intensities. The results depict the more prominent articulation together with the less probable but possible variations for each phoneme. The probability variation is encoded in the image by the brightness of different pixels. Note that there is no precise boundary between the more common and less common articulatory configurations, and the scale of variation has a fuzzy nature.

To obtain a typical image of each phoneme, the intensity properties of the digital image are used. The vocal tract shape is visible in the negative images in Figure 4.5–4.10 as a white cavity shaped by the contrast of colours in the boundaries of the vocal tract tissues and the air. An image representing a phoneme articulation for each speaker is calculated through pixel-by-pixel averaging over all the images corresponding to that particular phoneme, obtained during the image alignment process described in section 4.2.2. The resulting image depicts the average position of articulators for a particular phoneme. Each tissue (or other material) in the average image is represented by different shades of grey. As previously explained in chapter 3, this is due to the fact that position of the articulators in one image may overlap with the cavity (white) area in another image as a result of movement, and a brighter shade

of gray, often appearing as a blurring around the edges, results. In general, the white cavity area bounded by the grey colour of the edges of the mouth and tongue tissues can be regarded as a speaker's typical vocal tract shape when articulating a particular phoneme.

It is worth mentioning that using median instead of mean would be applicable for obtaining more prominent positions. However, by using median, the motions that are caused by the coarticulatory effects would most likely be ignored as *outliers*. Thus, the blurring around the edges that reflects the coarticulatory variation in running speech would be missed. Therefore, we decided to use the average position to obtain a typical articulation that depicts both the common configurations and the contextual effects on the common configuration.

Figures 4.5–4.10 (c)–(h) and (i)–(l) show average images of respectively 6 vowels and 4 consonants articulated by different subjects. The images are grouped based on the speaker rather than phone articulation, to facilitate the observation of variation in different phones' articulations by avoiding the complexities introduced by inter-speaker variability. Images (a) and (b) in each set of figures are negatives of two arbitrarily chosen single original images, from a sequence that corresponds to the articulation of phoneme /ɔ/ by each speaker. Images (c)–(l) in each figure are averages of a collection of images representing the articulation of a particular phoneme in different context for each individual speaker. Despite the fact that vocal tract shapes and manner of articulation vary from one individual to another, a general pattern can be observed in the average phoneme images, which is more or less in agreement with the expected vocal tract settings for each sound suggested in previous work based on X-ray studies of articulation [Fant 1960; Perkell 1969], as explained in detail in the following paragraphs.

Image (c) in each figure presents the mean articulatory configuration for the mid-central vowel /ə/. In spite of the blurring of the edges of tongue, a relatively uniform cross-sectional width in both upper and lower airways is observed in the image (c) of all speakers, consistent with previous observations of /ə/ by [Fant 1960,

p. 56 and 66]. Note that the blurring around the edges of the tongue has a roughly consistent width from the tongue tip towards the back of the tongue in most of the images, which suggests that the uniform tube shape for this phoneme is preserved in different contexts. There is a moderate (not too wide, not too narrow) gap between the lips creating a mouth opening for the articulation of the vowel. The minimal distortion in the images of /ə/ justifies its potential use as a "neutral" reference image in discussing the movement-related distortion in other images.

A robust pattern is observed in all the images Figure 4.5 (d) to Figure 4.9 (d) (but not Figure 4.10 (d) which represents phoneme /u/ of speaker C). This common setting consists of the high position of tongue body and its closeness to the palate in the pre-palatal or post-alveolar region. Comparison of image (d) with images (i), (k) and (l) shows that the narrow constriction in /i/ is quite different from that of /s/, /d/ and /n/. The vocal tract shape in these images has a horn-shape with narrow opening in the upper airway and a wide gap in the pharynx, a description that is in agreement with the anticipated vocal tract configuration for /i/ [Fant 1960, p. 114–116] and [Perkell 1969, p. 17–23]. The image in Figure 4.10 (d) is the average image of vowel /u/.[4] The backness of the tongue body and root is consistent with the anticipated articulatory configurations of /u/. A degree of similarity is in fact noticeable between the articulation of /ɔ/, pictured in image (h), and the articulation of /u/. However, a narrower lip opening is expected in articulation of /u/ [Perkell 1969, p. 17 and 43].

The vocal tract settings of the phones [ɪ] and [ˈɪ], images (e) and (f) in Figure 4.5 – Figure 4.10, suggest a fairly narrow constriction visible in the upper airway, with a wide aperture in the pharynx, which is in agreement with previous observations [Perkell 1969, p. 17–23]. The tongue body in all of the images is relatively front. However, the degree of front cavity occupation of the tongue is variable in different speakers, and a darker shade at the top of the tongue body is seen in the images of speakers A, P, M and C (Figures 4.6, 4.8, 4.9 and 4.10), compared to

[4]There is no data available for articulation of vowel /i/ for speaker C.

images (e) and (f) of the other two speakers (Figures 4.5 and 4.7). Nonetheless, the gap in the upper airway is consistently wider than the average articulation image of /i/, which is expected knowing that the height of the tongue for /ɪ/ has been found previously to be lower than /i/ and higher than /ə/ according to the International Phonetic Alphabet (IPA) chart.

For the open vowel /ɑ/ (Figure 4.5(g) – Figure 4.10(g)), the trends were opposite to what was observed for /i/ and /ɪ/ is observed with a relatively wide gap between the tongue body and palate in the front, and a narrower airway in the pharynx, creating a horn-shape with the wider end located at the front of the cavity, also suggested before in previous observations of this phoneme [Fant 1960, p. 114–116] and [Perkell 1969, p. 17–23].

The articulation structure in the average image of vowel /ɔ/ (images (h) of each figure) shows a constriction at the back of the throat between the tongue back and the soft palate. In spite of the blurring at the front seen in some of the images, such as images 4.6 (h) and 4.8 (h), the area of darker shades at the front of the tongue in all the images consistently takes a lower position compared to the back of the tongue.[5] The gap in the upper airway is at all times wider in comparison to the upper airways in images (c)-(e) across all of the figures. Another distinguishing articulatory feature observed in images (h), representing phoneme /ɔ/, for all speakers is the lip rounding that leaves a smaller gap at the lip opening of the vocal cavity relative to the mouth openings of the other vowels.

The canonical articulatory features observed for different vowels can be used to estimate the variations in the expected acoustic features of different vowels. We consider the uniform tube shape of /ə/, "neutral" tube shape, as a reference, and discuss the acoustic feature variations of different vowels from acoustic features of /ə/. For the vowels /i/ and /ɪ/, the decrease in the cross-sectional area at the front

[5]Note that in the above descriptions, the positions are described relative to the orientation of the speakers' heads instead of the absolute positions. That is, when the speaker's head is tilted in one direction, instead of the typical spatial x-y coordinate system, we must consider a tilted coordinate system, with x-axis being parallel to the palate and y-axis as a perpendicular line to the x-axis.

of the oral cavity, leads to a decrease in the value of first formant (F1), while the increase in the cross-sectional area at the back of the tube, increases the value of second formant (F2) [Mrayati *et al.* 1988]. The relatively bigger cross-sectional area at the front of the oral cavity in [ɪ] and [ˈɪ] compared to /i/ suggests slightly lower F2 and higher F1 values from the former two in comparison with /i/. In contrast, the large cavity in the front of the tube suggests a high first formant for /ɑ/. However, the value of the second formant falls accordingly as the cross-sectional area in the pharynx decreases. The expected acoustic features from the average articulatory configuration of /ɔ/ are a higher F1 linked to the large volume in front of the cavity and a lower F2 attributed to the constriction at the pharynx [Mrayati *et al.* 1988]. The narrowing at the lips observed in the average image of /ɔ/, has a dampening effect on the values of both formants, in comparison to all the other unrounded vowels discussed before.

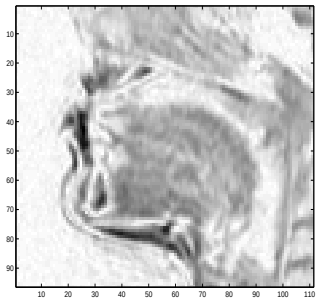
The configurations observed in the average images of consonants (Figure 4.5 – Figure 4.10 (i)–(l)) also agree with prior X-ray based observations [Fant 1960; Perkell 1969]. Where relevant, an ellipse is drawn around the area where the paramount and distinguishing feature of each sound's articulation is present. In images (i) and (k) of each speaker's figure, that respectively correspond to the average articulations of phonemes /s/ and /d/, a constriction of the tongue tip towards the alveolar ridge in front of the cavity is evidently visible. Despite the fact that no dramatic difference can be observed in the place of constriction and the articulators involved, in most of the images (i) of phoneme /s/, the tongue tip is darker in intensity compared to images (k) of phoneme /d/ for the same speaker. The darker shade illustrates a more robust (less dynamic) presence in a location over the images averaged. The difference of shades for /s/ and /d/ can be attributed to the fact that /s/ is generally a longer sound in duration than /d/, and consequently the tongue tip has been constricted to the alveolar ridge for a longer period of time.

The average image (j), corresponding to the articulation of phoneme /f/ in each figure, can be distinguished from the other consonant images in each set by focusing

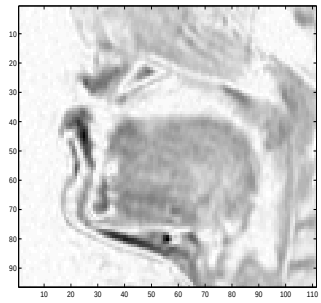
on the lower position of the tongue tip and more importantly the small gap between the lips. In fact in some of the images such as Figure 4.5 (j) and Figure 4.8 (j) hardly any gap can be discerned between the lips. The insignificant opening between the lips is a result of constriction of the lower lip to the edge of the upper teeth. Note that since calcified structures cannot be captured with MRI, the teeth are not visible in the images. The width of the opening at the lips in image (j) in each figure is more similar to the width of mouth opening in image (h) representing articulation of vowel /ɔ/.

Finally, image (k) in each figure representing the average articulatory configuration of phoneme /n/ is essentially different from the other consonants' average images (i)–(j) in the velar opening to the nasal cavity [Perkell 1969, p. 27 and 54]. The velum in /n/ is slightly lowered in most of the images (k) in each set of figures, and a narrow gap is apparent between the velum and the pharynx. The narrow nasal airway at the velic aperture is more visible in some speakers' figures such as speakers L, P, M and C (Figures 4.5, 4.8, 4.9 and 4.10) relative to the image (k) of speakers A and R (Figures 4.6 and 4.7), but is almost consistently visible across the images. The tongue tip closure at the alveolar ridge is similar to that observed for images of /d/.

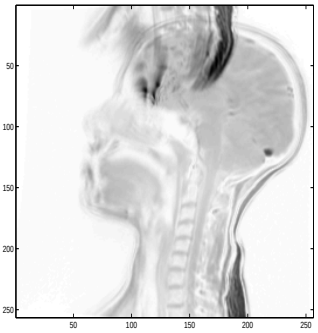
Figure 4.5: (a) and (b) Two arbitrarily chosen, magnified, original MRI negatives from the window of articulation images of /ɔ/, female speaker L. (c)–(l) Average articulation images. The ellipses show the areas where the paramount and distinguishing features of each consonant's articulation are present.



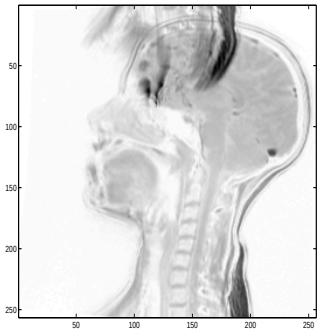
(a) an original negative instance of /ɔ/



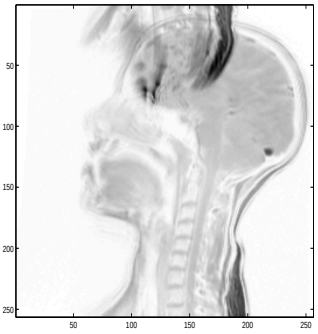
(b) another original negative instance of /ɔ/



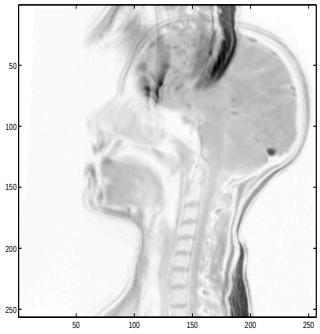
(c) average articulation of /ə/



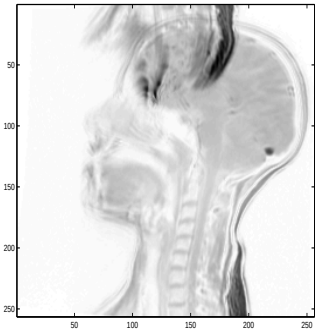
(d) average articulation of /i/



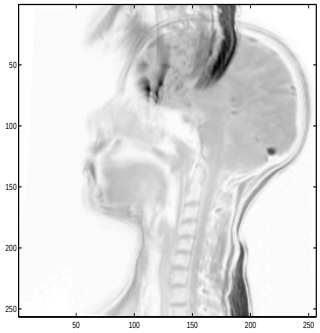
(e) average articulation of [ɪ]



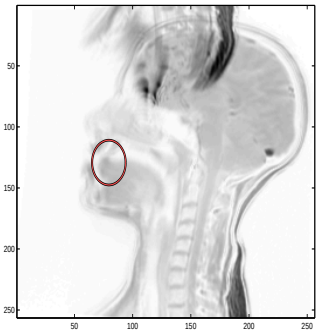
(f) average articulation of [ˈɪ]



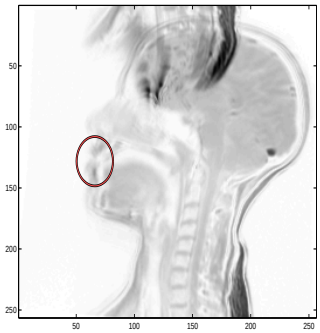
(g) average articulation of /ɑ/



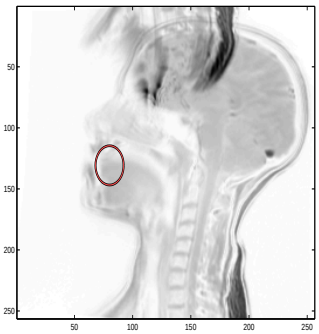
(h) average articulation of /ɔ/



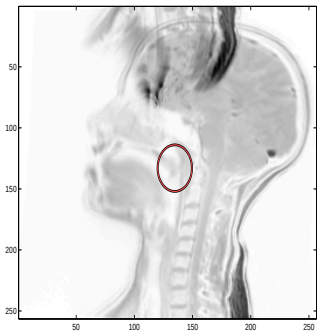
(i) average articulation of /s/



(j) average articulation of /f/

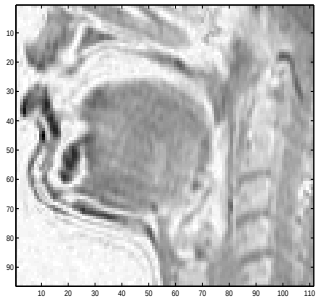


(k) average articulation of /d/

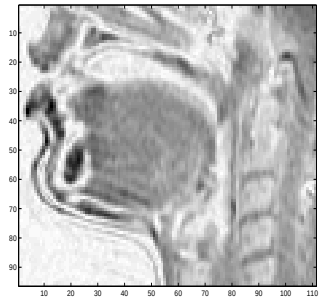


(l) average articulation of /n/

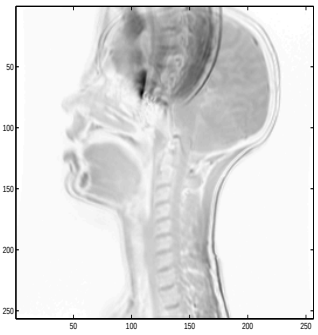
Figure 4.6: (a) and (b) Two arbitrarily chosen, magnified, original MRI negatives from the window of articulation images of /ɔ/, female speaker A. (c)–(l) Average articulation images. The ellipses show the areas where the paramount and distinguishing features of each consonant's articulation are present.



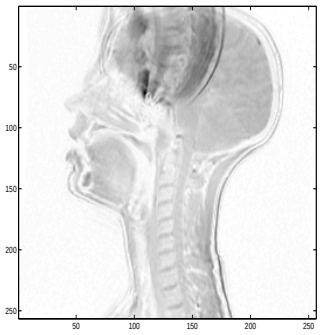
(a) an original negative instance of /ɔ/



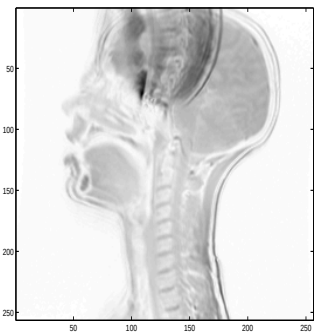
(b) another original negative instance of /ɔ/



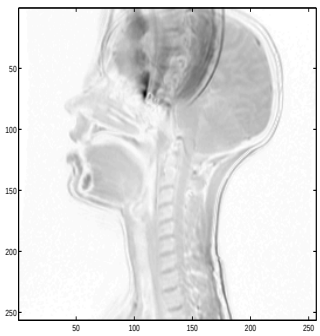
(c) average articulation of /ə/



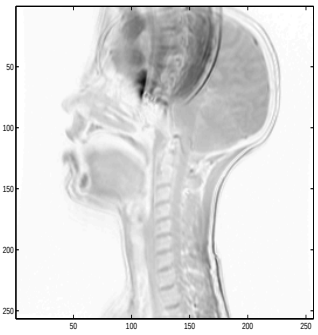
(d) average articulation of /i/



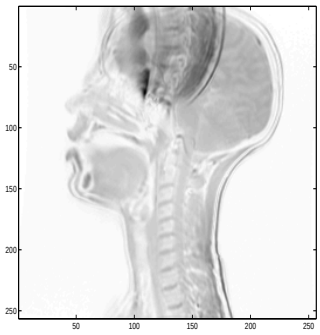
(e) average articulation of [ɪ]



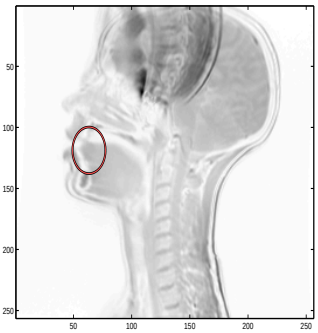
(f) average articulation of ['ɪ]



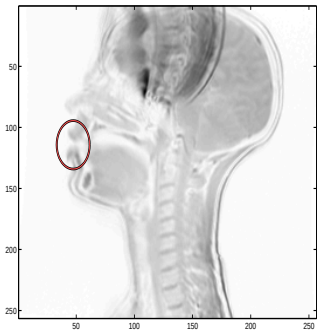
(g) average articulation of /ɑ/



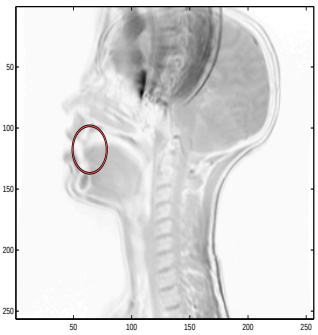
(h) average articulation of /ɔ/



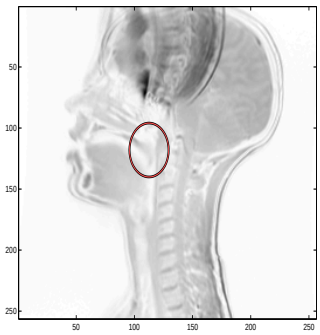
(i) average articulation of /s/



(j) average articulation of /f/

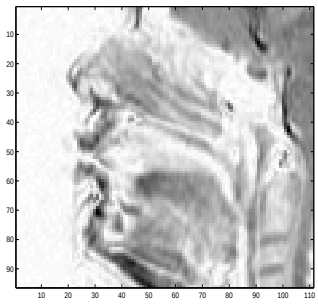


(k) average articulation of /d/

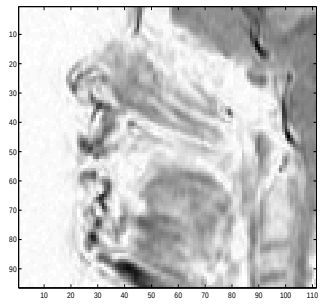


(l) average articulation of /n/

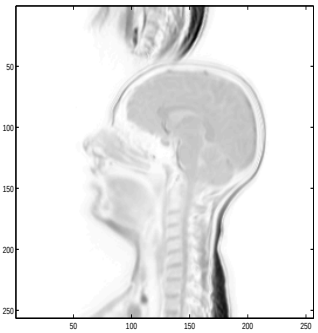
Figure 4.7: (a) and (b) Two arbitrarily chosen, magnified, original MRI negatives from the window of articulation images of /ɔ/, female speaker R. (c)–(l) Average articulation images. The ellipses show the areas where the paramount and distinguishing features of each consonant's articulation are present.



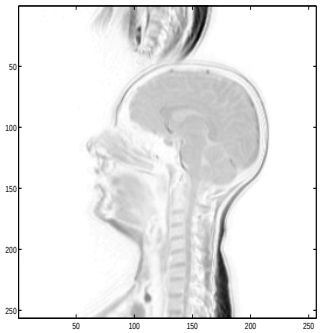
(a) original negative instance of /ɔ/



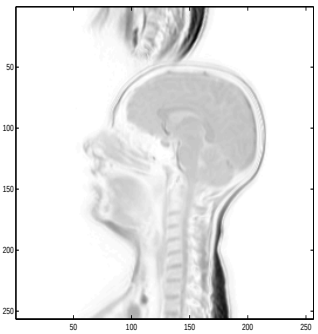
(b) an original negative instance of /ɔ/



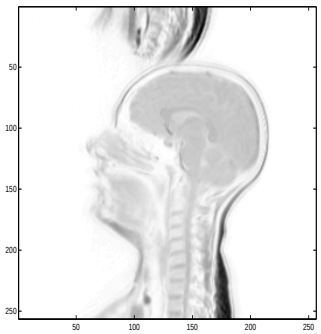
(c) another average articulation of /ə/



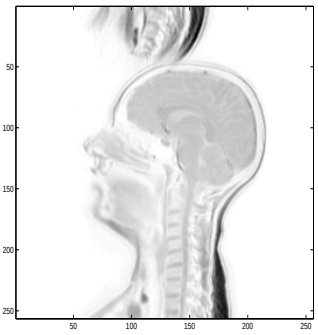
(d) average articulation of /i/



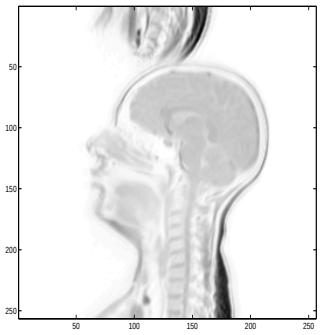
(e) average articulation of [ɪ]



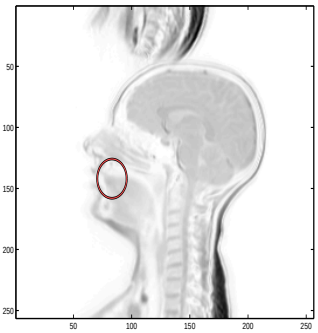
(f) average articulation of ['ɪ]



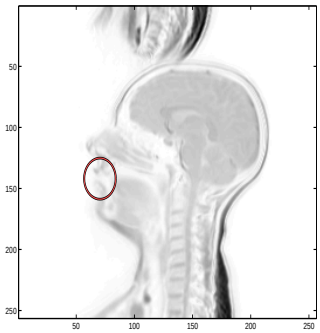
(g) average articulation of /ɑ/



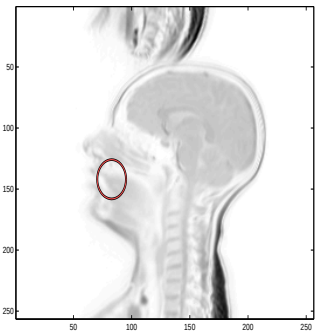
(h) average articulation of /ɔ/



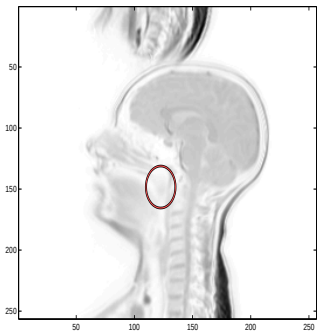
(i) average articulation of /s/



(j) average articulation of /f/

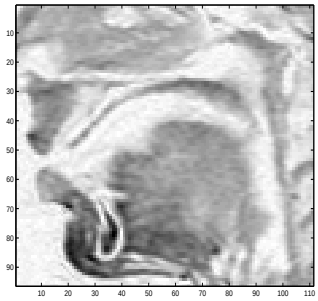


(k) average articulation of /d/

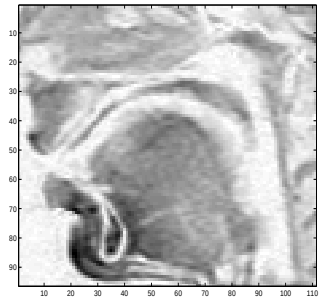


(l) average articulation of /n/

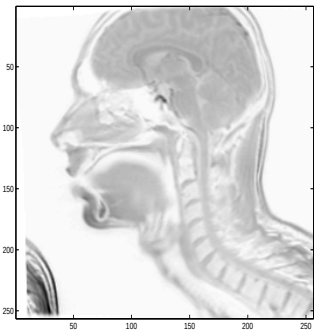
Figure 4.8: (a) and (b) Two arbitrarily chosen, magnified, original MRI negatives from the window of articulation images of /ɔ/, male speaker P. (c)–(l) Average articulation images. The ellipses show the areas where the paramount and distinguishing features of each consonant's articulation are present.



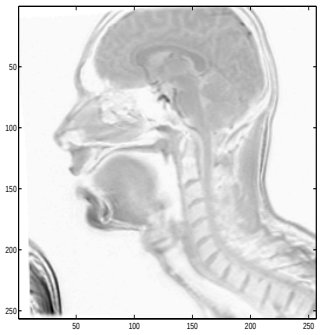
(a) an original negative instance of /ɔ/



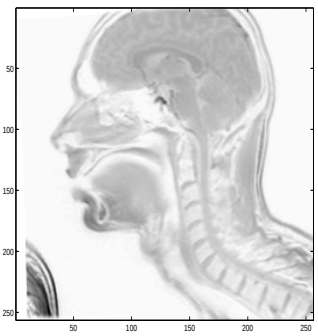
(b) another original negative instance of /ɔ/



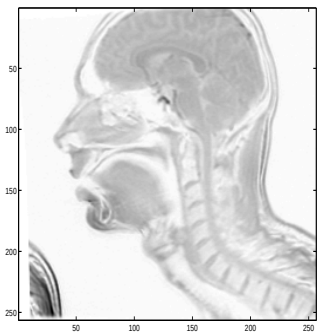
(c) average articulation of /ə/



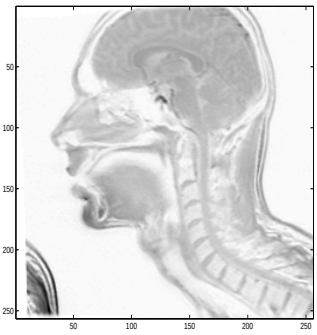
(d) average articulation of /i/



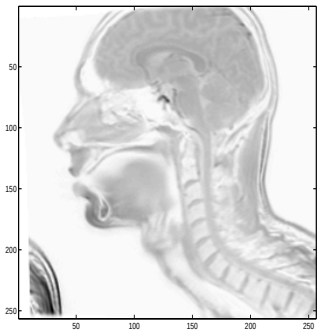
(e) average articulation of [ɪ]



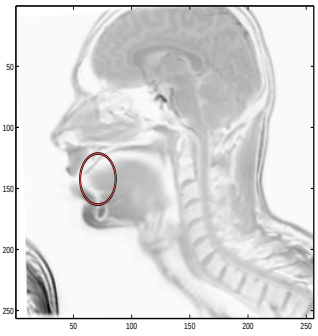
(f) average articulation of ['ɪ]



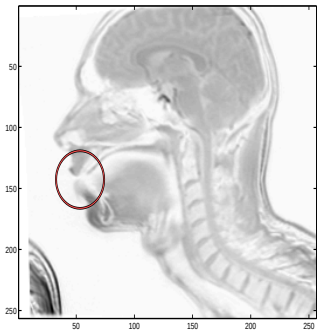
(g) average articulation of /ɑ/



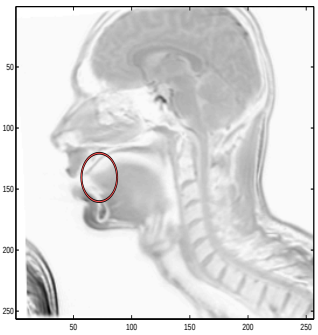
(h) average articulation of /ɔ/



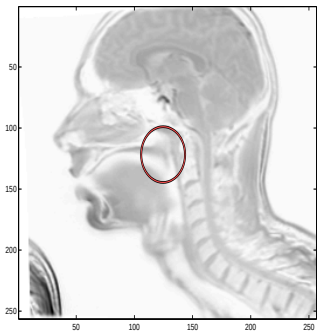
(i) average articulation of /s/



(j) average articulation of /f/

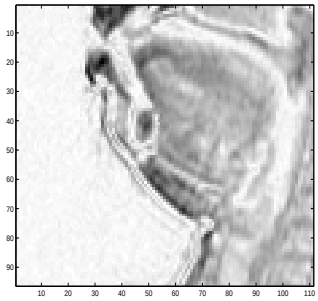


(k) average articulation of /d/

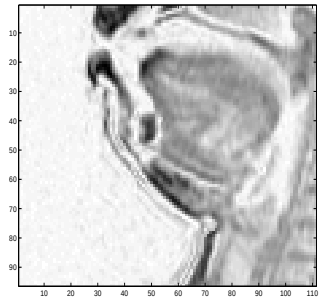


(l) average articulation of /n/

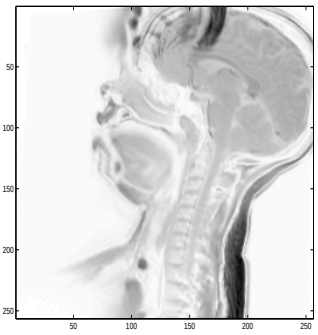
Figure 4.9: (a) and (b) Two arbitrarily chosen, magnified, original MRI negatives from the window of articulation images of /ɔ/, female speaker M. (c)–(l) Average articulation images. The ellipses show the areas where the paramount and distinguishing features of each consonant's articulation are present.



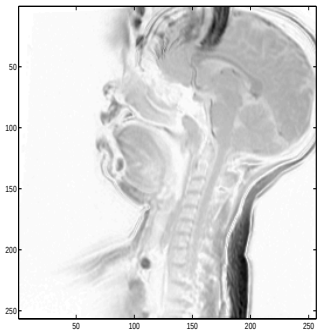
(a) an original negative instance of /ɔ/



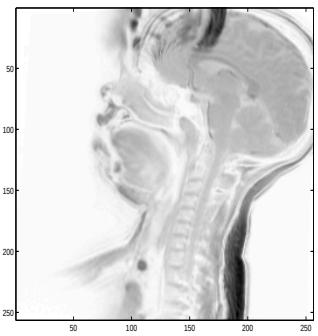
(b) another original negative instance of /ɔ/



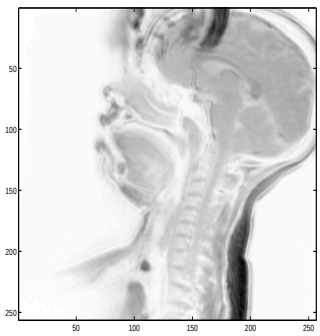
(c) average articulation of /ə/



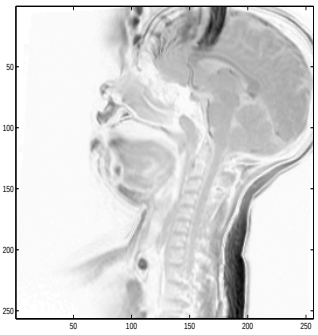
(d) average articulation of /i/



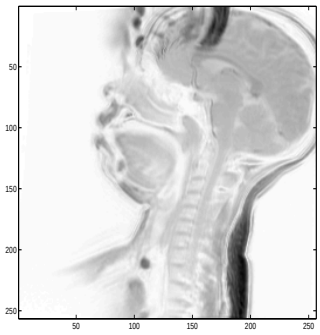
(e) average articulation of [ɪ]



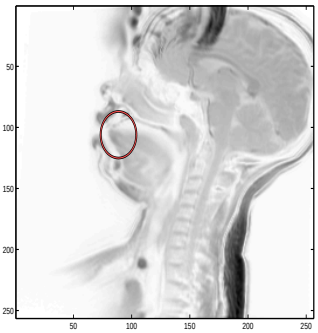
(f) average articulation of ['ɪ]



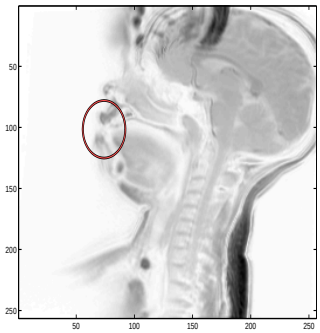
(g) average articulation of /ɑ/



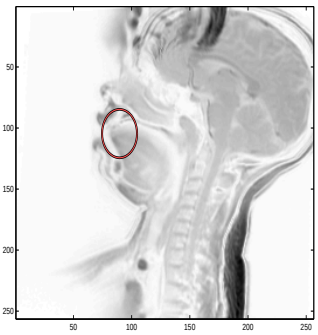
(h) average articulation of /ɔ/



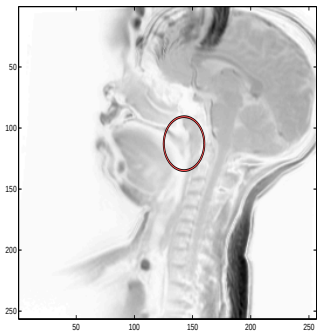
(i) average articulation of /s/



(j) average articulation of /f/

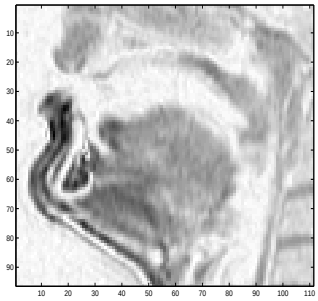


(k) average articulation of /d/

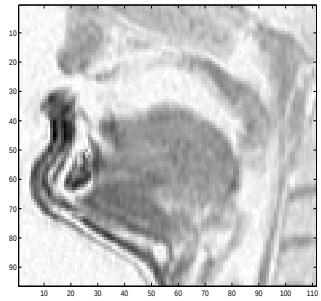


(l) average articulation of /n/

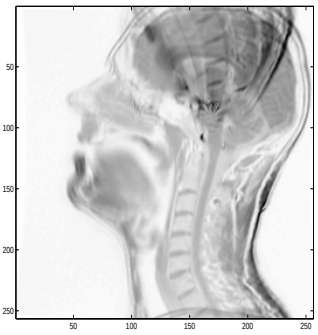
Figure 4.10: (a) and (b) Two arbitrarily chosen, magnified, original MRI negatives from the window of articulation images of /ɔ/, male speaker C. (c)–(l) Average articulation images. The ellipses show the areas where the paramount and distinguishing features of each consonant's articulation are present.



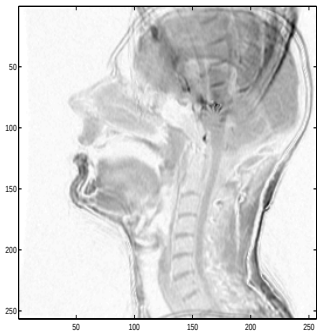
(a) an original negative instance of /ɔ/



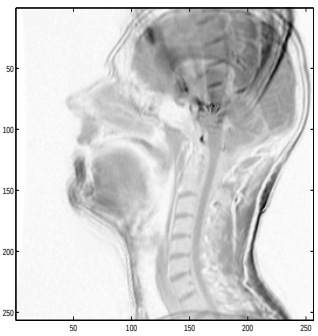
(b) another original negative instance of /ɔ/



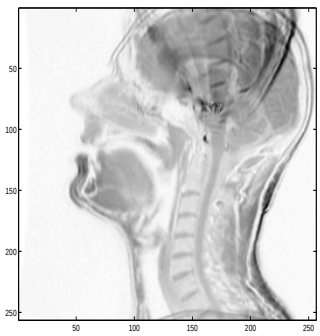
(c) average articulation of /ə/



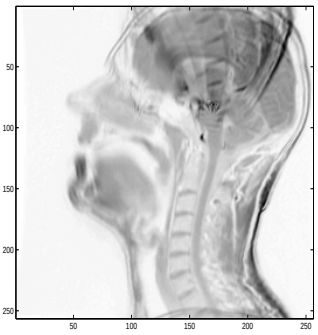
(d) average articulation of /u/



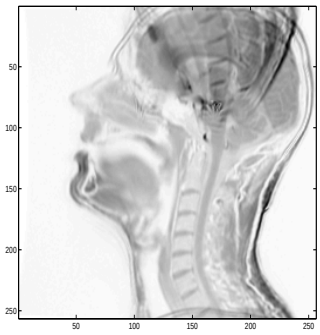
(e) average articulation of [ɪ]



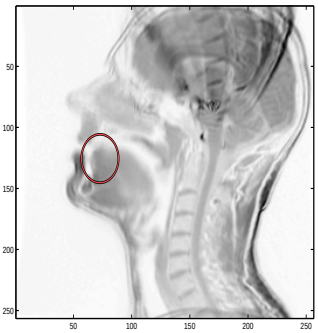
(f) average articulation of ['ɪ]



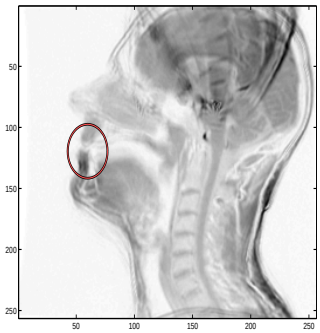
(g) average articulation of /ɑ/



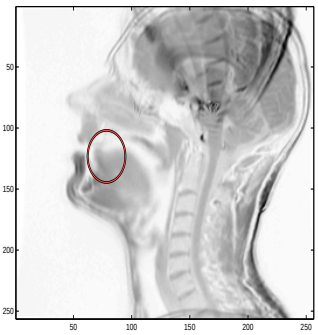
(h) average articulation of /ɔ/



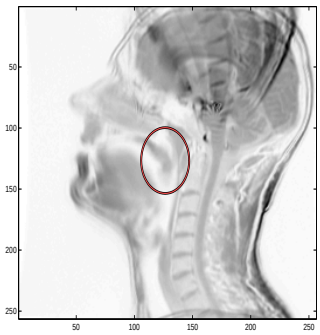
(i) average articulation of /s/



(j) average articulation of /f/



(k) average articulation of /d/



(l) average articulation of /n/

Discussion. In the above analysis of the vocal tract shapes, articulatory configurations and places of articulation across the different images illustrate patterns consistent with our articulatory expectations of the sounds. Thus, it can be concluded that the pixel intensity averaging technique is indeed a valid method for achieving a representative image demonstrating the "typical" speaker-specific articulation of the sounds, across different contexts.

However, some of the obtained average images prove to be less consistent in pattern, or at least the pattern is more difficult to discern in those images, for instance consider the narrow nasal cavity entrance in Figure 4.7 (k). Various factors can cause these inconsistencies in the images. One reason is the speaker-specificity of the average articulations, which introduces the inter-speaker variability factor in the comparisons of images. Another reason is the amount of motion blurring around the edges in average images compared to original images (a) and (b) in each figure. The lack of sharpness around the edges is a direct result of averaging over different instances of phoneme articulation in different context and the coarticulatory affects. The blurring in narrow regions of the airway can make it more difficult to observe an outstanding pattern. Another factor is the precision of the durations calculated by the forced alignment technique. As described in section 4.2.1, due to the high level of present noise in the MRI audio data, the acoustic models were initially trained on a large corpus of noise-free data, followed by a few rounds of re-training with the limited set of MRI audio data. The evaluation results proved a 93% agreement of the automatic labels to within 50 ms of the gold-standard ones (87% agreed to better than 30 ms). Alignment of the images to the audio is directly affected by the precision of the forced alignment, since the durations are computed based on the forced-alignment results. Each set of images corresponding to one sound may therefore erroneously contain some images that belong to neighbouring sounds. This adds noise to the typical articulation images which consequently makes the pattern harder to distinguish.

Although this method accommodates the variation originating from the sur-

rounding context into the "typical" articulation of the phoneme, it does not quantify or measure this variation. Variation in the vocal tract shape is the direct result of the fast movement and repositioning of the articulators from the posture appropriate to one sound to another posture. The rate of this variation and the manner of its correlation with the generated speech are important factors in speech production. In the next section, we investigate the correlation between articulation and the generated speech acoustics by quantifying the dynamics of the two and studying the relationship between them. We investigate the manner in which the amount of movement of articulators modifies the produced acoustics by introducing a parametrisation technique based on the image pixel intensity variations.

4.4 Parametrising the degree of articulatory movement

There have already been many attempts to parametrise the movement of the articulators and to apply those movements to speech coding [Schroeter and Sondhi 1991], speech synthesis [Sondhi 2002; Story 2001] and speech recognition [Wrench and Richmond 2000; Frankel *et al.* 2000]. These methods have invariably attempted to use detailed models of the articulators' positions to predict details of the temporal and spectral aspects of the signal.

Instead of analysing the exact movement of articulators, we propose a new approach, based on the outcome of section 4.3, for estimating the degree of articulator movement required for producing different speech sounds. In the proposed method, the dynamics of articulation is defined as the amount of movement in the vocal tract involved in articulation of a single unit. The goal here is to parametrise the articulation dynamics of each unit and to find the correlation between the degree of articulators' movements and the dynamics of the generated speech acoustics.

We look at the articulators' movements in terms of the amount of change in pixel intensities (i.e. brightness or darkness) in a sequence of MR images. The movements are parametrised by looking at the image pixels in the vocal tract area, the intensities of which vary across the images corresponding to the course of articulation of a

sound. Since changes in intensity only happen where there is a movement, the pixel intensity variation can be directly associated with the movement pattern of active articulators.

To evaluate this technique and to observe the correlation between the acoustics and articulation, dynamics (spectral variation) of the corresponding generated acoustics are similarly calculated, and the relationship between the two is investigated.

MRI Data. To calculate the degree of movement from the visual information, we use the averaging method explained in the previous section for generating "typical" vocal tract shapes. The durations of each instance of a phoneme were divided into thirds based on the assumption that the widely accepted 3-state phone models used in ASR would be sufficiently detailed to capture articulatory movement as well as acoustics. Examining the distribution of the number of images per unit, suggested that the minimum number of images allocated to each acoustic unit is three, and therefore dividing a sequence of three images into three parts means considering equal duration for all the three sub-phonemic segments. To be consistent, we generalised the assumption of equal one-third durations over the image sequences of all acoustic units. In special cases, where the number of images was not dividable by three, the (one or two) remaining image(s) were assigned to the middle one-third.

The images allocated to each one-third duration of a unit were averaged separately over the utterances of a speaker. Thus for each phone in the set of units articulated by an individual speaker, three average images were computed, each representing a third of the articulation process. For consonants these three stages can be interpreted as the approaching phase, actual articulation, and departure from the place of articulation. A similar interpretation can be applied to the three stages of vowel articulation: forming the tube shape for articulating the vowel, actual articulation and deforming the tube shape for the succeeding sound. Note that dividing the durations into thirds is solely for practical purposes in this application, and no strong

theoretical claim is made here about dividing the phoneme into sub-phonemes.

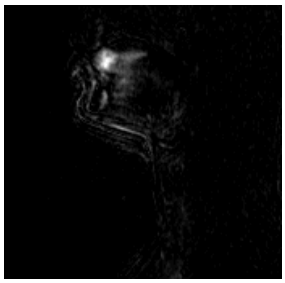
By taking the difference between the average MR images of the first, middle and final thirds of each phone, the amount of articulatory movement within an instance of a phone was quantified. The difference was computed by observing the variation in the intensity of the pixels between each pair of average images. The image intensity values are then re-scaled to (0,255) for consistency among all the images. These 'degree of articulation' images indicate the articulatory movements involved in production of each phone.

The average magnitudes of the differences between first and middle, and between middle and last, thirds of the phone [ˈɔ] for different speakers are shown in Figure 4.11 and Figure 4.12.[6] In each pair of images for individual speakers, areas with bright pixel intensities appear around the tip and front of the tongue, and the back of the tongue. These areas of higher pixel intensity values are formed because of the tongue retraction during the articulation of vowel [ˈɔ], that leads to a sudden change in the intensities of the pixels in those areas. However, not all non-zero pixels scattered across the image are *directly* related to the place and manner of articulation of the sounds, such as the bright pixels visible along the larynx in Figure 4.12 (a), (b), (e) and (f). These non-directly related non-zero pixels are the results of inevitable movements in other (than tongue) inter-connected tissues of the mouth, slight movements of the subject's head, and movements caused by breathing. We examined all of the difference images of individual speakers, and observed that the noise pixels, if present, were present in almost all of the difference images of a particular speaker, making the appearance of non-related non-zero pixels a speaker–

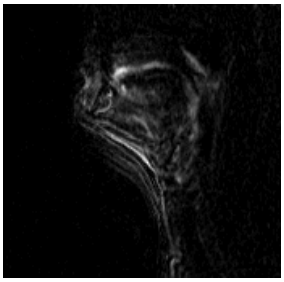
[6]Note that, for display purposes here, each image was scaled individually to its maximum brightness value, but the images used for the experiments were all re-scaled to grayscale 256 for consistency.

Figure 4.11: Images showing the average articulatory movement between the first and middle (left) and middle and final (right) thirds of the phone [ˈɔ] for speakers A, C and L. The intensity changes have been subtracted pixel by pixel; the bright pixels represent large changes in intensity values due to the movement of the articulators.

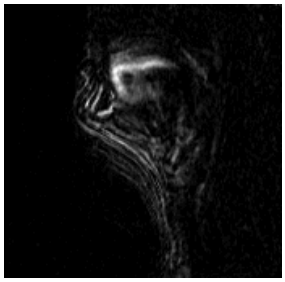
(a) speaker A, first–middle thirds subtraction



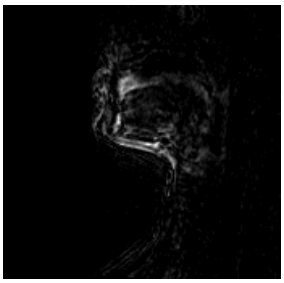
(b) speaker A, middle–final thirds subtraction



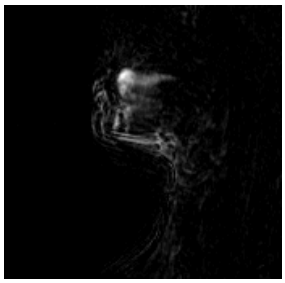
(c) speaker C, first–middle thirds subtraction



(d) speaker C, middle–final thirds subtraction

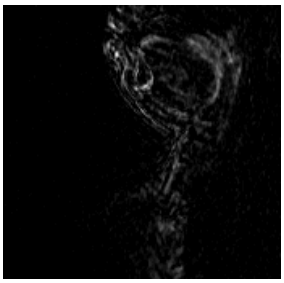


(e) speaker L, first–middle thirds subtraction

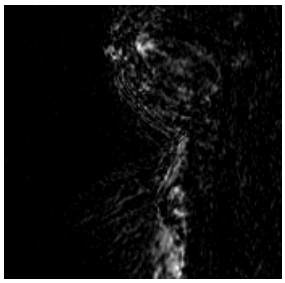


(f) speaker L, middle–final thirds subtraction

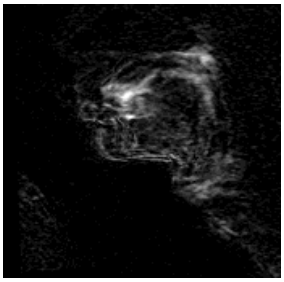
Figure 4.12: Images showing the average articulatory movement between the first and middle (left) and middle and final (right) thirds of the phone ['ɔ] for speakers M, P and R. The intensity changes have been subtracted pixel by pixel; the bright pixels represent changes in intensity values due to the movement of the articulators.



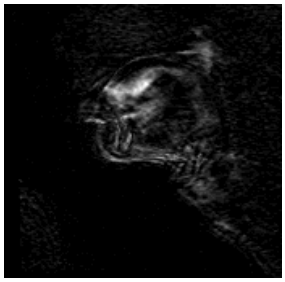
(a) speaker M, first–middle thirds subtraction



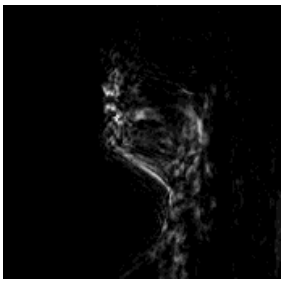
(b) speaker M, middle–final thirds subtraction



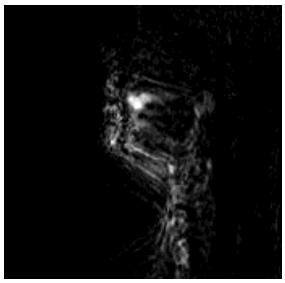
(c) speaker P, first–middle thirds subtraction



(d) speaker P, middle–final thirds subtraction



(e) speaker R, first–middle thirds subtraction



(f) speaker R, middle–final thirds subtraction

based phenomenon. Based on this evidence, we believe that the noise pixels do not affect the phoneme-based speaker-independent analysis that we intend to carry out, but could lead to inconsistencies if a speaker-based analysis were desired. The overall degree of articulator movement in each phone was quantified by summing the intensity levels indicated in images such as Figure 4.11 and Figure 4.12. These figures capture the articulatory movements in the first and final parts of each phone, respectively.

Acoustic Data[7]**.** For the acoustic data, a measure of spectral variation (referred to as acoustic dynamics) was devised, again by comparing the first, middle, and final thirds of each labelled phoneme. In this case, a modified symmetrical form of the Itakura-Saito distance [Itakura and Saito 1970; Gray Jr and Markel 1976] is calculated between the acoustic signals for the same regions as in the MRI analysis. The Itakura-Saito distance was chosen for its mathematical simplicity and well-documented behaviour. In addition, it can be calculated very easily using "linear prediction" analysis of the signal. The symmetric Itakura-Saito distance metric used in this work is defined as

$$Dist(PA(\omega), PB(\omega)) = f(PA(\omega), PB(\omega)).f(PB(\omega), PA(\omega)) - 1, \tag{4.2}$$

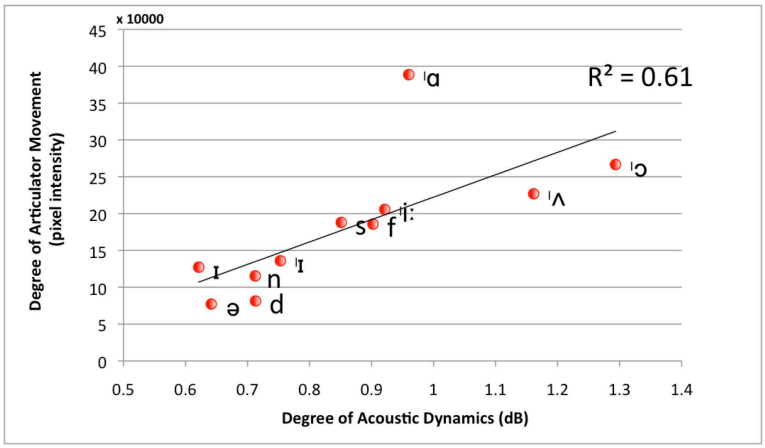
where $PA(\omega)$ and $PB(\omega)$ are power spectrums of signals A and B respectively [Baghai-Ravary 2010]. The function $f(PA(\omega), PB(\omega))$ is defined as:

$$f(PA(\omega), PB(\omega)) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{PA(\omega)}{PB(\omega)} - \log\left(\frac{PA(\omega)}{PB(\omega)}\right) - 1\right]d\omega. \tag{4.3}$$

A single symmetrical Itakura-Saito distance was calculated based on the whole of each third of the respective phone. These acoustic distance measurements were, like the MRI-based measure of articulator movement above, averaged over all in-

[7]Thanks to Dr. Ladan Baghai-Ravary for carrying out the acoustic analysis described in this paragraph.

Figure 4.13: Phone articulation movement measures averaged over six speakers: the degree of dynamics in the MRI data plotted against dynamics in the acoustic signal. The y-axis is the sum of the pixel intensities of the two images presenting the movement in the first and final parts of articulation of each phone. The x-axis is the sum of the spectral differences in the first and final parts of the acoustics of each phone.



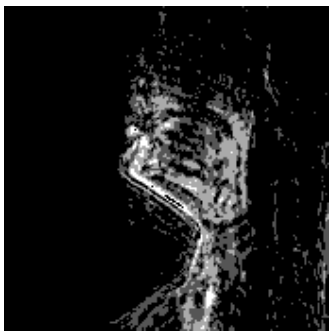
stances of each phone, spoken by an individual speaker, and their range normalised separately for each speaker. These normalised values were then averaged across all speakers to provide a single measure of 'degree of acoustic dynamics' (spectral changes) associated with the first and final parts of each phone.

The data for first and final parts of the respective phones are combined, by summing them; this is in order to capture the total degree of dynamics within each phone. To illustrate the relationship between articulator movement and acoustic dynamics, the measures described above are plotted in Figure 4.13.

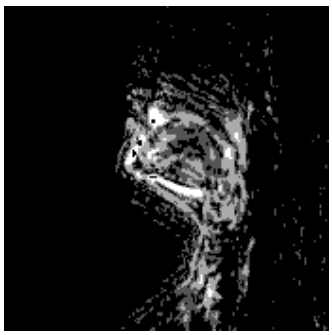
There is a clear correlation between articulator movement and acoustic dynamics, with a coefficient of determination (R^2) of 0.61, which is statistically significant ($t = 3.75$; $p \leq 0.01$). One phone, ['ɑ], has significantly greater articulator movement than would be expected from the associated acoustic dynamics, but the others are very clearly and directly related. The greater degree of articulatory movement for ['ɑ] is linked to its articulatory configurations; in practice more than one component of articulation are engaged in articulation of ['ɑ], e.g. tongue and jaw movements,

together with pharyngeal constriction. This can be observed by comparing difference images of articulation of [ˈɑ] and [ˈʌ] in Figure 4.14. Although approximately equal amounts of movement are observed at jaw and tongue regions in all images, more movement is present at the pharyngeal region in Figure 4.14 (a) and (b) in comparisons to Figure 4.14 (c) and (d): the pixels at pharyngeal region are consistently brighter in Figure 4.14 (a) and (b) relative to Figure 4.14 (c) and (d).

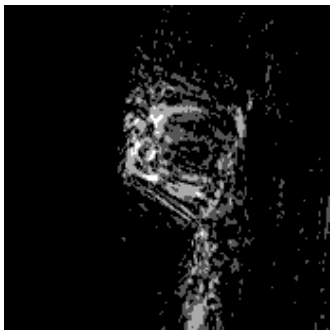
Figure 4.14: Images showing the average articulatory movement between the first and middle (a) and middle and final (b) thirds of [ˈɑ] and the first and middle (c) and middle and final (d) thirds of [ˈʌ] for speaker R. The intensity changes have been subtracted pixel by pixel; the bright pixels represent significant changes in intensity values due to the movement of the articulators. More movement is observed at the pharyngeal region in [ˈɑ]. The grey-levels have been scaled for display purposes.



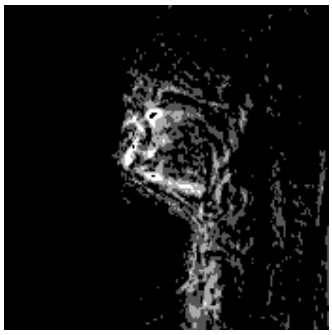
(a) [ˈɑ], first–middle thirds subtraction



(b) [ˈɑ], middle–final thirds subtraction



(c) [ˈʌ], first–middle thirds subtraction



(d) [ˈʌ], middle–final thirds subtraction

Discussion. From these observations it appears that although most of the phones in the table would normally be considered "steady" sounds, only three [ə], [ɪ] and [d] involve very little or no movement of the articulators during their production. This apparent little movement of unstressed vowels [ə] and [ɪ] may be an artefact of the short duration of the sounds obtained from the aligned acoustic data. The shorter durations lead to fewer frames being associated with the phone articulation and not much variation can be observed in short sequences of their corresponding image frames, which sometimes are very few in number because of the relatively low frame rate of the MRI data.

Another interesting observation is the distribution of consonants and vowels in the plot of Figure 4.13. As it is depicted in the plot, for vowels, there is no significant pattern between the type of the vowels (e.g. open or closed, front or back etc.) and their distribution in the correlation plot. In contrast, the consonants of the same type (sharing the manner or place of articulation) appear very close together in the plot. That is, the two stops /d/ and /n/ have very similar articulatory-acoustic dynamics correlation ratios. This is also true in the case of the two fricatives /s/ and /f/. However, the rather small number of observed phonemes in each consonant category prevents us from being able to prove the robustness of the correlation ratio within these categories.

4.5 Conclusion

To study the correlation between acoustics and articulation, the corresponding acoustic and articulation units are first specified. We chose phoneme-sized units to be the acoustic units studied in this research. The acoustic signals were forced aligned to the transcriptions using an HMM-based ASR system that is trained on a British English corpus in addition to the MRI audio data. The temporal information of the post alignment labels was used to align the MR images to the transcriptions and subsequently to the acoustic signal.

The diversity present in articulatory data of an acoustic unit was summarised

including the variation caused by surrounding context, and an average articulation image was obtained that represented the "typical" articulatory configuration a speaker employed when generating a particular sound. The consistency between the articulatory configurations observed in the average images and the articulatory configurations for different sounds (expected from prior studies [Fant 1960; Perkell 1969]) suggests that this method can be successfully used to picture the typical articulation of sounds. Moreover, this approach can be used to image articulatory configurations such as the hold phase of plosives, which is hardly possible with other methods such as static MRI.

Finally, a new technique was devised that used the degree of articulator movements for characterising the acoustics, instead of looking at the shape of the vocal tract and area functions. The results suggest that although the details of phoneme articulation are speaker-specific, a strong pattern of correlation can be observed between the degree of spectral variation (acoustic dynamics) and the amount of movement in the vocal tract.

In the next chapter, we review and introduce techniques for extracting the shape of the vocal tract from images, by determining the air–tissue boundary, for parametric representation and modelling of articulation.

Chapter 5

Segmenting Vocal Tract MR Images

5.1 Introduction

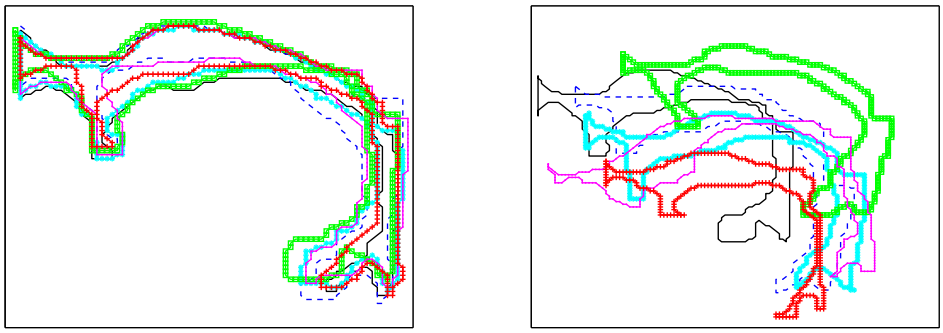
In the previous chapter, we used image pixel intensity features without parametric shape or edge information to extract useful articulatory information from sequences of MR images corresponding to production of different sounds. The intensity variation information were demonstrated to be informative for studying general patterns of articulation. Nevertheless, intensity fluctuation patterns do not provide us with a parametric description that can be analysed and further used for modelling articulation. Hence, a parametric feature extraction of the vocal tract shape or articulators is desired to compensate for the generality and limitations of pixel-based studies.

As we discussed in chapter 3, with large databases of MR images, we cannot ignore the need for automatic approaches for outlining the vocal tract shape. We also reviewed some of the existing techniques developed to detect the outline of the vocal tract in that chapter. In this chapter and the next, we attempt to extract the shape of the vocal tract automatically through image segmentation. Depending on the application, the extracted segment can be used to model either the vocal tract shape statistically or extract area function information.

We first describe the existing challenges for successful segmentation of vocal tract shapes from an MR image database of dynamic articulation (section 5.2). We then continue by introducing a selection of image segmentation techniques that are used to automatically extract the shape of the vocal tract and delineate the boundaries in images. In section 5.4, we demonstrate how the challenges in vocal tract boundary delineation influence our choice of segmentation technique. This is achieved by applying a commonly used region-based segmentation to our MR images and analysing the results, the successes and failures. Finally, the conclusions of this chapter are presented in section 5.5.

5.2 Shape extraction challenges

Automatic segmentation of vocal tract MR images involves addressing numerous challenges, including (i) high variability in the vocal tract shape introduced by many distinct sounds, (ii) high variability in the vocal tract shapes of different speakers, (iii) significant anatomical changes during the articulation due to movement of articulators, (iv) noise and blur introduced by dynamic behaviour of the vocal tract articulators and (v) the connectivity of the tract airway to other channels such as surrounding air (through the lip opening), nasal cavity (through the velum opening) or trachea (through the glottis). Figure 5.1 shows examples of variation in topological features of vocal tract shape within different phonemes for the same speaker (a), illustrating challenge (i), and between speakers for the same phoneme (b), illustrating challenge (ii), above. Blurring around edges is inevitable in dynamic MR imaging of speech due to the fast speed of articulation compared to the MR images' acquisition rate. The connectivity of the airway to other channels of air which share the same intensity characteristic as the vocal tract, makes it difficult to distinguish the boundaries of vocal tract (Figure 5.2). In addition, there is the challenge of "false" openings, where boundaries are missing due to the thin layer of tissue between the airways. An example of this can be found at the palate, as illustrated in Figure 5.2, where the thin tissue layer at the palatal region is not captured in the



(a) same speaker, different phonemes (b) same phoneme, different speakers

Figure 5.1: Variation in vocal tract topology (a) within phonemes (speaker A) and (b) between speakers (phoneme /ɔ/).

image, and thus the vocal tract airway is connected through the palate to the nasal cavity.

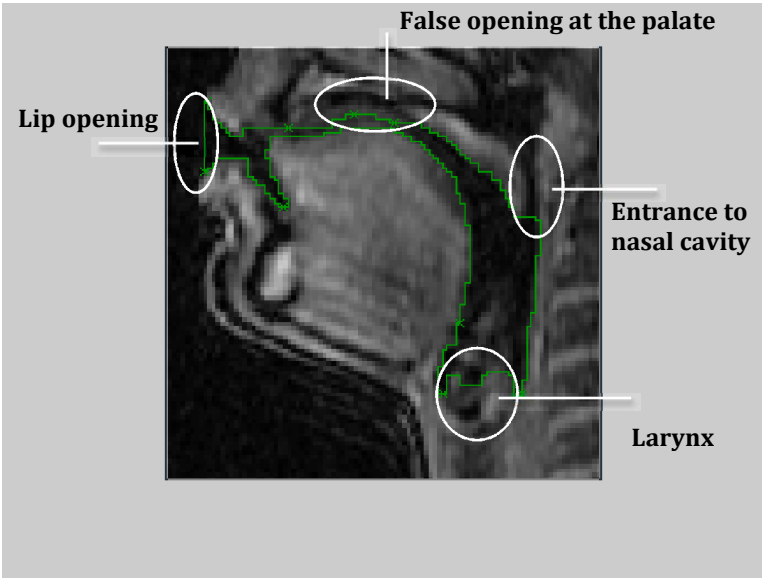


Figure 5.2: Vocal tract contour (green line) and openings to passages of air marked on the MR image.

To segment the vocal tract shape automatically, a technique is required that can address most, if not all, of the challenges discussed to extract the shape successfully. In following sections, we review some image segmentation techniques, and discuss

the suitability of each technique for this application.

5.3 Image segmentation methodologies

There is an extensive variety of methods to detect, extract and analyse features in images. Segmentation techniques aim at extracting the object information and representing them in the form of a geometric structure. There are generally two tasks involved in segmentation of objects in images: ***recognition***, which refers to finding the approximate position of the object of interest in the image, and ***delineation***, which refers to extracting the precise spatial boundaries of the object [Liu and Udupa 2009]. Most of the segmentation approaches tend to focus on one of these tasks more than the other, and the choice of which one to apply heavily depends on the application.

Segmentation categories. There is an extensive variety of segmentation methods to extract the shape or contours from images. Detailed treatment of all of these categories is beyond the scope of this thesis. We shall describe a few categories here that are applied in the field or in this thesis.

Region-based methods rely on image properties such as intensity of pixels in a region, texture, or spectral profiles, in detecting the boundary and segmenting the shape. Examples of these methods are thresholding [Saha and Udupa 2001] and fuzzy connectedness [Udupa and Samarasekera 1996]. Region-based methods rely on local information during segmentation and typically do not use global shape information.

Model-based approaches are based on the assumption that the objects of interest share similar topology. Therefore, a probabilistic model can be used to capture the general features of the object and its variations in order to build a *prior* for further segmenting similar images. Examples of these approaches include active shape models [Cootes *et al.* 1995] and active appearance models [Cootes *et al.* 2001].

Boundary-based methods, as the name suggests, aim at finding the optimum

boundary that can represent structure of the shape. An example of this category of approaches is the live wire segmentation technique [Falcão *et al.* 1998], that relies on input from the user and boundary information of the images to find the contour. Another example in this category is the snakes method [Kass *et al.* 1988] which is very popular in extracting deformable boundaries.

We provide a brief overview of a subset of these techniques next.

5.3.1 Thresholding

Thresholding is the most basic form of region-based segmentation, and is very widely used due to its simplicity. The general approach is to find a threshold and convert the grey-level image to a binary image based on that threshold. The crucial part of the method is to determine a reasonable thresholding value. A review of automatic global thresholding methods can be found in [Lee *et al.* 1990].

Thresholding has been used in the literature to find the boundaries of the vocal tract [Story *et al.* 1996; Engwall and Badin 1999]. It can successfully address challenges related to variation (challenges (i), (ii) and (iii) in section 5.2) as it is based on intensity features of the current image. The vocal tract openings, however, are usually blocked by manually placing a barrier to stop the region from growing outside the vocal tract boundaries, as reported in [Story *et al.* 1996]. Also, in traditional thresholding approaches, manual checking and corrections may be necessary [Engwall and Badin 1999], disqualifying the method for automatic shape extraction.

5.3.2 Fuzzy connectedness

Fuzzy connectedness (FC) segmentation [Udupa and Samarasekera 1996] has been developed to address the issue of data inaccuracy. FC segmentation is a region-based approach, from the family of thresholding methods, and used to find the connection map between the image spatial elements, *spels*, by defining a level for local "hanging-togetherness" of the elements. The strength of the affinity between the spels is estimated based on their spatial distance and their intensity similarity

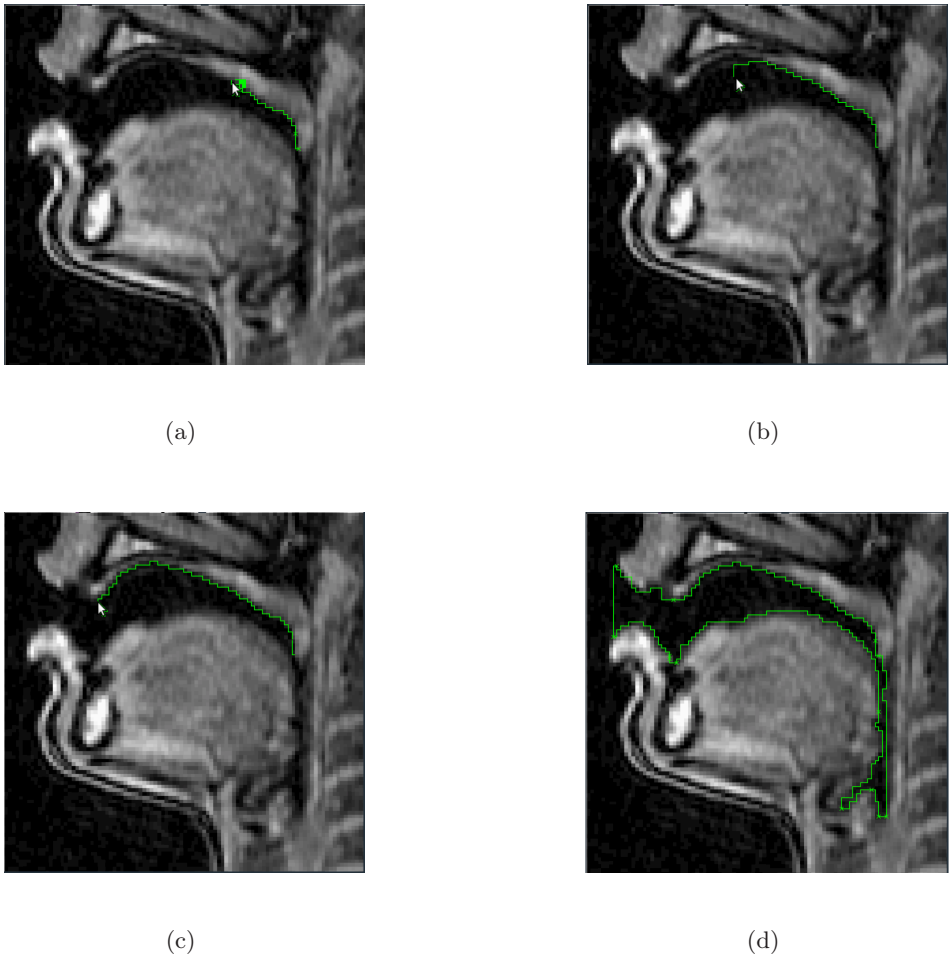
inside the target object. A threshold is then used on the strength of connectedness between the element to segment the final shape.

Fuzzy connectedness has been applied successfully in segmenting tissues in the presence of intensity gradation in MR and CT images. An advantage of using fuzzy connectedness for segmenting the vocal tract shape compared to the traditional thresholding segmentation is that fuzzy connectedness can address the blur and noise introduced in an MRI database of dynamic speech articulation, described as challenge (v) above.

5.3.3 Live wire

Live wire [Falcão *et al.* 1998] is a 2D user-steered boundary delineation technique purely based on image properties. The method uses the information (anchor points) provided by the user to find the optimal boundary in real time. The user initialises the algorithm by selecting a point on the boundary, then by moving the cursor, a globally optimal boundary is generated by live wire algorithm as explained in [Falcão *et al.* 1998], from the starting point to the current point. Live wire snaps to the boundary if the cursor moved by the user is close enough to the boundary, and a new starting point is generated at the snapping point. Live wire is dependent on human input for recognition, and therefore is not as efficient as ASM. Despite this drawback, because of the optimum real time boundary cost function, it is more accurate than ASM. Figure 5.3 shows snapshots of semi-automatic segmentation of a sample MR image of vocal tract from our database by a human expert.

Live wire is popular for being highly accurate as it agrees with the input from the expert. Other main advantages of live wire include being faster and more accurate than manual tracing, not requiring extensive training, and not requiring post-hoc correction. More importantly, live wire takes the orientation information into account when determining the boundary. This is very useful in distinguishing the boundary of similar objects even if they come very close together in the image domain.



(a) (b)

(c) (d)

Figure 5.3: Semi-automatic boundary delineation of the vocal tract in MR images. The cursor is guided by a human expert, and when it is close enough to actual boundary, live wire snaps to the boundary. If the human expert disagrees with the generated boundary, the cursor must be traced back to the last agreed upon point on the boundary, where the user fixes the point by a mouse click. Live wire then continues the optimum boundary search starting from the snapping point. For this figure, the live wire method in 3Dviewnix software [Udupa *et al.* 1993] was used.

The main drawback of live wire, as stated before, is the dependency on human input for recognition. This would make it a tedious approach to apply to large databases if it were used alone, requiring extensive manual input. In addition, if visual patterns are missing along the boundary, live wire cannot successfully follow the continuation in the boundary unless lots of points are selected.

5.3.4 Active shape models

Active shape models (ASM) [Cootes *et al.* 1995] are widely applied as a model-based approach in image analysis. Active shape models are trained on sample shapes to capture the shape model and its variation statistically. The shape model is then used as a prior for recognising the object of interest in unobserved images. The use of prior information allows ASM to recognise shapes despite the large variation and appearance in the images. The desired shapes are defined in form of a set of *landmarks*, labelled points that represent an object. The dominant pattern of the shape model is described by an average shape vector and a combination of its eigenvectors that covers the maximum variability in the set of training shapes. During the search stage, new target points are defined close to landmark points of the model (initialisation), and the model is updated using the variation range to best fit the object. The limitations imposed on the parameters of the model guarantees the similarity of the new segmentation to the training shapes by constraining them to the global (prior) shape model.

ASM is a popular method in medical imaging segmentation. In the area of articulation studies, it has been used by Vasconcelos *et al.* [2011] to model the vocal tract shape, and by Matthews *et al.* [2002] and Avila-García *et al.* [2005] to parametrise articulators such as lips and tongue. However, ASM has a few shortcomings that limit its application: sensitivity to initialisation; ignoring specifics of the current image in recognition phase (using only the prior information about the training landmarks that are in turn blurred due to statistical averaging process); and inaccuracy in boundaries, as segmented results are parametric descriptions of the shape,

and not the exact perceptual boundary of the object. These limitations make ASM weaker in delineating boundaries accurately compared to boundary-based technique such as live wire, particularly in this application. Although ASM can address the challenges of variability or missing boundaries along the tract airway, it is not successful in handling the blur and noise around the edges, and cannot determine the precise contours between the landmarks.

5.3.5 Oriented active shape models

Oriented active shape models (OASM) [Udupa *et al.* 2006] combine the powerful aspects of active shape models in capturing the statistical information of the shape and live wire in finding the globally optimal oriented boundary to find and segment the objects effectively. The orientedness property assumes that all the boundaries and boundary elements are oriented, thus all the object boundaries have well-defined "inside" and "outside". OASM brings together the strong features of a model-based segmentation method, ASM, in recognition, and of a boundary-based method, live wire, in delineation to segment the objects both accurately and effectively. The coordination between the two approaches in OASM helps them support each other and brings more accuracy and efficacy to both recognition and delineation. Live wire can produce a globally optimal boundary if the anchor points are close enough to the actual boundary in the image. ASM provides the landmark information through recognition, thus replacing the human operator support required for live wire. This makes live wire more efficient. On the other hand, live wire is used to find the minimum cost optimal oriented boundary for the shape model, which helps ASM in recognition. In addition, the boundary orientedness of live wire helps OASM distinguish the nearby object boundaries with similar properties that do not belong to the object of interest. By combining the two approaches, OASM can address the limitations of each of them as discussed in previous sections. OASM involves two-level dynamic programming [Eddy 2004], one level for recognising the shape of the object, and another level for finding the optimum delineation of the detected

object. If the initial recognition of the object finds a model close to the boundary of the shape in the image, the delineation step returns a globally optimal boundary that fits within the description of the shape model.

A few advantages of OASM over ASM and live wire, as stated in [Udupa *et al.* 2006], can be summarised as improved automation and robustness of recognition, higher accuracy, smaller training space and lower sensitivity to initialisation compared to ASM. The global optimal delineation attributes of live wire together with the statistical shape modelling of ASM, both embedded in OASM, help address the vocal tract shape variability challenges (i), (ii) and (iii) effectively. Furthermore, the fine delineation provides a more accurate vocal tract contour compared to other methods such as ASM, thereby resolving issues related to noise caused by motion blur (challenge (iv)). Limiting the parameters of the resulted shape model by global shape constraints forces the generated shape to fit within the actual boundaries of vocal tract shape, thereby addressing the missing boundaries as in challenge (v).

In the next section, we begin our automatic segmentation of MRI data by adopting a region-based method. We start our investigation with region-based approaches as they have been widely used in the literature to segment vocal tract MR images in small quantities.

5.4 Region-based vocal tract segmentation

A region-based segmentation approach is an appealing choice for extracting vocal tract shapes, as it can deal with the inter speaker/phoneme variability of the vocal tract shape, without requiring excessive training. The vocal tract portion of the image can be considered as an *object* of roughly uniform intensity, surrounded by tissues with distinct intensities. The pixel intensity and object homogeneity properties of the tract airway can be used as distinctive properties to segment the tract in the MR images. However, the lack of sharp boundaries due to the noise and blurring caused by motion blur and low signal-to-noise ratio affects the intensity homogeneity around the edges. As a result, the object boundary has a *fuzzy* characteristic.

Hence, any region-based method must also be able to handle the blur and noise that is introduced by articulator movements.

We begin our automatic shape extraction using the fuzzy-connectedness (FC) segmentation technique [Udupa and Samarasekera 1996]. FC segmentation is a region-based approach that has been developed to take the fuzziness of images into account when segmenting fuzzy objects. The vocal tract boundaries in a dynamic MRI database have a fuzzy nature as a direct result of articulators' movements, hence compared to other region-based thresholding techniques, fuzzy connectedness is a more suitable selection.

5.4.1 Fuzzy connectedness segmentation

In this section, we briefly describe the main principles of the fuzzy connectedness (FC) approach used in this application, as presented in [Udupa and Samarasekera 1996] for 2D images. A more thorough description of the general method can be found in that paper.

A fuzzy object has two dominant characteristics that lay the foundation for the fuzzy connectedness segmentation approach. Firstly, the so called "object" to be segmented in the images is composed of pixels with similar, but different shades of grey. Thus, generating an intensity characteristic component that describes the closeness of the pixels within the object. Secondly, the image elements that belong to a particular object in the image generally hang together in a certain way to form the object. Hanging togetherness is a grouping property, defined based on intensity homogeneity of the pixels in a group, that can define the grouping state of the pixels of an image.

The key idea of FC segmentation is to create a "connectivity map" that defines the strength of connectivity between every pair of pixels in the image through a fuzzy relation. In other words, a global fuzzy relation is defined for every pair of points that shows the hanging-togetherness of the pixels through out the image.

The strength of connectedness of pixels is locally defined in terms of a notion

referred to as "affinity". Two factors play crucial role in forming the affinity between two pixels: the spatial closeness and the intensity similarity. For two adjacent pixels, the affinity is decided based on the homogeneity of the region, and the distance of the intensity difference of two adjacent points from the intensity of object of interest. Every pair of points on the pixel grid of the image are connected through different paths on the grid map, where each path is formed by simply a sequence of links between two consecutive grid points. The strength of the path is equal to the affinity of the weakest link among its links. The connectedness strength between two grid points is then defined as the strength of the strongest path connecting them.

A global map of connectedness is generated based on the connectivity values of every pair of image elements. The resulting connectivity map is then thresholded according to the intensity of the object of interest and the rest of the image is filtered out. The algorithm is initialised by defining one or more "seed" points in the target object's region. The mean and standard deviation of the intensity values in the region of the object of interest is obtained with minimal training. In the following paragraphs, we briefly explain how the affinity and connectivity of the pixels are measured. The notation of the original paper [Udupa and Samarasekera 1996] is used here for describing the expressions.

Fuzzy Affinity. For two pixels c and d, affinity $\mu_\kappa(c,d)$ is a measure of local connectivity κ. This measure is defined based on the spatial adjacency, homogeneity and object components:

Spatial Adjacency Component $\boldsymbol{\mu_\alpha}$ is the function that incorporates the spatial closeness factor to the equation of connectivity. The adjacency component is calculated for two pixels c and d in the image domain as

$$\mu_\alpha(c,d) = \begin{cases} 1 & \text{iff } \sqrt{\sum_i (c_i - d_i)^2} \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

In other words, if pixels c and d are not adjacent in the image their affinity is equal to zero.

Object Feature Component μ_ϕ of affinity function is included to overcome the effect of slow intensity variation in adjacent regions, and in this thesis is expressed as

$$\mu_\phi(c,d) = \exp\left(-\frac{\max\{|f(c)-m_1|,|f(d)-m_1|\}^2}{2\,s_1^2}\right), \tag{5.2}$$

where m_1 and s_1 respectively represent the mean and standard deviation of the intensity values of the region. The functions $f(c)$ and $f(d)$ give the intensity value of the image at pixels c and d respectively.[1]

Homogeneity Component μ_ψ is introduced in the affinity function to overcome the slow intensity variations of the pixels in the same region and to measure the intensity gradient in the region. The homogeneity component is calculated as

$$\mu_\psi(c,d) = \exp\left(-\frac{(||f(c)-f(d)||-m_2)^2}{2\,s_2^2}\right), \tag{5.3}$$

where m_2 and s_2 are respectively mean and standard deviation of the intensity homogeneity of a region.

The ***affinity*** function is defined based on the above three components, where the first component reflects the spatial adjacency of the pixels, while the latter two characterise the "hanging togetherness" aspect of fuzzy connectedness. There exist several ways to compute fuzzy affinity from the above components [Ciesielski and Udupa 2010]. The affinity function for the present study is

$$\mu_\kappa(c,d) = \mu_\alpha(c,d)\sqrt{\mu_\psi(f(c),f(d)).\mu_\phi(f(c),f(d))}, \tag{5.4}$$

with μ_ϕ and μ_ψ defined respectively in (5.2) and (5.3).

Connectivity Map. For any two pixels c and d, the strength of connectedness $K(c,d)$ is computed by choosing the maximum affinity value among all the paths

[1]Function f represents pixel intensity in this application. In general, f can be an estimate of certain image properties (such as gradients and texture measures) [Udupa and Samarasekera 1996].

connecting pixels c and d. Each path p is generated from a set of points connecting the two pixels c and d, and the affinity of the path is set to the minimum of the affinities of each of the two consecutive points on the path. The strength of connectedness, denoted K, between two pixels c and d is defined as

$$K(c,d) = \max_{p} \min_{i} \mu_{\kappa}(e_i, e_{i+1}), \tag{5.5}$$

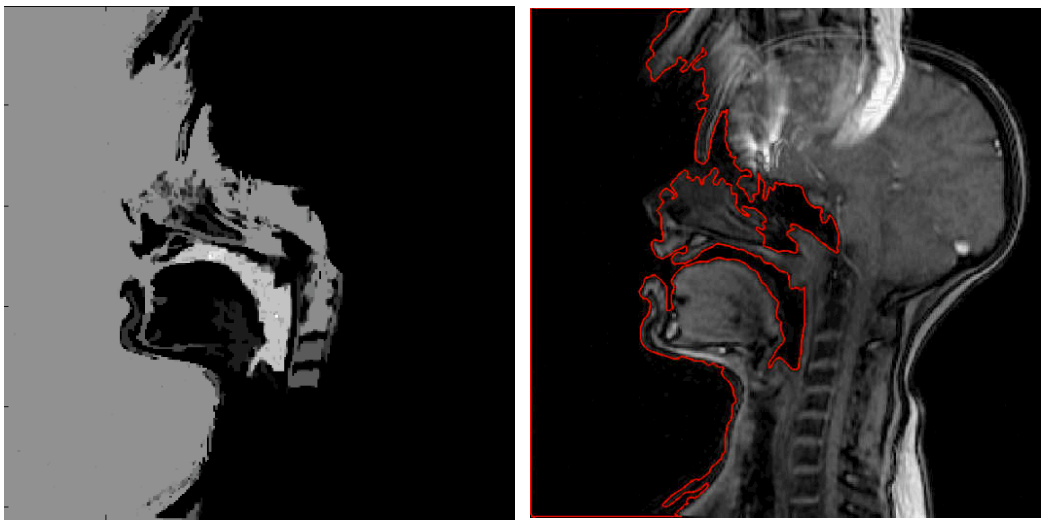
where (e_i, e_{i+1}), $1 \leq i \leq N$, are adjacent points on path $p \in P_{cd}$ of length N, where P_{cd} is the set of all paths from c to d, and μ_{κ} is the affinity function (Equation 5.4)).

The connectivity map is generated from the path between all the points in the image, and a threshold value T_{FC} is then applied to retrieve the object of interest from the image.

5.4.2 Application to vocal tract MRI data

The fuzzy connectedness technique, described above, was applied to MR images of a randomly chosen speaker from our database. The images were all standardised as described in chapter 3. First a few sample images were selected and the mean and standard deviation of intensity and gradient in the vocal tract region were estimated, by manually selecting a sample region inside the vocal tract (using MATLAB). For preliminary experiments, two seeds (i.e. pixels) were automatically placed in areas very close to the back wall of the pharynx and sublingual cavity. These areas were chosen as they usually have an intensity similar to the target object, and are less likely to be disturbed by movements of the articulators.

The preliminary results suggested that the algorithm *leaks* where the boundary is missing between the vocal tract and any area that has a similar intensity characteristics to the vocal tract cavity (basically air or bone areas, everywhere), particularly at the lips to the surrounding air. Figure 5.4 (a) shows the connectivity map and Figure 5.4 (b) shows the contours obtained after thresholding. The connectivity map divides the image into different regions, where the pixels that belong to the same region have approximately same intensity values. The leaking through



(a) connectivity map (b) contour after thresholding

Figure 5.4: FC segmentation connectivity map (a) is thresholded to obtain the vocal tract contour in (b). The lack of a boundary at the lips results in a region that grows out from vocal tract airway into the surrounding air (and into other areas connected to the surrounding air that share similar intensity values to the air, such as the nasal cavity).

the gaps mainly occurs at the opening of the lips where the vocal tract airway joins the surrounding air, and the velum, where the oral cavity joins the nasal cavity. To overcome the leakage to the surrounding air, we applied a filtering technique using speaker's profile, described below.

Problem of leaking, and forcing closure. To block the connection at the lips to separate the vocal tract from the surrounding air, we used the speaker's head profile to mask the rest of the image to a different grey value. The speaker profile was obtained with MATLAB, using the *imclose* command that performs morphological closing on the image, followed by the *imfill* command to set the intensity of the pixels inside the closed boundary to a uniform intensity value. Figure 5.5 shows an

example of speaker profile masking of the image.

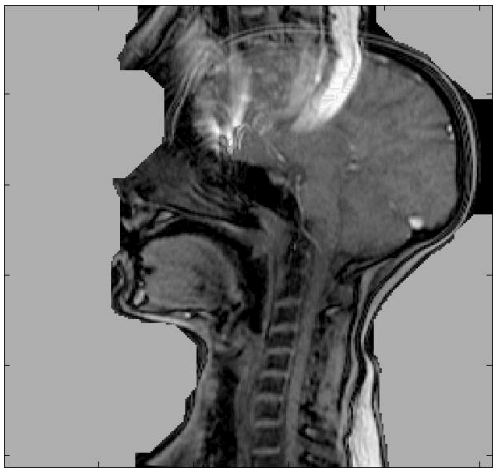
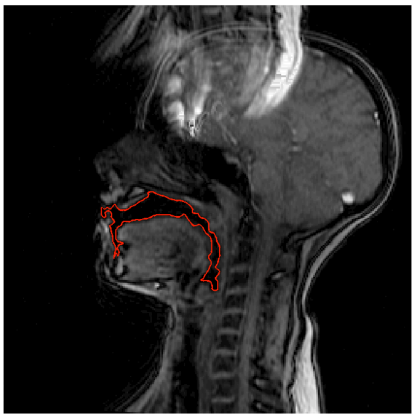


Figure 5.5: The speaker profile is used to mask the image and place a boundary at the lips to separate the vocal tract airway from the surrounding air.

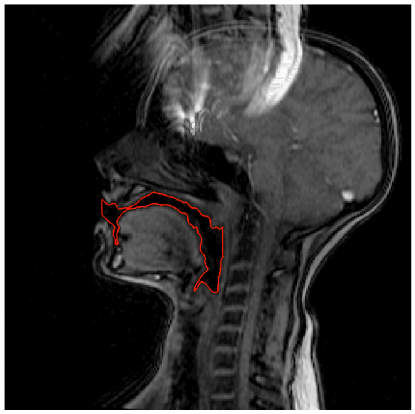
Masking the images at the lips successfully addressed the problem of leaking from the front of the cavity to the surrounding air. Figure 5.6 illustrates examples of FC segmentation applied on the masked images (Figure 5.6 (b) is the same MR image in Figure 5.4 (b)).

Although closing the front of the mouth by masking successfully stops the contours leaking though the gap at the lips, unfortunately, this is not the only area with missing boundary information along the vocal tract. A closer look at the images in Figure 5.6 confirms that in all of these MR images, the velum has closed the entrance to the nasal cavity. However, if the velum is lowered, the region will grow into the nasal cavity and to the areas in the brain that are not captured in the MR images due to the density of protons in the tissue.

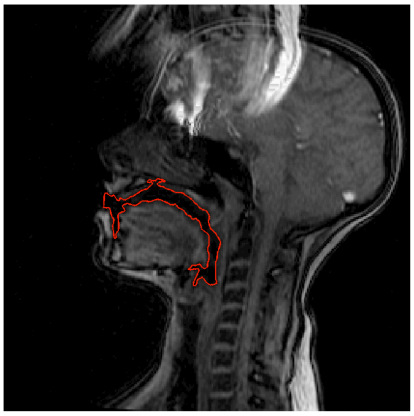
Figure 5.7 shows a few examples of leaking through the opening at the velum. It is worth mentioning that the leakage through the velum opening is not a problem per se; the nasal cavity is important in articulation and in theory the vocal tract must be extended all the way to the nostrils. However, due to the tissue texture and nature of the images, the leakage at the velum opening extends into the brain and fails to extend correctly to the nose. Unlike the opening of the mouth, the



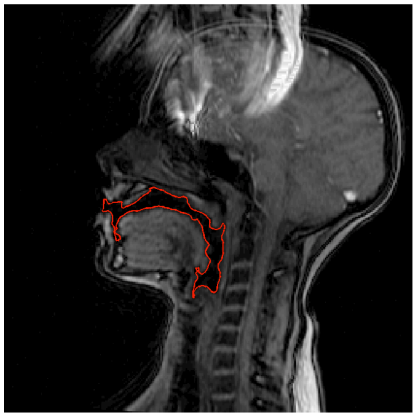
(a)



(b)



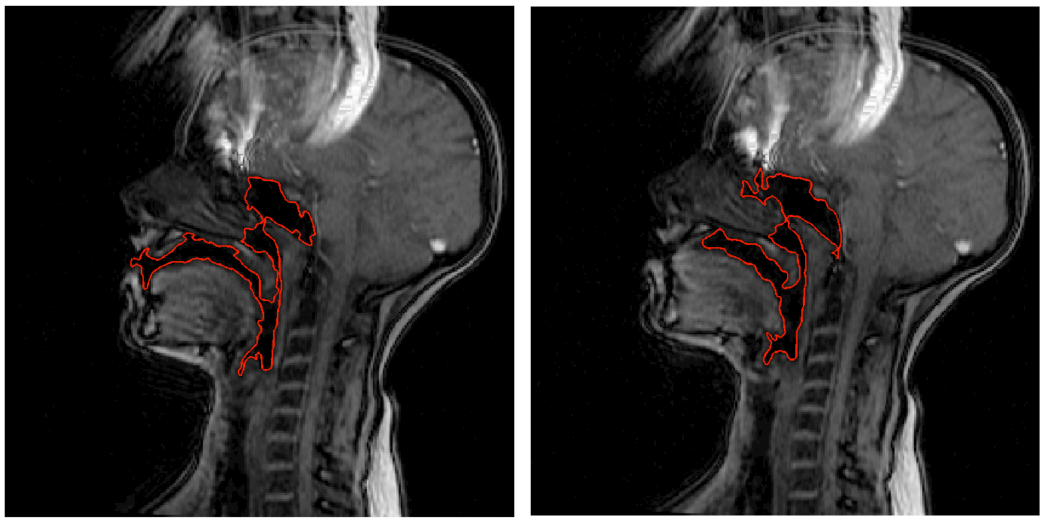
(c)



(c)

Figure 5.6: FC segmentation contours on MR images masked with speaker profile. The contour does not leak through the mouth opening to the surrounding air and the boundary of the vocal tract at the lips is fairly accurate.

entrance to the nasal cavity cannot be easily masked. The reason is that speaker profile is constant in all the image and does not change over the range of images in a sequence, whereas, the velum moves constantly during articulation of an utterance and consequently cannot be masked easily.



(a) (b)

Figure 5.7: FC segmentation examples where the region grows out of the vocal tract into the nasal cavity and penetrates into the brain region because of the missing a boundary at the velum opening.

The problem of leaking was also observed anywhere in the vocal tract where the difference in the tissue texture leads to a discontinuity in the boundary of the cavity, such as the thin palate at the roof of the mouth. Note that this does not necessarily occur in all the images as in some the palatal wall is more clearly captured. Figure 5.8 shows an example of leakage to the nasal cavity and brain through the missing boundary information at the palate.

5.4.3 Discussion

In previous sections, we demonstrated that a purely region-based segmentation, without any prior shape information, may not be a suitable choice for extracting vocal tract shapes. This is partly due to the specifics of the vocal tract where there are no physical boundaries between the tract airway and nasal cavity or surrounding

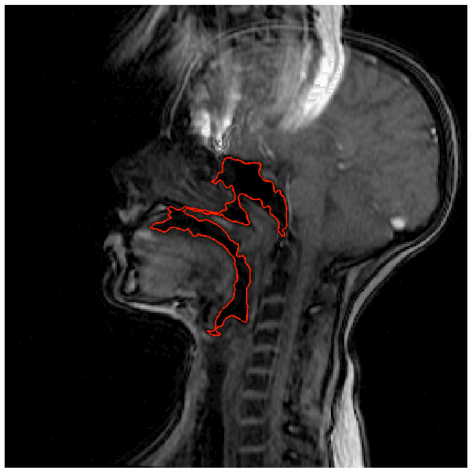


Figure 5.8: Example of leakage through a missing boundary segment at the palate.

air at the lips opening. The contours of the region-based segmentation can leak through these openings and grow out into other parts of the image. The leakage can also occur anywhere along the vocal tract where there is a gap in the boundary that connects the region to areas with similar intensity to the oral cavity.

An example is the hard palate, where the tissue layer is very thin and may appear very blurred in the image because of movement. Although the leakage can be avoided by using a blockage at more static places or where the profile is known, it cannot always be stopped where there are constant movements, such as the case of velum opening.

5.5 Conclusion

In this chapter, we discussed the existing challenges of extracting the vocal tract shape from images of a dynamic MRI database of speech articulation. We briefly introduced the segmentation techniques we shall use for segmentation.

We demonstrated that a purely region-based segmentation may not be a suitable choice for automatic segmentation of vocal tract images by applying fuzzy connectedness segmentation to our data. The results revealed that using FC segmentation, the boundaries of vocal tract *leaked* through lips or velum openings. We attempted to resolve the problem by blocking the mouth opening, a method traditionally used in vocal tract image analysis applications [Story *et al.* 1996]. But the results demonstrated that although this effectively addresses leakage through lips, leakage through any gaps in the boundary of the vocal tract, such as velum opening or a break in continuity of the boundary at the palate, remained an unresolved issue.

In the next chapter, inspired by the results from fuzzy connectedness segmentations, we propose applying a method that takes both the intensity features and the statistics of the shape model into account when segmenting vocal tract images. The method constrains the shape where boundaries are missing, and delineates with fine details where blurring occurs around the edges: the OASM framework.

Chapter 6

OASM-based Framework for Automatic Segmentation

6.1 Introduction

We concluded the previous chapter by demonstrating the leakage issues encountered when applying region-based methods to vocal tract segmentation. To remedy this problem, in this chapter we introduce a new framework for automatic segmentation of vocal tract images.

Oriented Active Shape Models (OASM) [Liu and Udupa 2009] can address requirements for vocal tract image segmentation by including both the shape statistics and intensity information in detecting the object boundary.

Similar to ASM, OASM requires a set of training images that describe the shape in the form of landmarks, to create a shape model representing the average shape of the object of interest and its variability. Manual positioning of landmarks is a labour intensive and time-consuming task, not to mention its vulnerability to errors and subjective judgments. Therefore, automatic approaches for tagging the landmarks on the training shape are more desirable.

We use the recursive boundary subdivision (RBS) landmark tagging approach [Rueda and Udupa 2009] to generate the initial training landmark sets. RBS landmark tag-

ging mathematically defines and characterises landmarks and ensures homology and correspondence by consulting all the training shapes in the training set simultaneously for selecting the landmarks. It uses fundamental attributes of shapes in finding the landmarks, and thus includes the variability in the training shapes. In addition, it is computationally simple and does not require initial registration of the shapes, and therefore is suitable for this application.

In this chapter, we propose a new framework for automatic segmentation of vocal tract MR images using RBS automatic landmark tagging and OASM shape extraction. The new framework does not require intensive human effort for tagging the landmarks and incorporates both image intensity and model information to find the optimal boundary in new images.

We continue this chapter by presenting an overview of the OASM-based framework for segmenting vocal tract images in section 6.2. The RBS landmark tagging approach and its application to our vocal tract shapes are explained in section 6.3. Section 6.4 describes the OASM segmentation method in detail. We finish this chapter by presenting the application of OASM to our database, followed by discussing the various settings applied to adapt OASM to this application (section 6.5).

6.2 Overview of OASM-based framework

Our proposed framework consists of two key stages (a) landmark tagging and (b) OASM training and segmentation. The flowchart in Figure 6.1 shows the overall structure of the framework. In stage (a), steps 1 and 2 in the flowchart, first the vocal tract shape is segmented in a set of training images. In this application, we use live wire user-steered segmentation in 3Dviewnix [Udupa *et al.* 1993] to segment the images. RBS automatic landmark tagging was then applied to the training data to define and characterise a set of landmark points for the training shapes. The training set is thus converted from a set of contours to sets of landmarks to describe the shape.

Stage (b) of the framework, corresponding to the application of OASM, involves two main phases: training and segmentation. We briefly describe each phase.

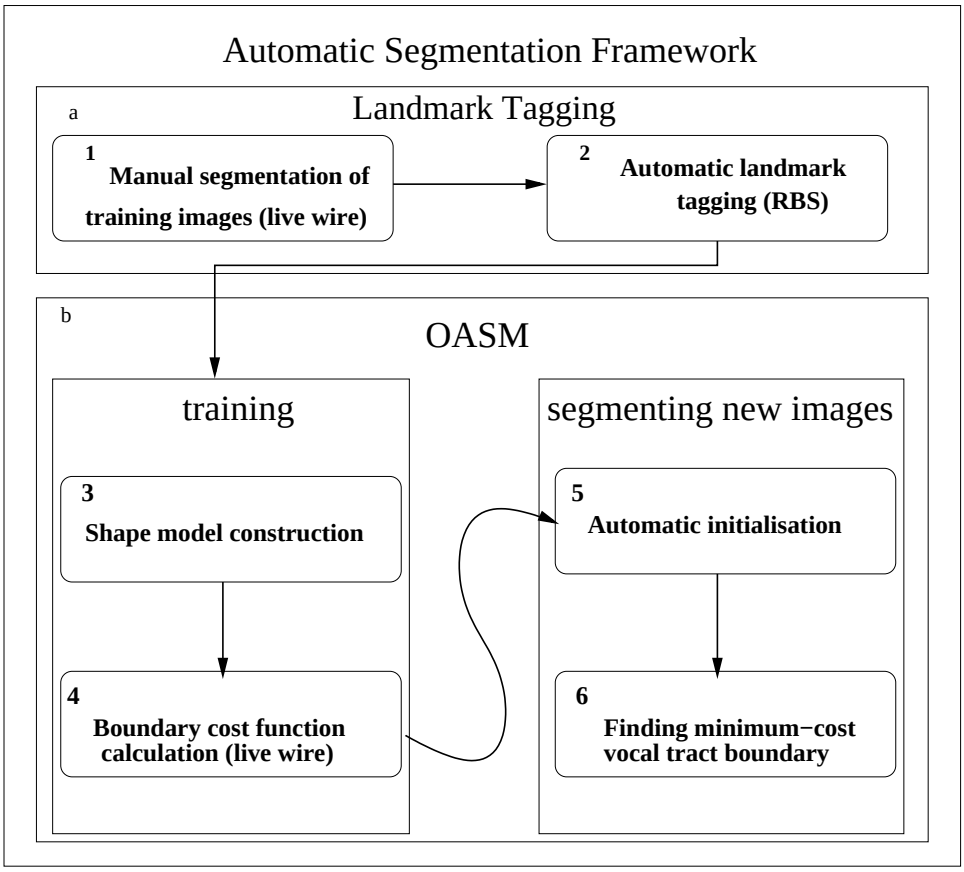


Figure 6.1: The flowchart of different steps involved in vocal tract segmentation with RBS and OASM.

The OASM training phase, referred to as training and model creation in [Liu and Udupa 2009], consists of steps 3 and 4 in the flowchart of Figure 6.1. In step 3, the training set is used to train and construct an ASM shape model (ASM training stage). Live wire is then used in step 4 to create a cost function based on the intensity and orientation information of the shape boundaries in the set of training images, i.e. live wire training step. An OASM is constructed from the generated shape model in step 3 together with the trained live wire cost function in step 4.

The OASM segmentation phase, searching and segmentation [Liu and Udupa 2009], consists of steps 5 and 6 in the flowchart of Figure 6.1. The constructed OASM is used to do a coarse recognition of the boundary in the domain of the image to be segmented, corresponding to step 5. The result of automatic initialisation is a

shape instance, that is subsequently used as the input to step 6: finding the globally optimal oriented vocal tract boundary. In this step, the shape instance is deformed to best fit the object properties and find the optimal boundary. This involves fine recognition and delineation using a two-level dynamic programming.

Next, we explain the RBS tagging approach used for labelling the training set, followed by an overview of the OASM segmentation algorithm and its implementation.

6.3 RBS automatic landmark tagging

Automatic landmark tagging approaches are typically completed in two stages; placing the landmarks on each individual image, and establishing the correspondence between landmarks across the set of images. The two stages are independent and do not necessarily have the same order in different methods. In one family of automatic landmark tagging approaches, the points are first locally selected on each individual image, and the correspondence between the points is established in a global manner (local-to-global) [Rohr 1999; Frangi *et al.* 2001; Davies *et al.* 2001; Walker *et al.* 2002; Thodberg 2003; Thodberg and Olafsdottir 2003; Souza and Udupa 2005]. In another category of automatic landmark tagging approaches, first the global operations are performed to extract the corresponding points and then the landmarks are localised for each single image (global-to-local) [Rueckert *et al.* 2001; Twining *et al.* 2002; Marsland and Twining 2003; Vaillant *et al.* 2004; Cootes *et al.* 2008].

The vocal tract shape constantly deforms along the sequence of frames, and therefore we need a method of landmark tagging that can find the corresponding landmarks even when the features move or the shape of the vocal tract in different frames is not exactly the same. We therefore chose a global-to-local approach for automatic landmark tagging, explained in the following subsection.

In this section we review the RBS tagging approach as described in [Rueda and Udupa 2011]. We illustrate its application to vocal tract contour landmark tagging through images depicting every step of the algorithm. Note that RBS is applied

on a set of *shapes* (contours of the objects) not the actual images. The vocal tract contours used for illustrations (and for training the OASM) are obtained with user-steered boundary tracking of images using live wire as demonstrated in section 5.3.3.

The RBS approach recursively and simultaneously subdivides the boundaries in a target set of shapes, and mathematically finds and characterises the landmarks on the subdivided segments. The recursive subdivision of the boundaries is a hierarchical procedure performed on all training shapes simultaneously. At each recursive step, the correspondence between the boundaries and subdivided segments is maintained using a specified similarity criterion, thus ensuring the homology among landmarks.

In 2D images, the first two segments on each shape boundary are defined by the two initial landmarks found mathematically. The line connecting the two landmarks divides the continuous boundary into two connected segments. A new landmark position is next estimated on each of the new subdivided segments. New pairs of points are defined using the newly found landmark and the previous landmark points. The lines defined by the new pairs, further subdivide the boundary segments into subsegments. This procedure of recursive subdivision is hierarchical and the correspondence between all the new subdivided segments in the training set is maintained. To select the new landmarks all the training shapes in the set are consulted to guarantee the homology between landmarks. The process terminates when no dividing lines can be created, according to a pre-determined criterion, to further divide the segment to subsegments. This process continues until the criterion is met in all of the corresponding segments over all the shapes. In the following paragraphs, we explain each step with mathematical details and illustrate them using the vocal tract shape examples.

For this application, principal components analysis (PCA) [Rueda and Udupa 2011] is used to find the initial two landmarks, to divide the boundaries recursively, and to find new landmarks. Each boundary/boundary segment B is defined in terms of a set of coordinates $(x_1, y_1, ..., x_n, y_n)$. PCA is used to define the bound-

ary/boundary segment B in terms of a geometric centroid $C = (\bar{x}, \bar{y})$, representing the mean location of the boundary coordinates, and the two dimensions where the variation in the shape is maximum. These two dimensions are determined using the eigenvalues and eigenvectors of the covariance matrix of the points in the shape boundary. We refer to the first two eigenvalues as $\lambda 1$ and $\lambda 2$, and to the corresponding eigenvectors as v_1 and v_2. These parameters define new axes with the geometric centroid C as the centre, and the two eigenvectors which define the two principal axes. The first eigenvector v_1 continued on both sides of the centre represents the major inertia axis of the boundary points, and v_2, continued on both sides of the centre, represents the second inertia axis.

The two initial landmark points found on the boundaries are the result of the outermost intersection of the first principal axis of the shape with the boundary B_i, $i = 1, ..., M$ where M is the size of the set of shapes. Figure 6.2 shows the process of finding the first two landmark points in a sample vocal tract boundary.

The eigenvector corresponding to the first eigenvalue does not necessarily have the same orientation in all of the images. To resolve such inconsistencies, the orientation of the first eigenvector in the first shape of the set is used as a reference

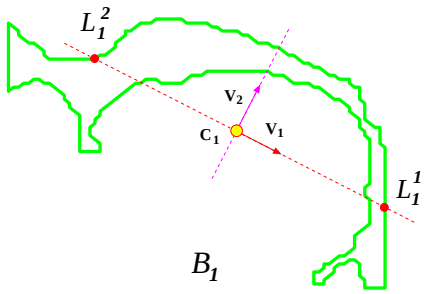
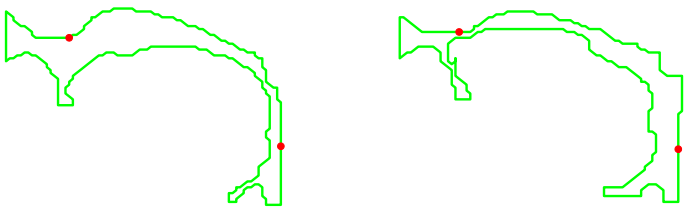


Figure 6.2: The first two landmarks L_i^1 and L_i^2, estimated by PCA for the boundary B_i. The two directions corresponding to maximum variations in the shape are depicted respectively by $\{\mathrm{v}_1, \mathrm{v}_2\}$ in each boundary B_i. The initial landmarks are found by extending the first eigenvector in both directions to meet the boundary of the shape (outermost intersection). The yellow circle represents the centroid of the shape.

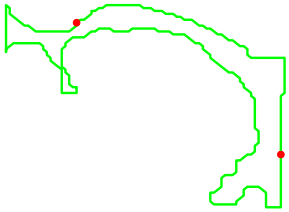
to find the orientation of all the other first eigenvectors in the rest of the shapes. This is important for guaranteeing the correspondence of the landmarks. Figure 6.3 shows the initial two landmarks obtained with PCA in a set of training images of the vocal tract.

The initial landmark points L_i^1 and L_i^2 create a line that divides each boundary B_i into two connected segments B_i^1 and B_i^2 (Figure 6.4). This time, PCA is used to find a single new landmark on each of the new segments. This time the direction of the second principal axis (second eigenvector v_2) is used to find the new landmark L_i^3 on the boundary segment B_i^1. Figure 6.5 illustrates the process of finding the third landmark point on the sample boundary of Figure 6.2.

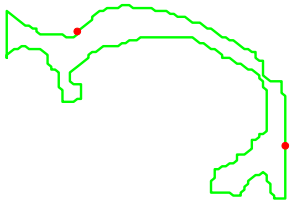


(a) B_1

(b) B_2



(c) B_3



(d) B_4

Figure 6.3: The initial landmarks estimated for four different vocal tract contours using PCA.

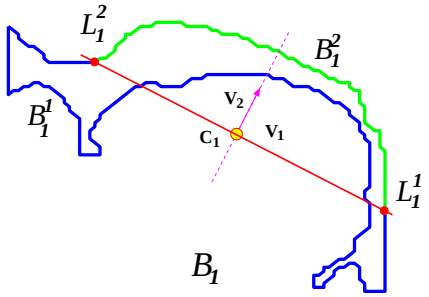


Figure 6.4: The line connecting landmarks L_i^1 and L_i^2 divides the boundary into two connected segments B_i^1 (blue contour) and B_i^2 (green contour).

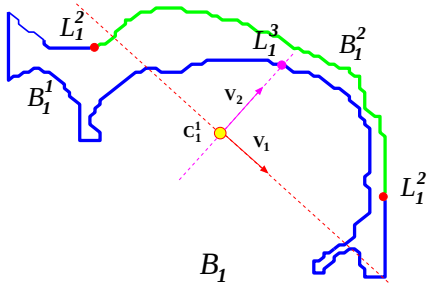


Figure 6.5: A new landmark point L_i^3 is found on boundary segments B_i^1 (blue contour). PCA is applied on the boundary segments B_i^1 to find the second eigenvector, v_2. The point where the second principal axis meets the boundary is chosen as the new landmark point.

The new landmark couples with each of the previous landmark points to form new pairs $L_i^1 L_i^3$ and $L_i^3 L_i^2$. The new pairs are subsequently used to divide the segments to two new subsegments B_i^{11} and B_i^{12} (Figure 6.6).

The rest of the landmarks are recursively found by dividing the new segments into subsegments and calculating where the second principal axis of the new subsegments meets the boundary. The process is terminated when a stopping criterion is met. In this thesis, the stopping criterion is defined by a parameter δ. At each recursion, the division of boundaries continues only if one or more than one of the λ_2 values

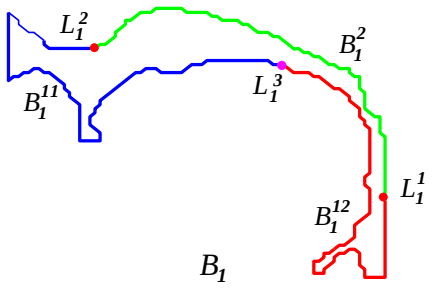


Figure 6.6: The new landmark L_i^3 divides the boundary segment B_i^1 to two connected subsegments B_i^{11} (blue) and B_i^{12} (red).

of the new segments are greater than δ. The value of δ is a positive real number ($\delta > 0$), where greater values of δ achieve smaller number of landmarks.

The final set of landmarks automatically found for the vocal tract contours of multiple speakers and one speaker are respectively presented in Figures 6.7 and 6.8. In both sets it can be observed that although the majority of landmarks correspond across the images, some inconsistencies exist around particular areas such as the opening of the lips or the sub-lingual cavity. This is mainly due to the extreme variability existing in the vocal tract shapes across speakers and within phonemes. Also, the procedure of finding and selecting landmarks in RBS is based on distance and variation criteria. As it is not based on the anatomical structure, the opening of the lips is not considered a separate entity by RBS to expand or shrink (it has no anatomical landmarks). The landmarks on the parting of the lips may be positioned on the lips in another image. This issue can be resolved to a good extent by tuning the number of landmarks, by setting δ to different values, to make the algorithm select more prominent landmarks. In the illustrations in Figures 6.7 and 6.8, we have set $\delta = 1$.

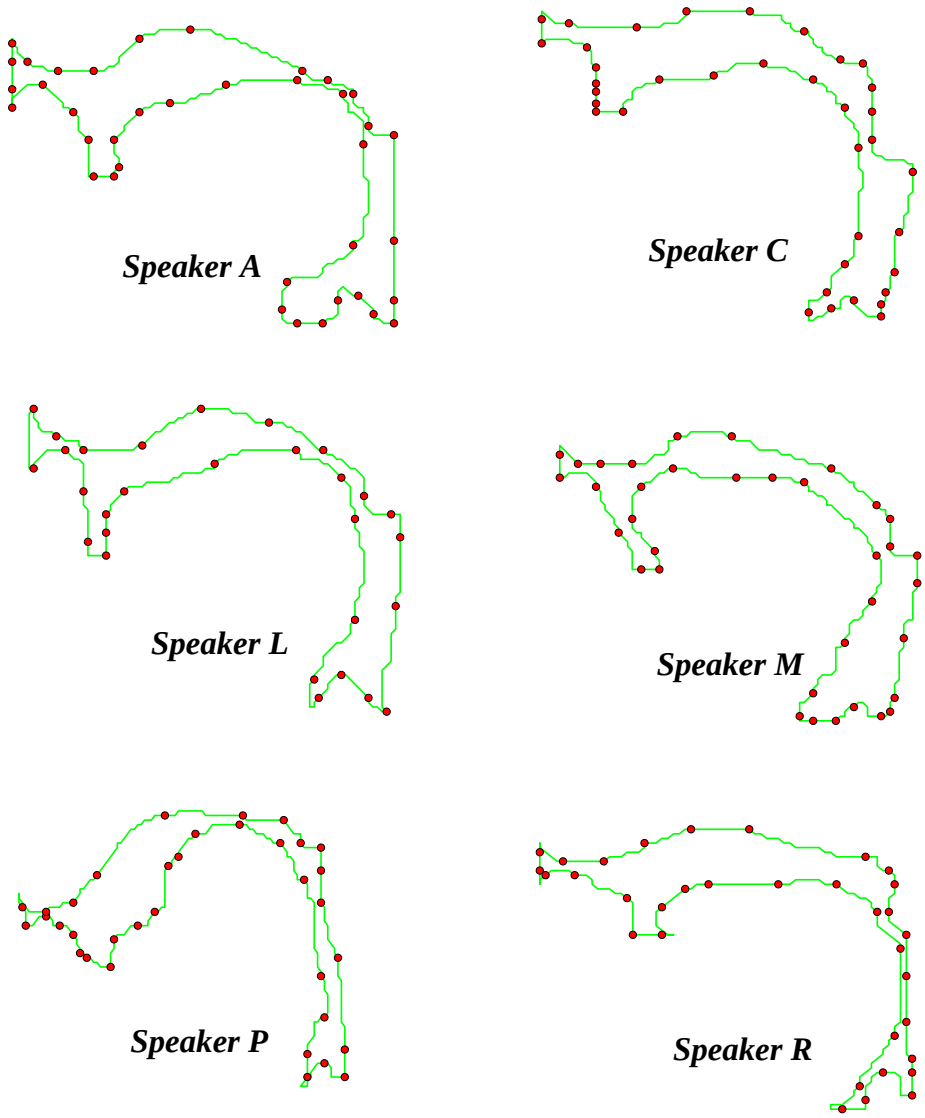


Figure 6.7: Landmarks found on vocal tract contours of different speakers. All the shapes correspond to the articulation of phoneme /ɔ/ ($\delta = 1$). Note the variability in the vocal tract shapes.

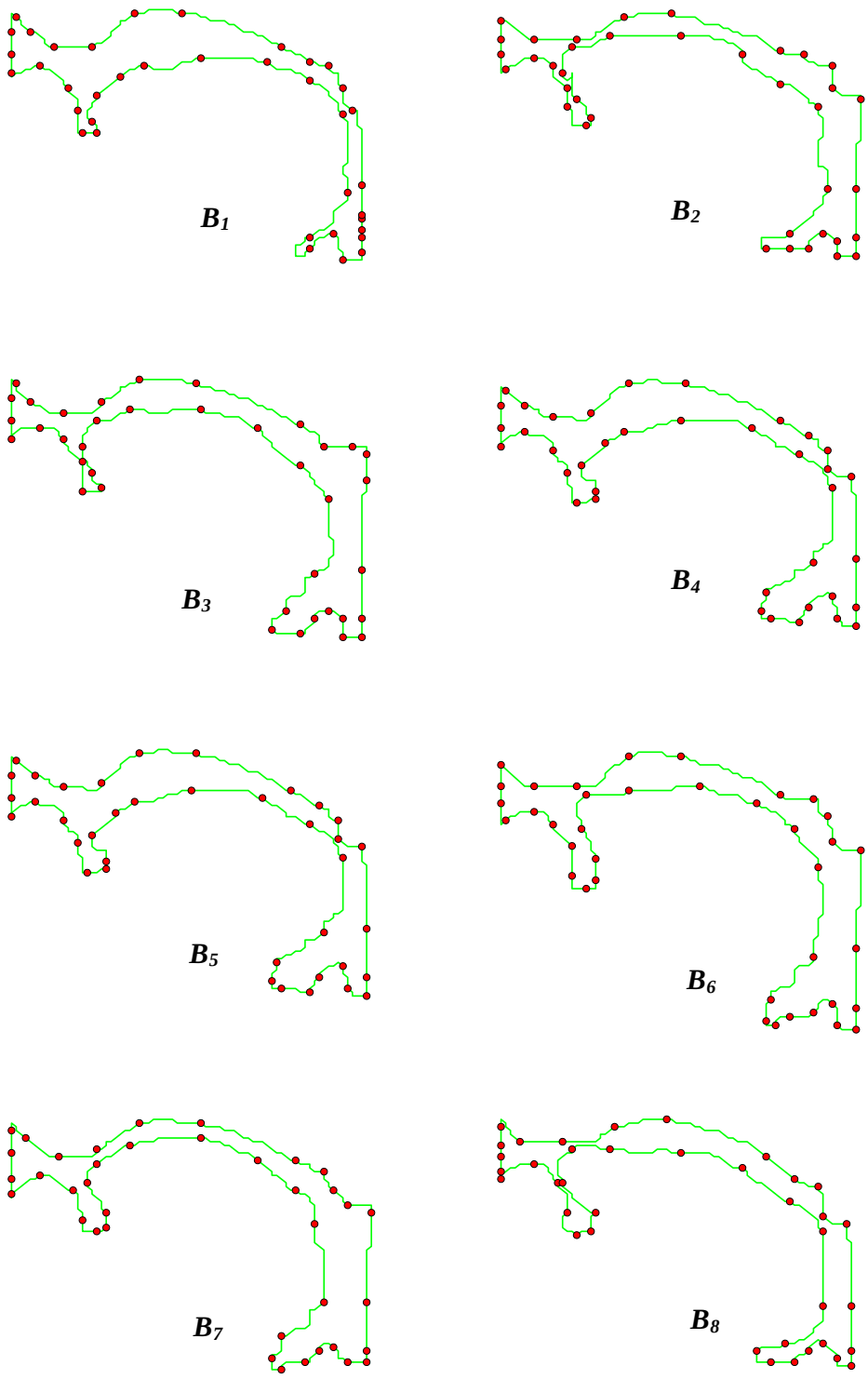


Figure 6.8: Landmarks found on a selection of vocal tract contours of speaker A ($\delta = 1$). The contours $\{B_1, B_2, B_3, B_4, B_5, B_6, B_7, B_8\}$ respectively correspond to the following phonemes: {/ɑ/, /s/, /ə/, /ʌ/, /f/, /d/, /ɪ/, /n/}.

6.4 Oriented active shape models

The main steps involved in the OASM algorithm, in this application, are presented below. We explain each step in detail.

Structure of the OASM algorithm

---

1. Training
   (a) Shape model construction: ASM training.
   (b) Boundary cost function calculation: live wire training.
2. Segmentation (of new images)
   (a) Coarse recognition: Automatic initialisation.
   (b) Fine recognition and delineation: Finding the best oriented boundary.

Shape model construction. Like ASM, OASM involves an initial training step, where a shape model is generated using a set of N training shapes. Each shape $\mathbf{x}$ is represented by a set of n landmarks

$$\mathbf{x} = (x^1, x^2, ..., x^n). \tag{6.1}$$

Note that $\mathbf{x}$, bold lower case x, notation is used to represent the shape, and italic lower case x^i refers to landmark i, $i = 1, ..., n$.

To generate a generic shape model M, the variations between the shapes caused by scaling, rotation, and location are discarded by performing scaling, rotation and translation, and aligning all of the shapes in the training set together (procrustes).

ASM then generates a shape model on the aligned training data by applying PCA. The resulting model consists of a mean shape $\bar{\mathbf{x}}$, the covariance matrix S, a set of eigenvalues λ_j of the covariance matrix, and the corresponding eigenvectors ϕ_j.

ASM retains a few modes of variation that cover most of the variation in the shapes within the training set. Typically t, the number of variation modes (eigenvalues), is decided based on the the percentage of the variance of the training shapes retained within the shape model (e.g. 90% – 95%).

The shape instance $\mathbf{x}$ can then be reconstructed by deforming the mean shape $\bar{\mathbf{x}}$ by a linear combination of the retained eigenvectors as

$$\mathbf{x} \approx \bar{\mathbf{x}} + \Phi\mathbf{b}, \tag{6.2}$$

where $\mathbf{b} = (\beta_1, \beta_2, ..., \beta_t)^T$ represents a vector of weights for the matrix Φ of retained eigenvectors $(\phi_1|\phi_2|...|\phi_t)$. The weights β_j, $1 \leq j \leq t$ are typically chosen to allow the variation in the shape within a suitable range constrained by the shape model

$$-3\sqrt{\lambda}_j < \beta_j < +3\sqrt{\lambda}_j. \tag{6.3}$$

The pixel intensity profiles of the points around each landmark on the mean shape model are collected to estimate the structure of the image around that landmark. The derivatives of the intensity grey-level values of a few pixels around each landmark (on a line perpendicular to the shape model) are calculated and normalised to obtain an intensity profile value g_p^i for the landmark x^i, $1 \leq i \leq n$ in shape instance $\mathbf{x}_p$. The mean intensity profile $\bar{g}^i$ and the covariance matrix S_g^i for each landmark are then computed over the range of the images in the training set. The Mahalanobis distance metric D is used to find the intensity profile of the segmented shape instance $\mathbf{x}_s$ in the image that matches the shape model best by minimising the distance

$$D(x^i) = (g_s^i - \bar{g}^i)^T (S_g^i)^{-1} (g_s^i - \bar{g}^i)^T, \tag{6.4}$$

where g_s^i is the intensity profile value of landmark point x^i in the segmented shape $\mathbf{x}_s$.

This distance metric is used in ASM to find the best segmented shape in the image. The OASM approach combines the above distance metric from ASM with the oriented boundary cost structure to generate an OASM.

Oriented boundary cost function. The cost of each shape is calculated by introducing a new entity, *bel* (boundary element), as the constructing unit of a boundary. A bel $b = (p, q)$ has an orientation and a location parameter, where the location is the common edge of p and q, and the *orientation* is the direction a bel can take so that pixel p is situated inside the boundary and pixel q falls outside the boundary of the shape (Figure 6.9). A set of features assigned to each bel are used to estimate the probability of the bel being part of the boundary. The features include intensity on the immediate interior of the boundary, intensity on the immediate exterior of the boundary, and different gradient magnitudes at the centre of the bel. The cost function of a bel $c(b)$ is the weighted sum of these features.

The costs of all the bels of a boundary are accumulated to find the total cost value. Hence, the problem of finding the boundary with the smallest cost is addressed by finding the best set of bels for the boundary that minimises the total cost estimate of the shape. A dynamic programming approach is implemented in OASM with the goal of finding the minimum-cost path in a directed graph, where the boundary pixels are the vertices and the bels are the edges of the graph.

OASM uses live wire to find the minimum cost path between any two successive landmark points on a shape instance $\mathbf{x} = \{x^1, x^2, ...x^n\}$ from the model M. The cost

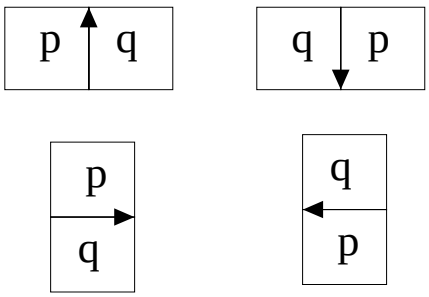


Figure 6.9: The four possible types of boundary elements (bels) based on the orientation of the pixel edge on the boundary of the object.

$\kappa(x^i, x^{i+1})$ is the minimum cost between two successive landmark points x^i and x^{i+1}, obtained by summing the cost of the bels on the best oriented path $< b_1, b_2, ..., b_h >$ connecting the two points

$$\kappa(x^i, x^{i+1}) = \sum_{l=1}^{h} c(b_l), \tag{6.5}$$

where h is the number of bels on the oriented path. The cost associated with the boundary instance $\mathbf{x}$ can be expressed as

$$K(\mathbf{x}) = \sum_{i=1}^{n} D(x^i)\kappa(x^i, x^{i+1})D(x^{i+1}), \tag{6.6}$$

where $D(x^i)$ is the Mahalanobis distance of the intensity profile of landmark point x^i. An OASM can now be defined by a pair (M, K), M being the statistical shape model and K being the oriented boundary cost function associated to model M. In the recognition step, OASM assigns a cost structure K to the shape model M, where every shape instance $\mathbf{x}$ generated by model M through statistical variations has a determined cost value $K(\mathbf{x})$. The cost structure is defined by calculating the cost of the live wire segments between every two successive landmark points in the shape instance $\mathbf{x}$. The shape instance with the smallest cost, $\mathbf{x}_o$, is selected as the closest shape to the original boundary in the images to be further delineated.

Coarse recognition. The recognition step, also referred to as *automatic initialisation* step, looks for a shape in the image I that is closest to the actual position of the object in the image. A pseudo-exhaustive search is performed across the image domain to find the shape instance with minimum boundary cost. The assumption is that the cost of an oriented boundary at a pose (i.e. location, scale, and rotation) closer to the actual shape is indeed smaller than the cost of the oriented boundary at other poses. Note that in the original implementation of the algorithm, and in this application, the search by the pose is only based on the location, by moving the geometric centre of the shape model across the image to find the best position. The

size of the window in the image domain to be searched and the sampling rate of the geometric centre points affect the computational cost of the automatic initialisation. For example, in a 256×256 pixels image, if a window of size 156×156 is considered as the search range (excluding the 20% left-most, top-most, bottom-most and right-most pixels in the image), with uniform sampling of the centre points at intervals equal to 5% of the number of pixels in a row/column, approximately 170 object positions must be searched. Although this approach has been reported to be efficient in finding the optimal boundary in clinical applications where the orientation and scale variation is small due to a predefined imaging protocol, a location-based search may fail in applications where variation in the orientation and scale is large and not negligible [Liu and Udupa 2009]. In such scenarios, the recognition can be enhanced by either reducing the amount of variation in orientation, i.e. categorising the shape model classes, or by limiting the search space and consequently restricting the coarse search (location-based) and focusing on fine search. We discuss this further in subsection 6.5.2.

Fine recognition and delineation. The final step of the OASM technique is finding the optimal boundary of the shape. That is, assuming that the initialised boundary $\mathbf{x}_{init}$ is close enough to the actual boundary, new landmark positions are found with the objective of minimising the sum of the costs of the oriented boundary paths between new landmarks. To do this, first a set of neighbouring pixels around each landmark point x^i, $1 \leq i \leq n$ from the shape $\mathbf{x}_{init}$ are determined by running a line perpendicular to the boundary at x^i and sampling at $2 \times m$ points $P^i = (p^1, ..., p^m, x^k, p^{m+2}, ..., p^{2m+1})$, on this line ($m > l$, where l is the number of points used in profile estimation for ASM training). The algorithm next finds the minimum cost paths between each of the points in sets P^i and P^{i+1} (corresponding to two consecutive landmarks x^i and x^{i+1}), referred to as fine recognition. The problem of finding the optimum boundary then becomes the problem of finding the set of minimum paths between successive landmarks so that the final closed boundary is

continuous and the total cost of the boundary is minimum (delineation). This is achieved by redefining the problem in terms of a graph search problem, comprising all the possible landmarks determined in the last step as vertices. The arcs of the graphs represent the paths between the points in P^i and P^{i+1}, and each has an associated value as the cost of the path. A two-level dynamic programming algorithm was implemented to find the path with the smallest possible total cost, between a point p^i in P^1 and the same node in P^{N+1} where $P^{N+1} = P^1$.

The above procedure of delineation continues until a convergence criterion is met. Here the criterion is the maximum distance between the corresponding landmark points of the obtained oriented boundaries during two consecutive executions of delineation, that must be smaller than a threshold.

6.5 Application to MRI database

In this section we explain the data and settings for training and testing the proposed approach. We used the MR images of six different subjects (four females and two males) for these experiments. For each speaker, two sets of images were selected separately for training and testing purposes. While for training the images were chosen to include maximum variety in terms of articulation manner, the test set was *randomly* sampled from the speakers' image sets excluding the training images. For each speaker, 12 training images were used, one for every phoneme in the database (refer to chapter 3 and 4 for a description of the database and available phonemes).[1] The same number of images were set aside for evaluating the performance. The initial training images were semi-automatically segmented by using the user-steered segmentation based on live wire in the 3Dviewnix software [Udupa *et al.* 1993]. Figure 6.10 shows an example of user-steered segmentation.

The vocal tract user-steered segmentations (examples in Figure 6.11) were then used as the input for the RBS automatic landmark tagging approach described in

[1]For some of the speakers, only 11 distinct phonemes were available in the database, therefore we selected the 12th image in the training set to be another image /ə/ to have uniform tube shape that bring less distortion to the image structure.

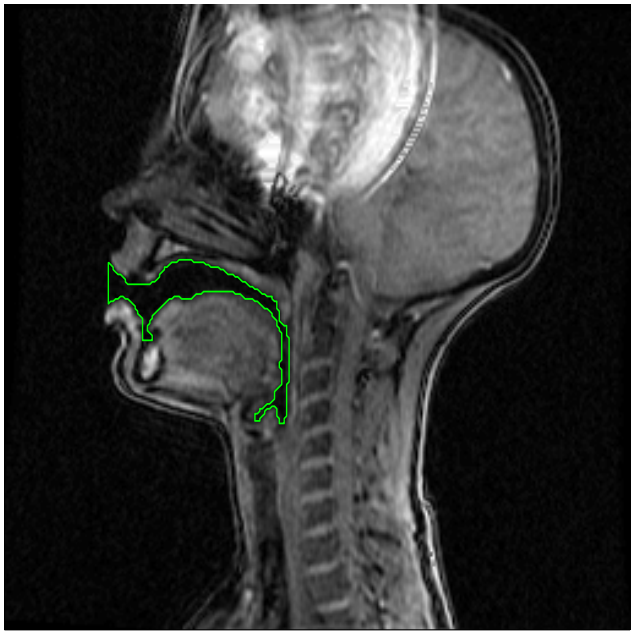


Figure 6.10: User-steered segmentation of the vocal tract shape with live wire using 3Dviewnix software [Udupa *et al.* 1993].

section 6.3. Different δ values were selected to extract different sets of landmarks.

OASM was trained using the generated sets of landmarks and their corresponding MR images. The ASM shape models were trained to include sufficient number of modes to embed 90% of variation in the original vocal tract shapes.

A study of preliminary segmentation results revealed a few important factors that have crucial roles in performance of the proposed framework in this application. These factors largely have roots in the specifications of the dynamic MRI databases and vocal tract shape. A key factor is the extent of shape variability present in the database, in both training and test sets. The variation is caused by both articulation differences and anatomical differences of the speakers' vocal tracts. Ideally, we need a general model that contains all the variation possible in real cases; however, this is a very complex problem due to the extreme variation of shape structure within phonemes and across speakers. A single OASM may not be sufficient to

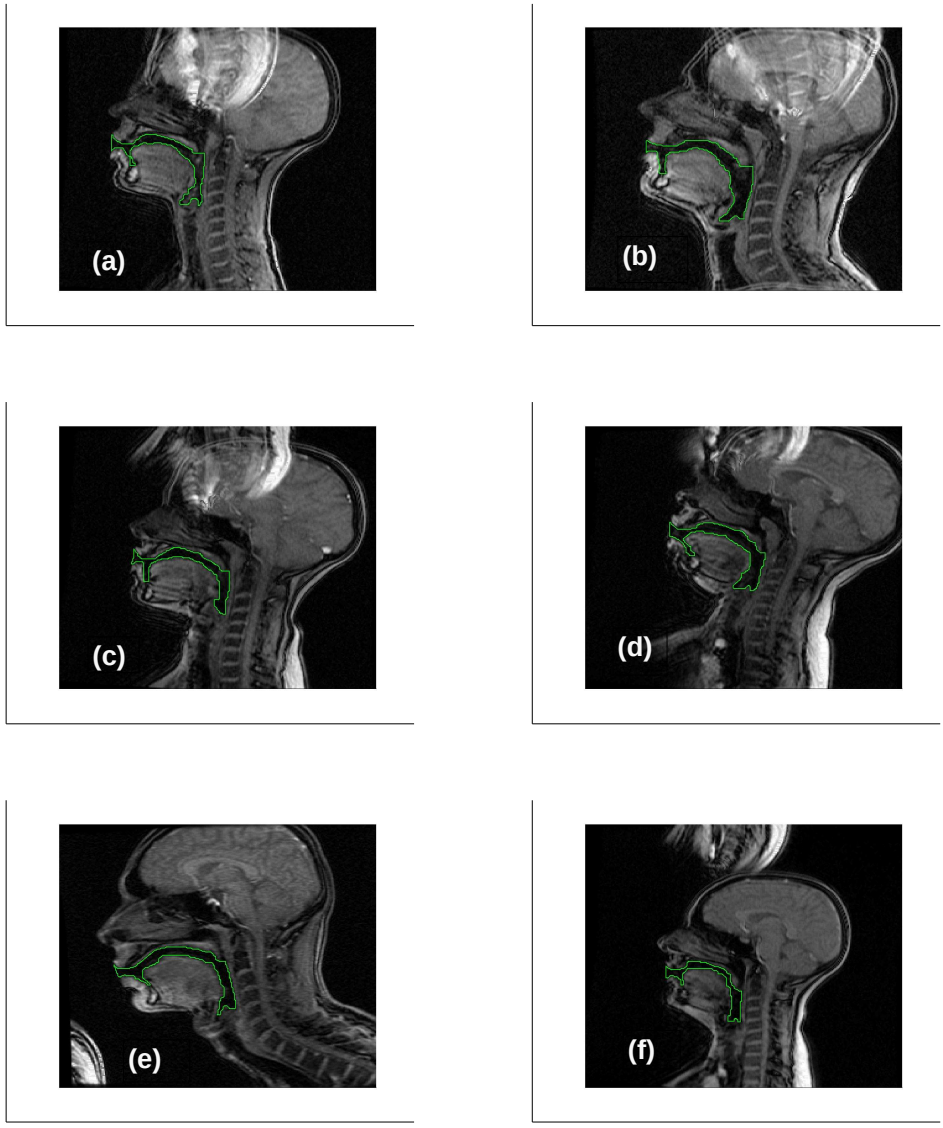


Figure 6.11: Semi-automatic (user-steered) segmentation of the vocal tract shapes of different speakers articulating the same phoneme /ə/.

generalise to all the cases, and a grouping to generate different shape classes might be necessary. Another factor is the structure of our MR images. We examined the cases of recognition failure, where the automatic initialisation goes wrong and the initial shape is not close enough to the actual position of vocal tract. Our observations suggested that in the majority of the cases, an object (a cavity) with very similar properties to the vocal tract in terms of orientation and structure is

present in the image. If the boundary cost of a shape fit to this cavity is lower than the boundary cost at the actual position, the OASM tends to pick the wrong shape and position for initialisation.

In the following sections, we discuss each of these factors and explain how we addressed them. We demonstrate how we resolved their consequences by presenting examples of their performance on test images.

6.5.1 Multi-speaker vs speaker-specific training

Each speaker in our training set has specific vocal tract features which make it different from others. Typically, speakers' vocal tracts are different in scale, orientation and relative position in the image. Examples of these differences can be observed in Figure 6.11. These difference in principle should not be a problem as the OASM algorithm uses scaling, rotation, and translation to address the inconsistency in *poses* of the shapes.

However, differences in scale indirectly affect the OASM performance by changing the number of landmarks for each. If we use the vocal tract shapes of all the speakers to find the landmarks automatically, the difference in scale between speakers results in having too many landmarks for one speaker, and maybe too few landmarks for another one. The excessive number of landmarks makes the boundary rigid for live wire to find the optimum boundary; in contrast too few landmarks may not captures the shape structure properly.

Furthermore, in training the OASM models, the large variation in scale, orientation and relative position of vocal tract is added to the rather extreme variation that already exists in the training set due to the articulatory configurations of each phoneme. Therefore, a single model that can capture all the categories of variation may not sufficiently represent the crucial details of the vocal tract. As described in the previous chapter, the differences heavily modify the topology of the vocal tract.

In addition, despite the fact that images were standardised in order to resolve the inconsistency in tissue intensity, the tissue textures were still different between

speakers. This can decrease the live wire boundary delineation accuracy as it is strongly dependent on the intensity training during the user-steered segmentation. Figure 6.12 presents a selection of examples that show how different variation aspects affect the outcome of OASM segmentation.

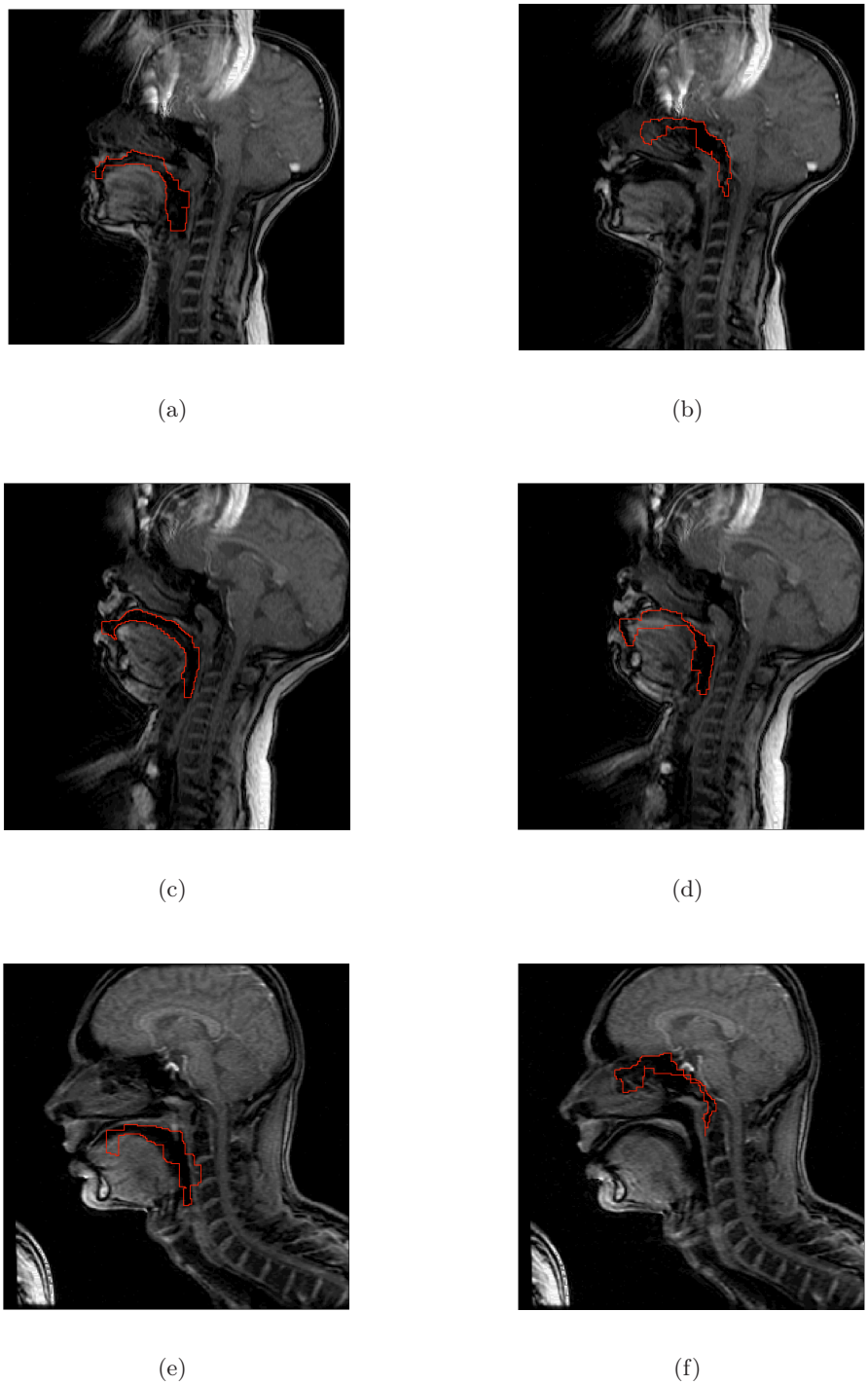
Given the requirements of the application, therefore, it seems inevitable to have to group the vocal tract shapes according to some criterion (phoneme-based or speaker-based) and create multiple OASMs instead of a single OASM to represent all. We tried to resolve this issue by classifying the shapes in speaker groups and training one OASM for each speaker. That is, the landmarks are extracted separately from the training shapes of each speaker, and an OASM is constructed from those landmarks individually per subject. We decided on speaker categorisation of OASMs to minimise the variation in terms of head and vocal tract orientation and tissue textures (brightness differences).

This approach has some advantages and disadvantages: in addition to resolving the orientation issue, one big advantage of this approach is that the model only has to account for variability due to articulation manner and not the speakers' variability in terms of anatomy. A main disadvantage is that the model cannot be generalised to delineate the images of a new speaker because it was not trained for that. However, training one OASM for each speaker is not exceedingly time-consuming, but it does require some human skill and expert knowledge.

6.5.2 OASM without automatic initialisation

We thoroughly examined *failure* instances of OASM segmentation after and before grouping the shape models based on speakers. By failure, we refer to scenarios where the achieved segmentation has zero or negligible overlap with the true segmentation (original vocal tract). The results suggested that these cases were mostly affected by the automatic initialisation outcome, rather than the performance of OASM in delineation and fine recognition.

For each case, we plotted the cost related to each shape at each location across



(a) (b)

(c) (d)

(e) (f)

Figure 6.12: OASM segmentation results. Images (a) and (c) are examples of OASM performing relatively well. Images (b) and (f) show a case of mis-recognition due to orientation variation. Images (d) and (e) show OASM failing in delineation as a result of inconsistency in tissue intensity.

the search area of the image. We observed that the cost function is minimum at a location in the image that is not the correct position of the vocal tract for these cases where it failed.

There are two main reasons for this outcome. Liu and Udupa [2009] argue the automatic initialisation step of OASM is very basic and is only based on a coarse location search across the image. Therefore, the automatic initialisation fails when dealing with large variation in scale and orientation. The multi-speaker OASM therefore is likely to fail in initialisation due to the amount of scaling and variation. The speaker-specific training can successfully address this problem.

The other main reason behind automatic initialisation failure is related to specifications of vocal tract images. Figure 6.13 shows the cost plots of the correct and wrong initialisation of the shapes, in speaker-specific models. Images (a) and (c) show the value of the cost function where darker intensities correspond to lower cost values obtained for image elements. As depicted in Figures 6.13 (b) and (d) there is a cavity area in the images with similar properties and orientation to the vocal tract. The OASM initial recognition thus fails to distinguish the two and occasionally returns the initial shape at an incorrect position, because the cost value tends to be smaller at that region (Figure 6.13 (c)). This was similarly reported in [Avila-García 2007] in which the tongue contours were fitted on the boundaries of the skull, as it was the best match returned by the algorithm.

To adapt the initialisation to fit this application, we modified the OASM automatic initialisation step, limiting the coarse recognition step to ASM recognition at the specific location, the location of the mean ASM shape model, while leaving the fine recognition and delineation as before (OASM two-level dynamic programming search). That is, instead of doing a coarse search across a range of images, we limit the initial recognition to the position of the average shape model. Having categorised our image modelling by speakers, the images in each testing group are expected to have the same profile in terms of approximate location as the constructed model. An essential advantage of OASM is that it returns the globally optimal boundary

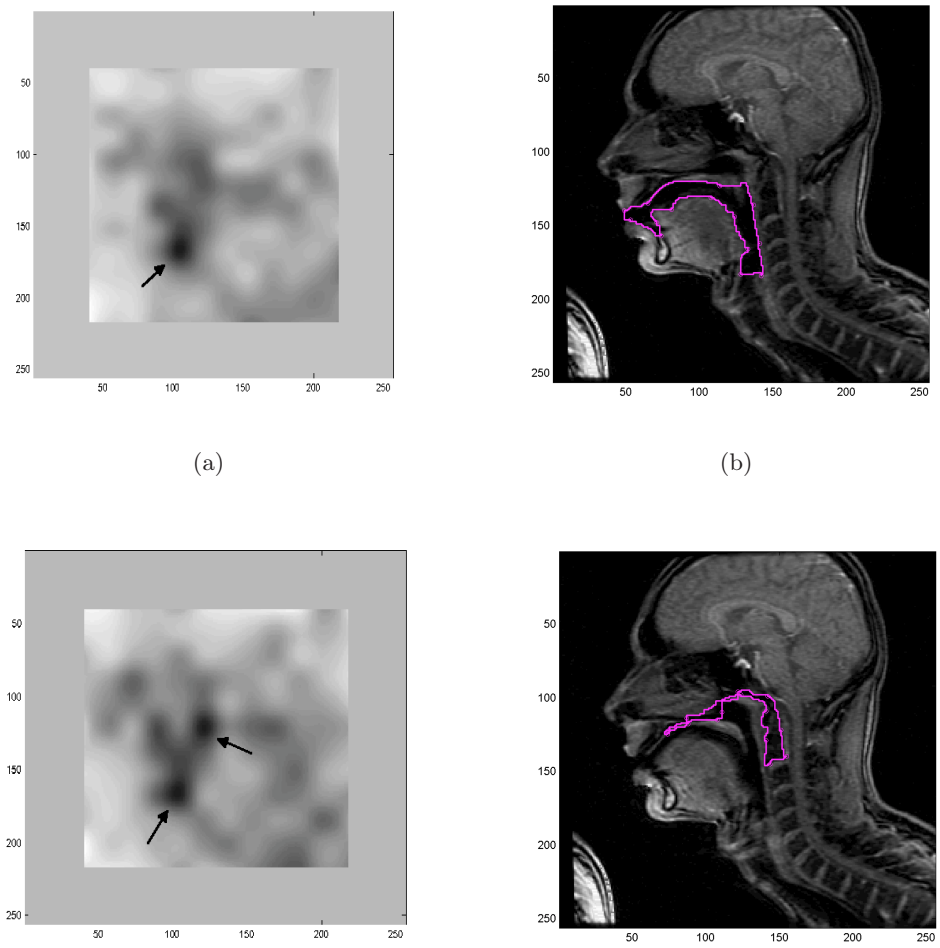


Figure 6.13: OASM boundary cost plots (left) and corresponding segmentations (right). The darkest areas on cost images (a) and (c) represent the smallest boundary costs.

if the initialised shape is close enough to the actual boundary in the images. Consequently, we were able to start the OASM fine recognition and delineation on the mean shape model constructed and assume that it is the closest location to the vocal tract. Note that limiting the search is only applicable in speaker-specific models as the location of the object is nearly fixed.

6.6 Conclusion

In this chapter, we proposed a new framework for segmenting vocal tract MR images by combining two existing methods in the literature. Unlike purely region-based segmentation methods, the segmentation method in this framework is able to delineate the vocal tract by using a shape structure that bounds the airway and prevents any leakage from the airway to the surrounding air. The proposed framework addresses the main challenges in automatic extraction of vocal tract shape, such as inter-phonemic variation of the shape, the blurring around the edges of the tract airway and connectivity of the airway to other channels of air such as nasal cavity or surrounding air. Except for the user-steered segmentation of training images, the rest of the procedures are done automatically, adjusting the model for applications on large MRI databases of articulation.

By analysing the preliminary results we noticed that the recognition phase in OASM fails in some of the images due to inter-speaker orientation variability and presence of cavity areas similarly oriented to the vocal tract in the images. We attempted to address these issues by categorising the shape models by speakers. To further improve the results, we restricted the OASM coarse recognition phase, by using the initial mean shape model position for each speaker, and bringing the landmarks closer to the boundary by ASM recognition.

In the next chapter, we extensively evaluate the performance of the proposed framework by analysing the segmentation results, in terms of

(a) how the segmentation method performs qualitatively and quantitatively from an image processing and analysis perspective,

(b) how the extracted vocal tract regions relate to previous results of area function measurements, and

(c) vocal tract acoustics.

Chapter 7

Evaluation of Automatic Segmentations

7.1 Introduction

In this chapter, we present the evaluation results of the vocal tract shape extraction framework introduced in the previous chapter.

Our evaluation experiments were carried out on the segmentations of test images that were not included in the training set. A set of user-steered "gold standard" segmentations was produced by an expert for the test images. The qualitative analysis shows how good the segmented shape is, given the expected vocal tract shape for each phoneme. The quantitative analysis uses a selection of distance-based and region-based metrics to compare the segmentation results with the gold standard numerically.

To analyse the effectiveness of the algorithm in speech production modelling specifically, we carried out a set of articulatory assessments. We first compared the gold standard segmentations and OASM segmentations in terms of their midsagittal distances along the vocal tract. The area functions were compared between the speakers and also with the area functions suggested in previous work in the literature. Finally, an acoustic analysis of the estimated area functions was undertaken to test

whether the extracted shapes are acoustically reasonable as well.

In the remainder of this chapter, we first review settings for the experiments and measurements. The metrics of quantitative evaluation are presented in section 7.2. In section 7.3 we briefly review the training dataset, and explain experiment settings and parameter tuning for training the models. The qualitative analysis is explained in section 7.4.1, followed by quantitative evaluations in section 7.4.2. Section 7.5 includes the articulation modelling evaluations of OASM segmentations. A comparative study of midsagittal distance between OASM segmentations and gold standard is presented in 7.5.2. The area function comparisons with previous work are presented in section 7.5.3. Finally, the effectiveness of the obtained area functions is evaluated in section 7.5.4 by analysing the formant frequencies synthesised from those area functions.

7.2 Evaluation metrics

The segmentation results were statistically evaluated using a few methods that are extensively used in the field of biomedical image processing for measuring the quality of segmentation. We use region-based evaluation [Udupa *et al.* 2006] to evaluate the precision and accuracy of the segmentation results. The difference between gold standard segmentations and OASM segmentations are analysed using distance-based metrics [Heimann *et al.* 2009] reported in millimetres (mm). Below we explain the region-based and distance-based metrics followed by the numerical analysis of the performance of the proposed approaches.

Region-based evaluation. The region-based evaluation focuses on assessing the precision and accuracy of the segmentation method [Udupa *et al.* 2006]. An OASM segmented vocal tract shape V_{OS} is compared to a gold standard (user-steered) segmented boundary V_{GT}, obtained by user-steered delineation carried out by an expert once.

Precision is a measure of reproducibility of the results and is calculated by

measuring the ratio of common area of tissue between V_{OS} and V_{GT} and the total amount of tissue present in the union of V_{OS} and V_{GT} as

$$P = \frac{|V_{OS} \cap V_{GT}|}{|V_{OS} \cup V_{GT}|}. \tag{7.1}$$

Accuracy is a measure of precision of the delineation. It is defined in terms of *true positives* (TP) and *true negatives* (TN).

The area of tissue in the correct segmentation that is covered by the delineation algorithm is referred to as delineation *sensitivity* and is represented by

$$TP = \frac{|V_{OS} \cap V_{GT}|}{|V_{GT}|}. \tag{7.2}$$

The delineation *specificity* of the algorithm, represented by TN, is equal to the ratio of the reference region U that does not belong to the object and that was excluded from the segmentation result to the reference region that does not belong to the object, defined as

$$TN = \frac{|U - V_{OS} - V_{GT}|}{|U - V_{GT}|}. \tag{7.3}$$

The higher values for TP and TN, the better delineation performance of the method.

Dice similarity is a measure of mutual overlap between two shapes, defined as

$$D = \frac{2|V_{OS} \cap V_{GT}|}{|V_{OS}| + |V_{GT}|}. \tag{7.4}$$

Distance-based evaluation. The distance-based methods are used to calculate the errors in the segmentation performance [Heimann *et al.* 2009]. These measures are reported in millimetres (mm).

Maximum Symmetric Contour Distance (MSD) uses Euclidean distance to estimate the maximum distance between a pixel c_{OS} on the OASM segmented

vocal tract contour C_{OS} and the gold standard vocal tract contour C_{GT}, defined as

$$ED(c_{OS}, C_{GT}) = \min_{c_{GT} \in C_{GT}} ||c_{OS} - c_{GT}||, \tag{7.5}$$

where c_{OS} and c_{GT} are contour elements belonging to the contours C_{OS} and C_{GT}, respectively. The MSD value, also referred to as Hausdorff maximum distance [Huttenlocher *et al.* 1993], is then calculated as

$$MSD(C_{GT}, C_{OS}) = \max\left(\max_{c_{OS} \in C_{OS}} \Big(ED(c_{OS}, C_{GT})\Big), \max_{c_{GT} \in C_{GT}} \Big(ED(c_{GT}, C_{OS})\Big)\right). \tag{7.6}$$

Note that the Hausdorff distance is symmetric and not oriented.

Average Symmetric Contour Distance (ASD) is the average of the distances between the two contours, defined as

$$ASD(C_{GT}, C_{OS}) = \frac{1}{|C_{GT}| + |C_{OS}|}\left(\sum_{c_{GT} \in C_{GT}} ED(c_{GT}, C_{OS}) + \sum_{c_{OS} \in C_{OS}} ED(c_{OS}, C_{GT})\right). \tag{7.7}$$

Better delineation leads to smaller ASD (0 mm for perfect delineation).

Root Mean Square Symmetric Contour Distance (RMSD) is similar to ASD, the only difference is that it penalises large errors. It defined as

$$RMSD = \sqrt{\frac{1}{|C_{GT}| + |C_{OS}|}} \times \sqrt{\sum_{c_{GT} \in C_{GT}} ED(c_{GT}, C_{OS}) + \sum_{c_{OS} \in C_{OS}} ED(c_{OS}, C_{GT})}. \tag{7.8}$$

Smaller values of RMSD suggest smaller deviation of OASM segmentations from the gold standard.

Failures. The failures in segmentation are defined as cases where no or minimal overlap is observed between the segmentation results and experts' segmentations. The cases of failure are reported in terms of percentage of overlap between the segmentations. The minimum overlaps are reported separately and are excluded from further region-based and distance-based evaluations.

7.3 Experiment settings and results

We first present an overview of the train and test data. Different available settings for the experiments are then explained and their enforcement effects are analysed through the results.

7.3.1 Data description

The training data to construct the model and the testing data for evaluation were described previously in section 6.5. The test set images were semi-automatically segmented by a human expert to generate the "gold standard" or "ground truth" set. The segmentations were carried out with user-steered live wire boundary tracking method in 3Dviewnix software [Udupa *et al.* 1993] (similar to training images).

To examine the validity of the expert segmentations, we compared them with the user-steered segmentations of another expert on a selection of our test images (12 images, 2 from each speaker). Note that this was a limited set only for validating the user-steered segmentations used in the experiments, and they are not directly utilised in the experiments/evaluations. The comparison of segmentations of the two experts proved an average agreement (in terms of precision) of 80% between the two with a standard deviation of 5%. The disparity is due to the fact that although both segmentations are correct and precise, there are very fine details, particularly around the epiglottis and the lip area that do not have a clear boundary definition (even in the literature). This, in fact, is one of the main reasons that image processing based evaluations, alone, may not be sufficient for analysing the results and further evaluations, in terms of midsagittal distances or area functions are necessary.

7.3.2 Settings

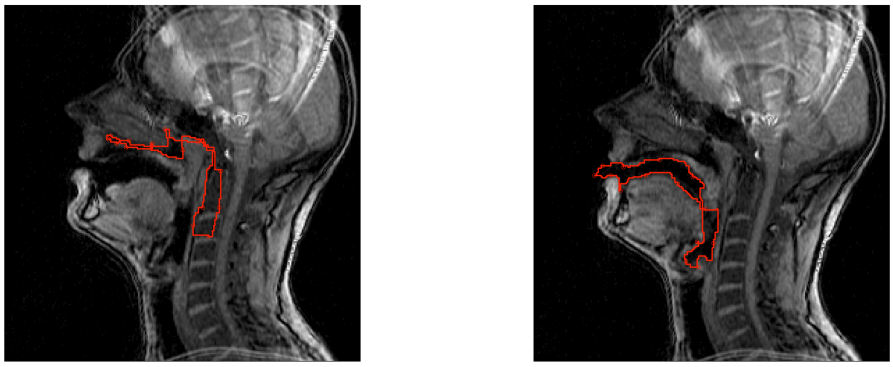
A few parameters need to be determined for both training and segmentation steps in OASM. For the majority of parameters, we used the original values suggested reported in [Liu and Udupa 2009]. Table 7.1 lists the values of different parameters used, together with a brief description.

Table 7.1: Parameters used in OASM training and segmentation.

parameter	description	value(s)
l	number of points selected on each side of the contour in the appearance aspect of the model during training	3
m	number of point selected on each side of landmarks of the model during segmentation (search range)	{6,9}
δW	constant parameter in the size of rectangular search window in 1st level of dynamic programming	10

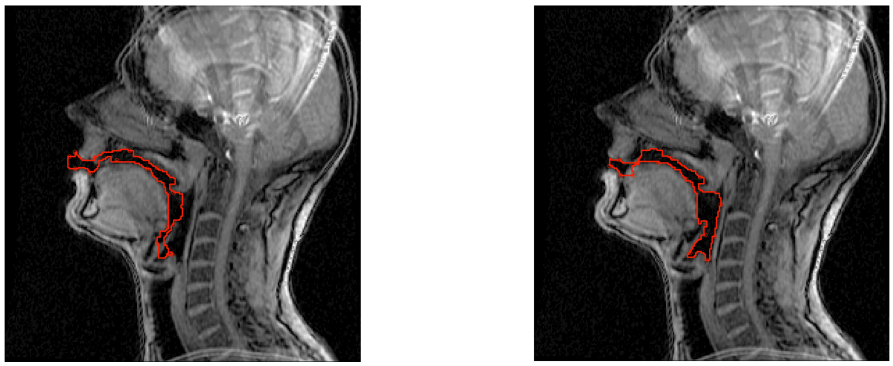
Search range during segmentation. We look at modifying the search range m during segmentation in the OASM algorithm. A quick evaluation of the OASM segmentations suggests that in this application the search range m affects both recognition and delineation. Figure 7.1 illustrates the variation in recognition when different search range values are used with the same number of landmarks. Figure 7.2 shows the sensitivity of delineation mechanism to the search range.

Landmarks. Ideally, we need a compact set of landmarks that are the key anchor points for live wire boundary tracking, and represent the shape of the vocal tract effectively. We tried to find the optimum set of landmarks experimentally by setting δ in RBS landmark tagging to different values and observing the generated set of landmarks. If the landmarks roughly corresponded across the range and the landmark positions were satisfactory in terms of representing key articulatory features, that set of landmarks was chosen as the training set for OASM.



(a) $m = 6$ (b) $m = 9$

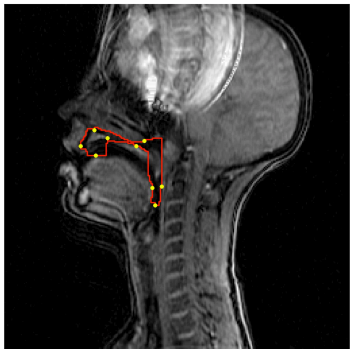
Figure 7.1: OASM recognition of the same image with the same number of landmarks, using two different search range values.



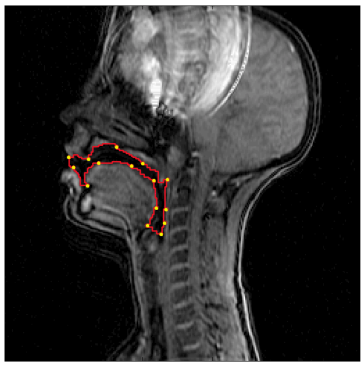
(a) $m = 6$ (b) $m = 9$

Figure 7.2: OASM delineation of the same image with the same number of landmarks, using two different search range values.

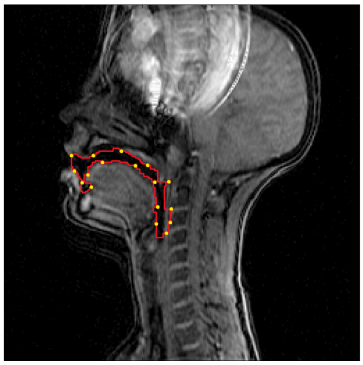
We extensively examined the segmentations with different number of landmarks. The images in Figure 7.3 show a few examples of OASM segmentations of the same object, i.e. same image, with different number of landmarks. Picking the right number of landmarks plays an important role in improving the quality of both recognition and delineation. The suitable number of landmarks is different between speakers, as the size of the vocal tract varies considerably between speakers. This is illustrated in Figure 7.4 where a reasonable segmentation result is obtained for two different speakers with speaker-specific OASMs. While 12 landmarks are sufficient



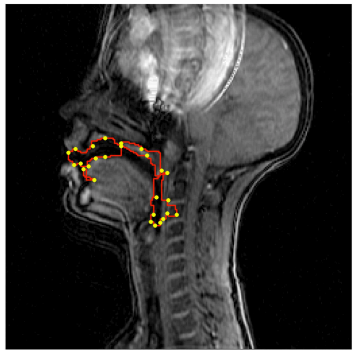
(a) 9 landmarks



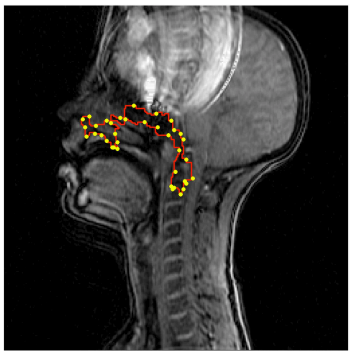
(b) 15 landmarks



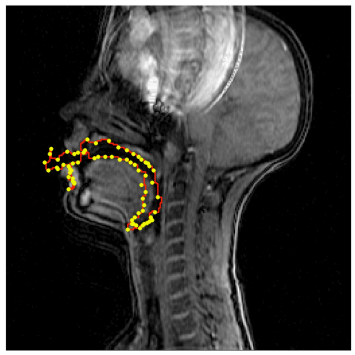
(c) 17 landmarks



(d) 19 landmarks

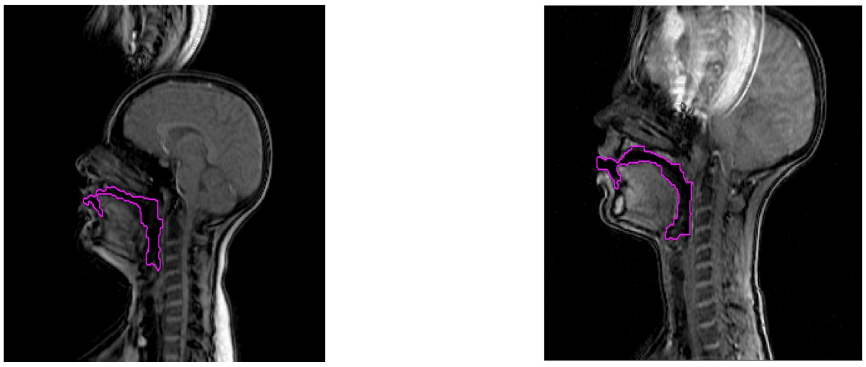


(e) 26 landmarks



(f) 37 landmarks

Figure 7.3: OASM segmentations ($m = 6$). A moderate number of landmarks results in better recognition and delineation (15 to 17 in this example).



(a) speaker R, 19 landmarks

(b) speaker A, 12 landmarks

Figure 7.4: OASM segmentations for images of speaker A and speaker R. For speaker R, 12 landmarks were used to describe the training shapes, while training shapes of speaker A were represented with 19 landmarks. The contours are obtained by OASM segmentation without automatic initialisation.

to describe the vocal tract shape of speaker R, a minimum of 19 is required by the algorithm for speaker A.

The preliminary results confirmed the sensitivity of the algorithm to a few parameters such as search range and number of landmarks, and hence the need for parameter tuning. However, according to the inventors of OASM [Liu and Udupa 2009], if we have an appropriate number of landmarks, the model should be independent of the search range. We therefore use the robustness of the outcome relative to the search range to determine the sufficient number of landmarks. If the number of landmarks is suitable for constructing an OASM shape model, then the outcome should not be sensitive to the search range. We use this fact to complete our quantitative evaluations.

7.4 Image-based evaluations

In this section, the results of qualitative and quantitative image-based evaluations of OASM segmentations are presented. For qualitative evaluations, the OASM segmentations were plotted over the original MR images where the target object was to be found. The objective is to see how well the obtained contour describes the shape of the vocal tract and separates it from the tissues. The OASM segmentations were compared quantitatively with the user-steered segmentations (gold standards).

7.4.1 Qualitative analysis

We carried out a subjective evaluation on all 72 test images, with different experimental settings. Our observations suggested that if OASM successfully finds the initial position of the shape, it can precisely delineate the vocal tract. The details can indeed be very fine at sensitive areas such as the opening of the lips or the sublingual cavity.

Figure 7.5 shows an example of a successful OASM segmentation. The generated boundaries include most of the features crucial in describing the vocal tract shape, such as the width of the gap between the lips, the posture of the velum, the height of the tongue, and the width of the cavity all along the vocal tract from glottis to lips. Some features such as length of the cavity cannot be as precisely estimated by OASM segmentations. Figure 7.6 shows examples of OASM segmentation results for different speakers. Figures 7.6 (b) and (e) are two examples where the contour does not capture the details at the glottal region, and consequently the estimated length of the obtained vocal tract contour is slightly shorter than the actual length. We discuss this further in section 7.6.

We also examined the effect of including all the inter-speaker variance by constructing one OASM shape model using the training images of all speakers in the dataset. Figure 7.7 shows a few examples of segmentations using this OASM model plotted over the original images. As these examples illustrate, the segmentation method sometimes may not successfully find the optimum boundary. The results in

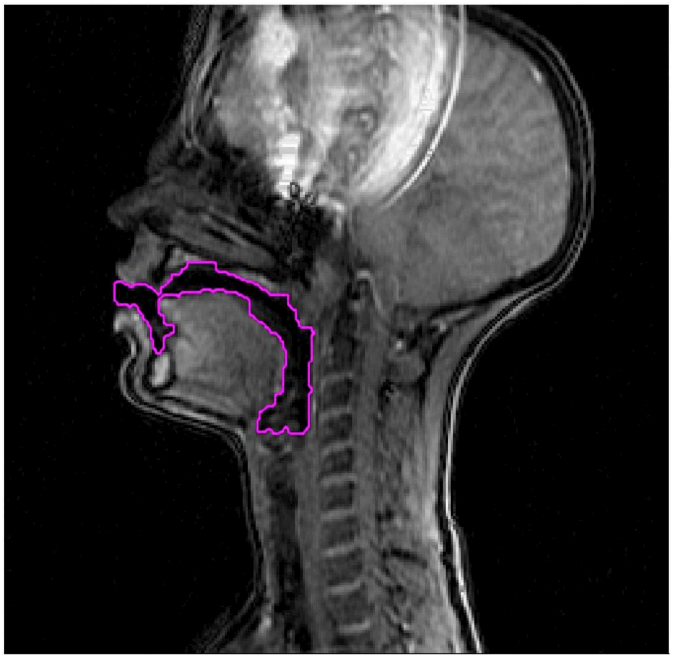
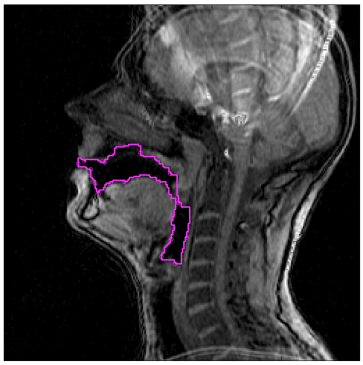


Figure 7.5: OASM segmentation. The delineation is very fine even at the challenging regions such as the opening of the lips, or the sublingual cavity. Contours in magenta represent the results of OASM segmentation without automatic initialisation.

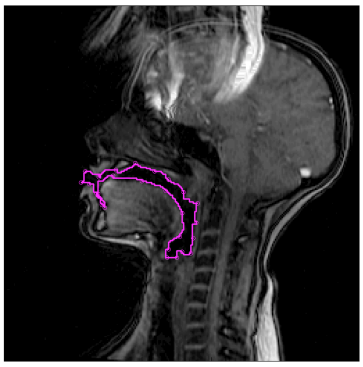
the landmark analysis earlier in this section suggested that the sufficient number of landmarks varies for different speakers, due to anatomical differences. This is more clearly observed in multi-speaker models, where a particular number of landmarks may work fine for one speaker, but is not the optimum value for other speakers' vocal tract shapes. Thus, OASM may perform better in segmenting some speakers' images and not perform as well on other speakers, even when applied to the same phoneme. In addition, as discussed in the previous chapter, by training on multiple speakers, too many degrees of variation are introduced into the model that might not be very well characterised due to the structure of data. These variations include differences in tissue density, significant anatomical differences between speakers, and different orientations of the vocal tract.

We limited the amount of variation by categorising the shapes per speaker, i.e. constructing an OASM individually for each speaker. As expected, the speaker-specific models proved to work better in terms of both recognition and delineation.

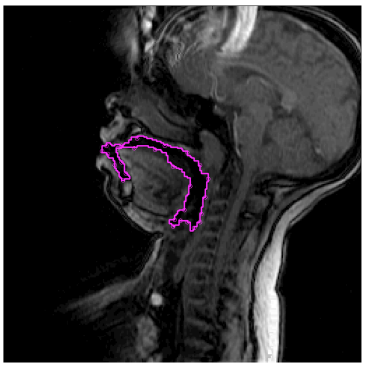
(a) speaker A, 26 landmarks, $m = 6$



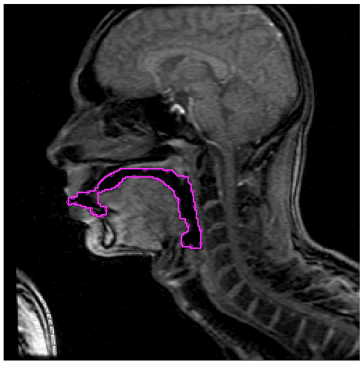
(b) speaker C, 16 landmarks, $m = 9$



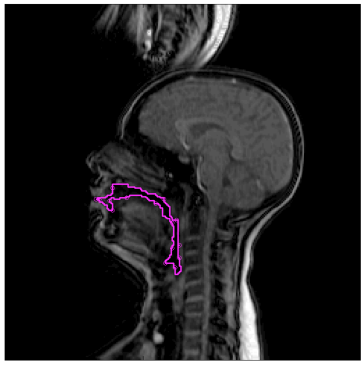
(c) speaker L, 26 landmarks, $m = 9$



(d) speaker M, 27 landmarks, $m = 6$

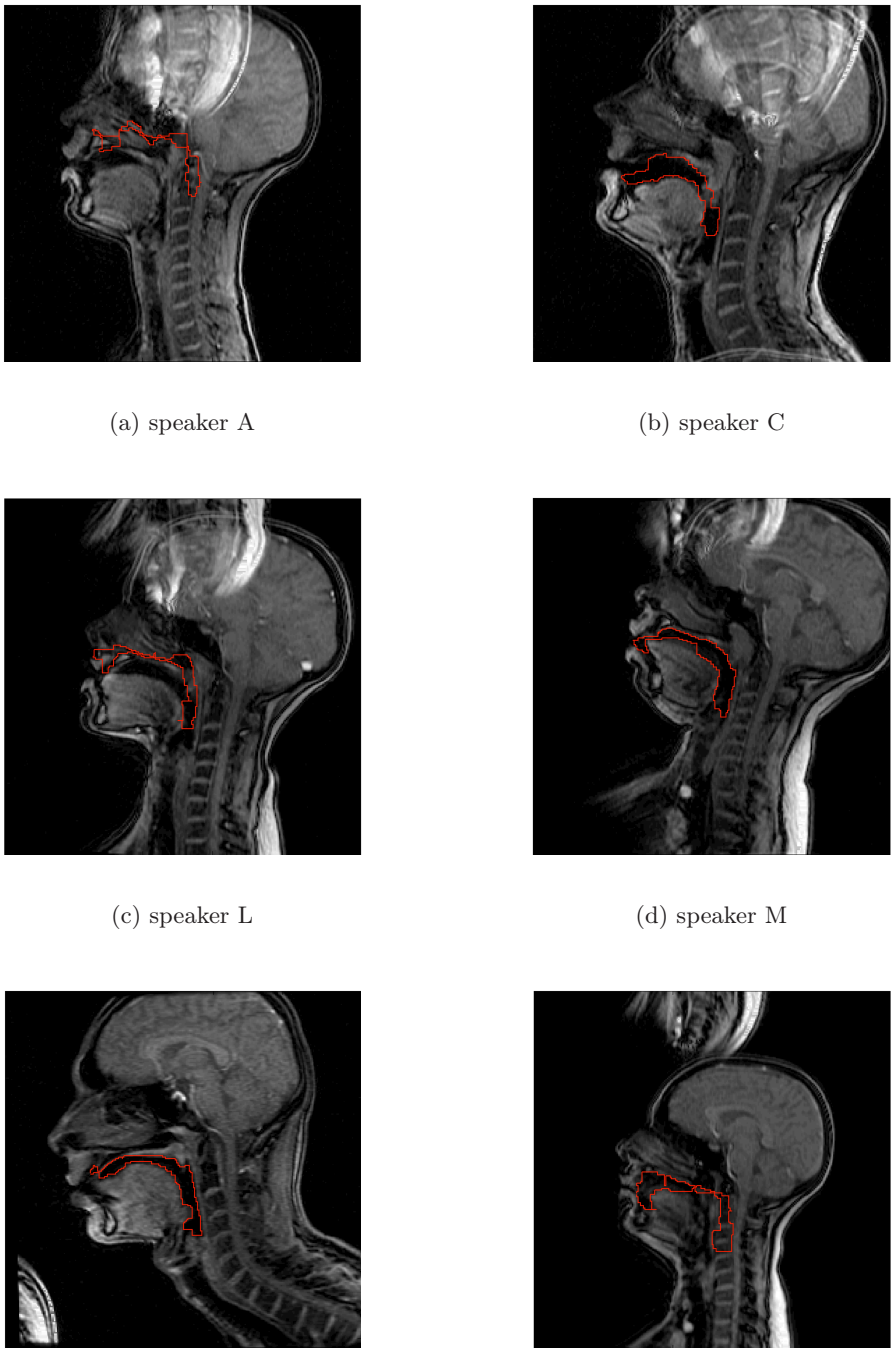


(e) speaker P, 26 landmarks, $m = 6$



(f) speaker R, 15 landmarks, $m = 6$

Figure 7.6: Segmentations without the OASM automatic initialisation step for phoneme /ɔ/. Speaker-specific OASMs were individually trained for each speaker but including all the phonemes in the database. Contours in magenta show the results of OASM segmentation without automatic initialisation.



(a) speaker A

(b) speaker C

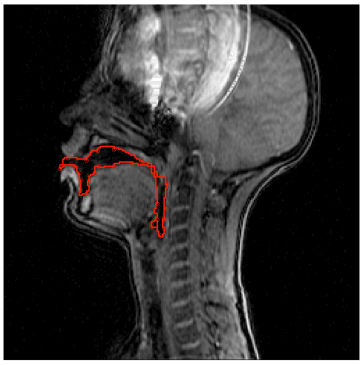
(c) speaker L

(d) speaker M

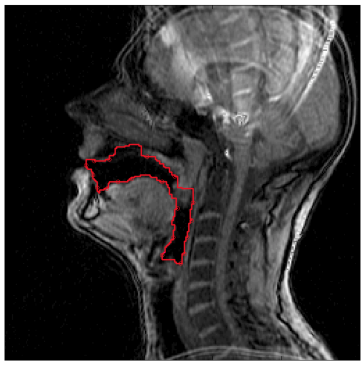
(e) speaker P

(f) speaker R

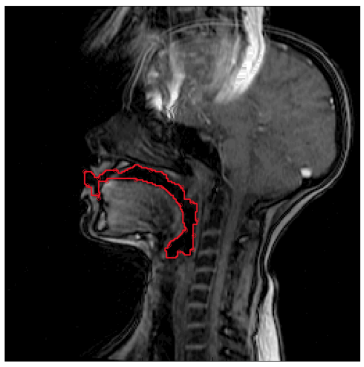
Figure 7.7: OASM segmentations ($m = 9$, 14 landmarks) for the phoneme /ɔ/. One OASM was trained for all the speakers, using images from all phones in the database. OASM automatic initialisation was used. Contours in red show the results of OASM segmentation with automatic initialisation.



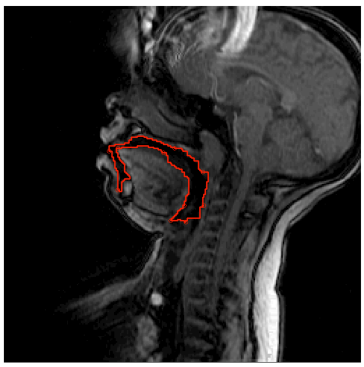
(a) speaker A, 26 landmarks



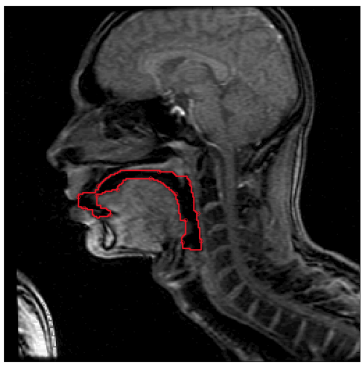
(b) speaker C, 19 landmarks



(c) speaker L, 19 landmarks



(d) speaker M, 10 landmarks



(e) speaker P, 14 landmarks

(f) speaker R, 9 landmarks

Figure 7.8: OASM segmentations for the phoneme /ɔ/, $m = 9$. Multiple OASMs were trained individually for each speaker but including all the available phones in the database for each speaker. OASM automatic initialisation was used. Contours in red show the results of OASM segmentation with automatic initialisation.

The speaker-specific OASMs are not affected by inter-speaker variation (struc-

ture and orientation of vocal tract) and therefore can characterise the variation in terms of articulation better. The speaker-specific OASM segmentation results for the images in Figure 7.7 are given in Figure 7.8. It is evident from the images that the segmentations are improved in speaker-specific OASM segmentations. Note that both Figure 7.7 and 7.8 present results of the OASM with automatic initialisation.

7.4.2 Quantitative analysis

In this section, we compare the automatic segmentation results with the user-steered segmentation using the metrics introduced in section 7.2. We first report the number of failures with different experimental settings, such as training multi-speaker or speaker-specific OASMs, or performing/not performing the coarse recognition in automatic initialisation. We refer to the latter case as OASM without automatic initialisation step. We then investigate the precision evaluation, followed by accuracy and distance-based metrics for each setting.

The number of landmarks was chosen experimentally after extensive experimentations, and observing the sensitivity of the model towards the search range, as explained in section 7.3.2. The chosen number of landmarks are shown in Table 7.3.

Failures. The cases of failure are reported separately first and they are excluded from further evaluation steps. The failures are decided based on the overlap of the segmentation results with the gold standard segmentation, and are reported at 0%–10%, 0%–20% and 0%–50% overlaps. Table 7.2 shows the number of failures out of 72 test images in different settings used in OASM training and segmentation stages.

Table 7.2: Number of failures at different ranges of overlap between OASM segmentations and gold standard.

Model settings	overlap (less than)		
	10%	20%	50%
Multi-speaker OASM with automatic initialisation	5	10	21
Speaker-specific OASMs with automatic initialisation	1	2	6
Speaker-specific OASMs without automatic initialisation	0	0	0

Table 7.3: The mean and standard deviation of precision for OASM without automatic initialisation. The precisions are calculated over segmentations with different search range parameters (m={6,9}). Speaker-specific OASMs were trained using all the phonemes in the database. The number of landmarks was individually chosen for each speaker.

Speaker	Best Precision (%)	landmarks	Worst Precision (%)	landmarks
speaker A	88.98 ± 5.04	26	82.33 ± 5.89	36
speaker C	85.88 ± 7.49	14	78.60 ± 8.22	38
speaker L	83.83 ± 1.90	26	75.79 ± 6.13	14
speaker M	89.47 ± 7.20	27	80.26 ± 9.00	14
speaker P	84.73 ± 5.49	16	80.86 ± 5.43	37
speaker R	88.13 ± 8.74	9	75.15 ± 15.24	35

The count of failure instances is a proof of our argument in chapter 6, that adaptation and modification of OASM is inevitable in this application. In the rest of quantitative evaluations, we only report the results for the settings with the minimum rate of failure, speaker-specific OASMs without automatic initialisation.

Precision. The reproducibility of the results with two different settings for the search range parameter $m = \{6, 9\}$ was evaluated with a precision calculation. The precision average and standard deviation values are presented in Table 7.3. Two figures are given for each speaker: the best precision and the worst precision. As the numbers suggest, there is not a very big gap (7–8 percent) between the best and worst precision values. Note that different precision values were obtained for different number of landmarks for each speaker. The higher value of precision suggests more consistency in the results with two different search ranges. Therefore, for each speaker, we chose the number of landmarks that resulted in the highest precision values in Table 7.3 for the rest of the experiments.

Region-based and distance-based evaluations. The evaluation results of region-based and distance-based metrics are presented in Tables 7.4 and 7.5. These are the evaluations for the models in Table 7.3 that resulted in the highest precision over the search range.

Table 7.4: The mean and standard deviation of region-based evaluations for speaker-specific OASMs without automatic initialisation.

	Region-Based(%)		
Speaker	Sensitivity (TP%)	Specificity (TN%)	Dice (%)
speaker A	95.06 ± 2.85	99.32 ± 0.11	76.51 ± 4.84
speaker C	85.16 ± 3.87	99.69 ± 0.17	86.17 ± 2.74
speaker L	90.06 ± 3.10	99.61 ± 0.10	85.01 ± 2.87
speaker M	94.14 ± 1.83	99.61 ± 0.18	87.71 ± 4.09
speaker P	87.58 ±8.07	99.47 ± 0.05	81.75 ± 4.88
speaker R	84.86 ± 4.10	99.64 ± 0.11	78.24 ± 3.59

We evaluated the sensitivity and specificity of the models by comparing the true positives and true negatives in Tables 7.4. Relatively high values are obtained for all the speakers in terms of specificity and sensitivity, suggesting good performance in terms of delineation. The Dice similarity values vary from 76.51% to 87.71% for different speakers, meaning that the minimum overlap between the OASM segmentations and gold standard is more than 75%.

The distance measurements provided in Table 7.5 suggest that the average maximum distance between the boundaries in gold standard and OASM segmentations is 11.43 mm with an average ASD of 1.69 mm, and an average RMSD error of 2.49 mm. These results suggest that the proposed framework can automatically extract the vocal tract shape from MR image with relatively good accuracy in delineation, with the gold standard and OASM segmentation having an average overlap (Dice similarity) of 82.56%. This is remarkable considering that vocal tract shapes are rather complex with extreme variations and very fine details.

In order to evaluate the application of this framework to studying speech articulation and to analyse how well the segmented shapes represent vocal tract structure, we need to analyse the differences between gold standard and OASM segmentations and their importance in articulation research. In the next sections, we extend our analysis to evaluations in terms of articulation modelling.

Table 7.5: The mean and standard deviation of distance-based evaluations for OASM without automatic initialisation.

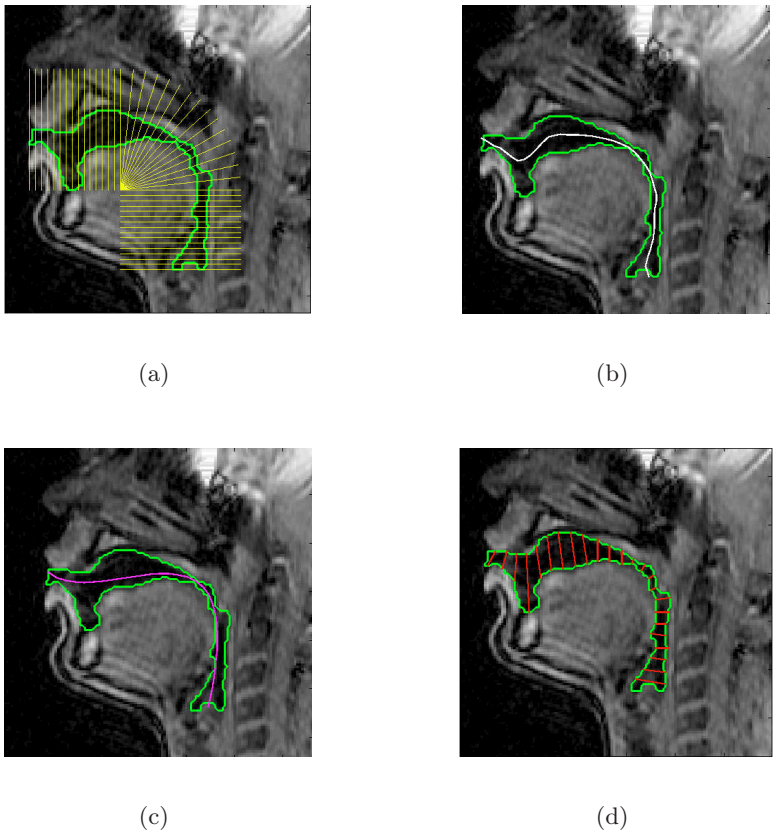
	Distance-Based (mm)		
Speaker	MSD (mm)	ASD (mm)	RMSD (mm)
speaker A	8.63 ± 1.71	1.60 ± 0.26	2.25 ± 0.49
speaker C	14.68 ± 3.35	1.60 ± 0.41	2.82 ± 0.80
speaker L	8.98 ± 2.32	1.26 ± 0.25	2.01 ± 0.48
speaker M	7.30 ± 1.79	1.12 ± 0.25	1.65 ± 0.41
speaker P	15.26 ± 3.51	1.88 ± 0.31	3.07 ± 0.67
speaker R	13.73 ± 3.23	1.92 ± 0.23	3.17 ± 0.47

7.5 Articulation modelling analysis

In order to calculate the area functions, a gridline system was fitted to the image in midsagittal plane. The initial gridline consists of three parts: a set of parallel horizontal lines at the pharynx, a set of vertical parallel lines at the front of the oral cavity, and a semi-polar grid system in the middle to connect the two. Figure 7.9(a) displays an example of the grid superimposed on a sample MR image and the obtained OASM contour. The centroids of the gridlines are used to generate a centre line for the vocal tract (Figure 7.9(b)). The vocal tract midline is then generated using regression on the centre points (Figure 7.9(c)). Next, the midline is divided into 24 equal parts to obtain 24 cross-sectional areas that are equally distanced from each other (Figure 7.9(d)). We chose 24 experimentally using the images of all speakers (the distance between the cross sections are neither very close nor extremely far). Finally, the midsagittal distance is obtained by calculating the length of the line perpendicular to the midline at the centre of each area that connects the inner and outer walls of the vocal tract.

7.5.1 From midsagittal width to cross-sectional areas

The availability of data in other planes has facilitated generation of area functions that map the midsagittal width to the cross-sectional areas. Traditionally, the vocal tract area functions are parametric models of midsagittal width at different distances along the vocal tract [Mermelstein 1973; Perrier *et al.* 1992; Sundberg *et al.* 1987],



(a) (b)

(c) (d)

Figure 7.9: Steps in calculating the midsagittal width. The vocal tract contour is obtained with OASM segmentation as explained in chapter 6. (a) 45 gridlines are superimposed on the image, (b) the centre line is calculated using the centre points from the edges on each gridline, (c) the vocal tract midline is smoothed using regression on the centre line, and (d) the perpendicular lines to the vocal tract midline are the cross-section lines. The length of each line is the midsagittal width at each distance.

as reviewed in chapter 2. The classic approach, and perhaps the most popular in the field, is Heinz and Stevens [1964]'s (α,β) model defined as

$$A(x) = \alpha d(x)^{\beta}, \tag{7.9}$$

where A is the cross-sectional area, x is the distance along the vocal tract midline from the glottis, and d is the midsagittal width at distance x. The parameters α and β are the transformation parameters estimated from cross-sectional areas for each speaker specifically. The transformation parameters are generally adapted

by researchers for each particular study and, in general, the optimum parameters are speaker specific. However, since our data are 2D, we did not have the means to optimise and adapt the parameters, we therefore simply used the parameters for male and female speakers proposed in [Soquet *et al.* 2002], as have been used in other studies [Busset and Laprie 2011; Laprie *et al.* 2013]. The vocal tract is divided by Soquet *et al.* to 8 distance regions: larynx, low pharynx, mid-pharynx, oropharynx, velum, hard palate, alveolar region, and labial region, following the proposals of Mrayati *et al.* [1988]. Figure 7.10 shows an example of such division of a vocal tract image.

We carried out our articulation analysis in the following order: first, the mid-sagittal widths of OASM contours are compared to the user-steered segmentations, for all the phonemes. Next, the area functions from OASM segmentation of vowels are compared with the area functions in previous work. Finally, to evaluate the results acoustically, we analyse the estimated vocal tract resonances obtained from the area functions, and compare them with expected formant frequencies, according to the prior literature on acoustic phonetics.

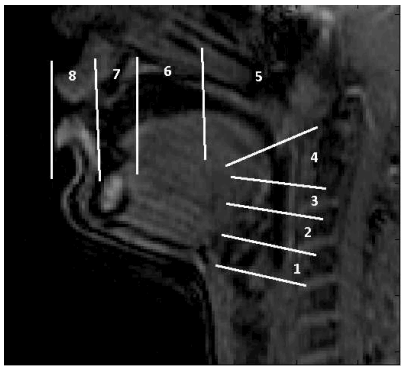


Figure 7.10: Place regions marked on an MR image of articulation of the vowel /ɔ/ by speaker A: (1) larynx, (2) low pharynx, (3) mid-pharynx, (4) oropharynx, (5) velum, (6) hard palate, (7) alveolar region, and (8) labial region.

7.5.2 Midsagittal width analysis

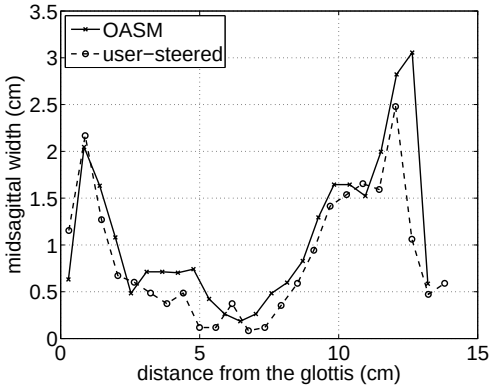
To compare the similarity of the articulatory configurations of the segmented shape and user-steered segmentations, the midsagittal widths at different distances from the glottis were compared. A selection of comparison plots of the user-steered segmentations and OASM segmentations are presented in Figure 7.11.

To compare the midsagittal differences numerically, the RMS error was calculated for each test image and averaged individually for each speaker. Table 7.6 presents the means and standard deviations of the RMS error for each speaker in centimetres. The average RMS errors are between 0.45 cm and 0.58 cm, with an average of 0.51±0.17 cm for all speakers.

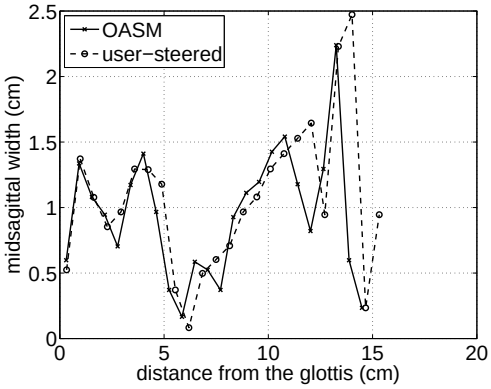
A closer look at the plots reveals that large differences in RMS errors occur at cross-sectional areas near the glottis and at the front of the vocal tract close to the lips. The bar plots in Figure 7.12 present the RMS errors at three different sections of the vocal tract: glottal region, low pharynx to alveolar region and labial region. The middle bar in each group, corresponding to the region from low pharynx to alveolar ridge, is more stable, ranging on average between 0.3 cm and 0.4 cm across subjects, proving OASM segmentation performs more robustly in that region. The highest RMS errors relate to the opening of the lips, where the boundary is not properly defined. The test of significance (paired-sample t-test) between the distribution of average errors in the three regions of vocal tract did no show statistically significant differences between the average errors at 5% significant level ($t = 1.35$ and $p = 0.23$

Table 7.6: The mean and standard deviation of RMS error for different subjects, averaged over all the test images for each speaker in midsagittal width.

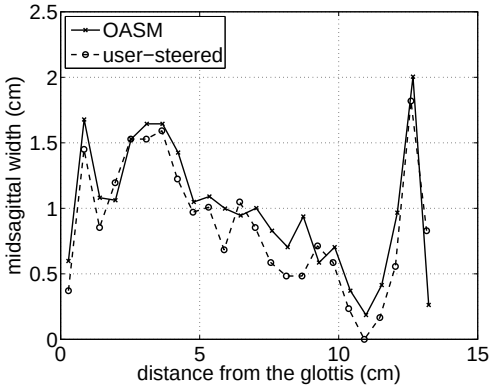
Subject	mean (cm)	standard deviation (cm)
speaker A	0.5383	0.1547
speaker C	0.4501	0.1392
speaker L	0.5062	0.1635
speaker M	0.5865	0.2129
speaker P	0.5776	0.1999
speaker R	0.5119	0.1723
average	0.5284	0.1737



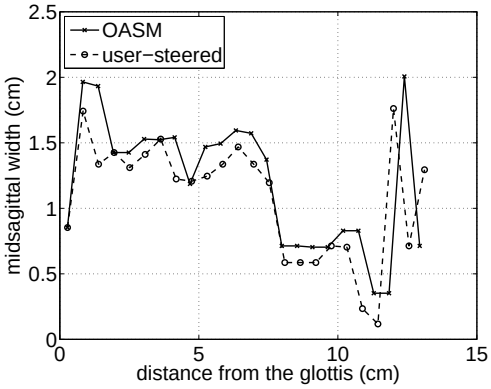
(a) speaker A, phoneme /ɔ/



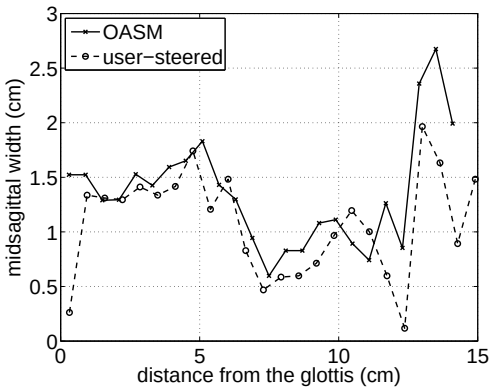
(b) speaker C, phoneme /ɔ/



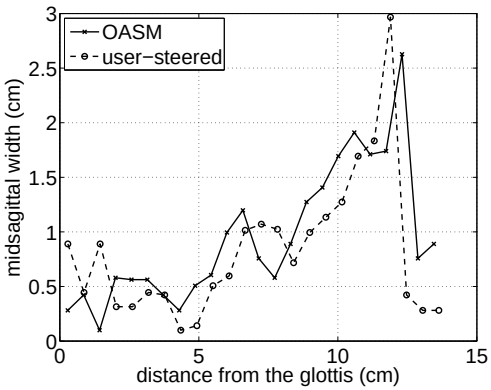
(c) speaker L, phoneme /ɪ/



(d) speaker M, phoneme /ɔ/



(e) speaker P, phoneme /ɔ/



(f) speaker R, phoneme /ʌ/

Figure 7.11: Plots of midsagittal width comparisons between the user-steered contours and OASM segmentations.

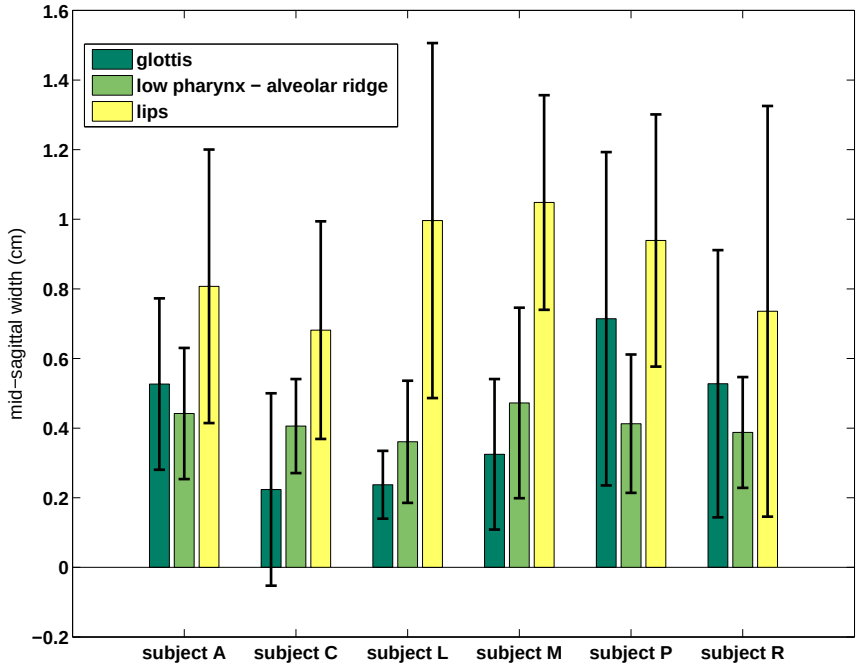


Figure 7.12: RMS error means and standard deviations of subjects grouped by region in the vocal tract: glottis, low pharynx up to alveolar region (middle region), and lip opening region. The difference of midsagittal width is highest at the lips where the boundary is not precise. The RMS errors in the region from low pharynx to alveolar ridge are similar.

for middle and labial regions, and $t = 0.87$ and $p = 0.42$ for glottal and middle regions).

7.5.3 Area function analysis

The majority of area function estimates in the literature are obtained from static MR images. To compare the area functions obtained from OASM segmentations, we selected the middle image in the duration of each phoneme articulation (from the images that are not included in the training set). The differences caused by coarticulatory effects is thus minimised when comparing the phoneme articulations. We compared the area function analysis of the six vowels {/ɑ/, /ʌ/, /ɔ/, /ɪ/, /i/, /u/}.

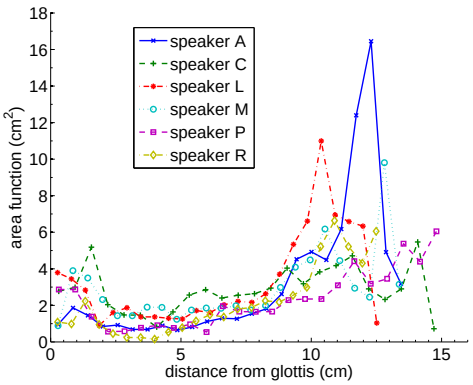
We first carried out a comparative analysis between our speaker's area functions. The plots in Figure 7.13 display the corresponding area functions of the speakers

for different vowels. Note that since the length of the vocal tracts are different, the curves' peaks and valleys may not exactly align and occur at different distances from the glottis in each speaker's plot. Nevertheless, for each vowel, the plots seem to follow a similar shape.

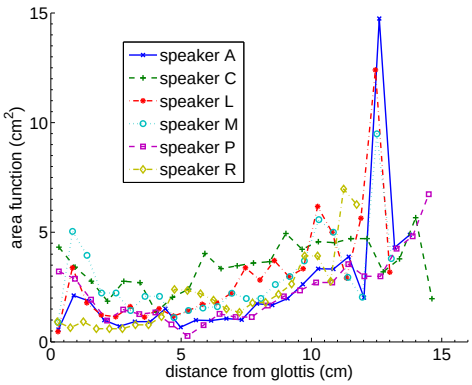
For /ɑ/ and /ʌ/ a wider aperture is expected at the front of the cavity, in contrast to a rather narrow airway at the pharynx. For /ɔ/, the plots mostly start at high values (no more than for the other phonemes), but drop and stay at low values for almost half of vocal tract length, suggesting a constriction at the oropharynx and velum area. This is followed by a sharp rise at the front of the cavity after the palatal area, followed by dramatic fall just at the furthest distances from the glottis, confirming the rounding of the lips (except for speaker P where the lips are affected by coarticulation). For /ɪ/, as opposed to /ɔ/, the general pattern of the area functions suggest a small gap at the front of the cavity (approximately 10 cm from the glottis), and a relatively big aperture at the pharynx. The patterns are clearer for /i/ and /u/ as there are fewer speakers (images of these phonemes were only available for a few speakers). Similar to /ɪ/, the plots suggest wider areas at the back of the tract and constrictions at the front of the cavity. The plots of /u/ illustrate a similar trend to /i/, however, the length of the tract seems to be slightly longer (compare speaker A's plots in the two diagrams). If we follow the patterns of the speakers individually, a longer vocal tract length can be observed in Figure 7.13 (c) and (f), due to the extension caused by rounding of the lips.

A comparison of phoneme area function patterns with area functions previously presented in the literature obtained from static MRI is presented in Figure 7.14. Note that for facilitating the comparisons we used the area functions of one male and one female speaker from the database.[1] The patterns of plots for each phoneme, agree along most of the vocal tract length, with exceptions of regions closest to the glottis and the lips. Note that the absolute values are not comparable in these plots since the physical features of the speakers are different and the area functions also

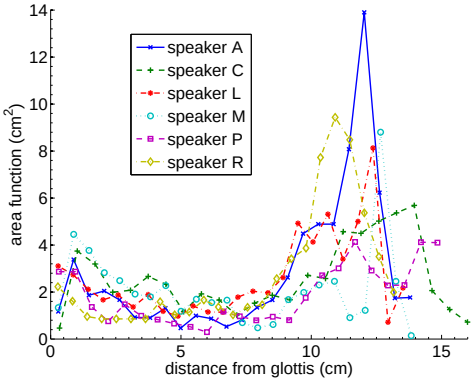
[1]We avoided the speakers with unrealistic sublingual cavity area estimates caused by missing teeth information to make the comparison simpler.



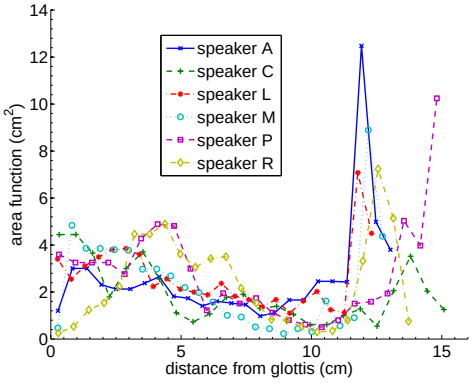
(a) phoneme /ɑ/



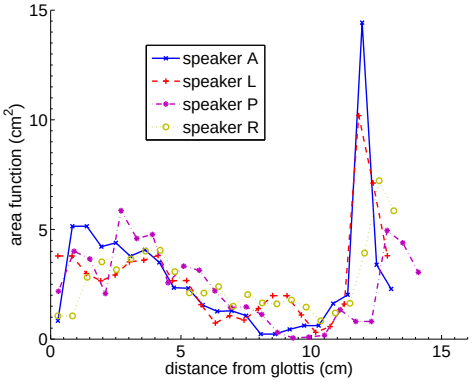
(b) phoneme /ʌ/



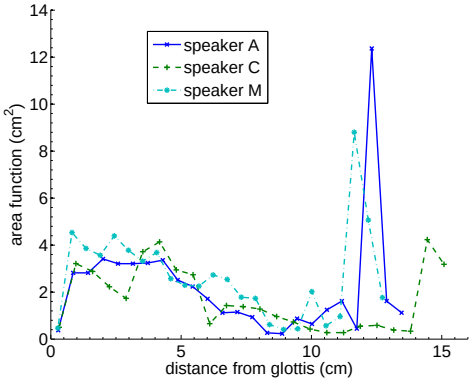
(c) phoneme /ɔ/



(d) phoneme /ɪ/

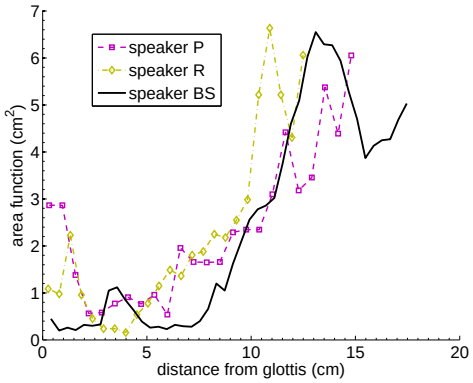


(e) phoneme /i/

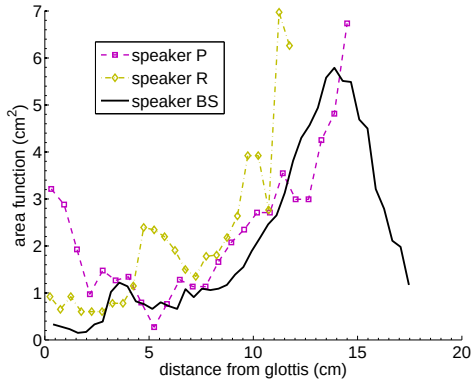


(f) phoneme /u/

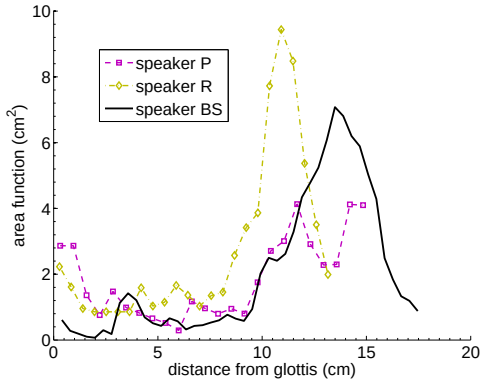
Figure 7.13: Area function plots for middle images of each phoneme, calculated using the area functions obtained from OASM segmentations.



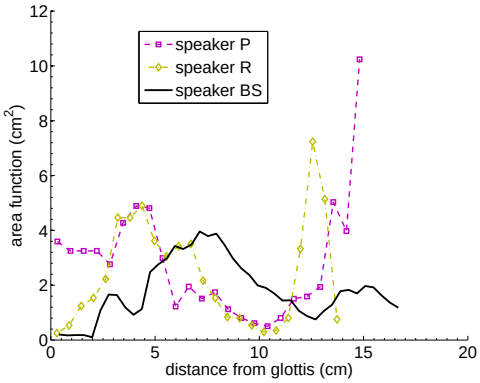
(a) phoneme /ɑ/



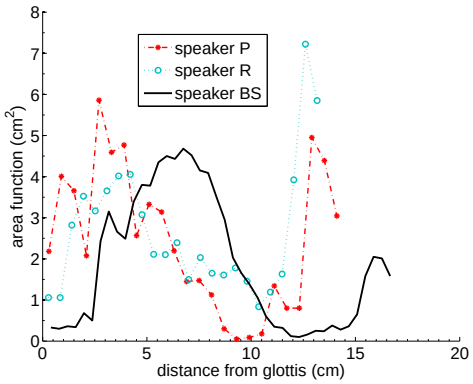
(b) phoneme /ʌ/



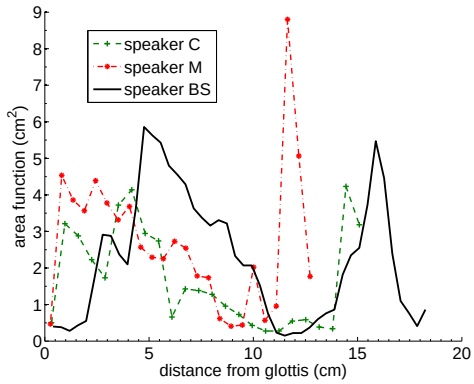
(c) phoneme /ɔ/



(d) phoneme /ɪ/



(e) phoneme /i/



(f) phoneme /u/

Figure 7.14: Area function plots for middle images of each phoneme, calculated using OASM area functions. Speaker BS's data is from [Story *et al.* 1996].

have been calculated using different methods. The expected patterns of openings and constrictions create different apertures and narrowing along the airway, and are generally consistent between our results and the data presented in Story *et al.* [1996].

7.5.4 Formant frequencies

To analyse the success of the segmentation approach further, we examined the vocal tract resonance frequencies estimated from the area functions. Unfortunately, estimation of *true* formant frequencies from the audio signals in the database is very difficult due to the MRI machine's noise and the narrow bandwidth of the microphone. We estimate the formant frequencies from the obtained area functions using VTAR software [Zhou *et al.* 2004]. The glottal source for estimation was set to the Rosenberg-Klatt [Klatt and Klatt 1990] glottal model. To analyse the obtained formant frequencies, we compare them with the formant frequencies in [Story *et al.* 1996]. Similar to area functions, the absolute values are not actually comparable, as the subjects are different, and the glottal source model is different. In addition, the values in [Story *et al.* 1996] are obtained from static MRI, rather than from running speech. We only use these values to perform a general comparison and assess the capability of the model in generating reasonable formants. Observing the formant frequencies and absolute values individually is not very informative, as it is the ratio of the formant frequencies to each other that is important in generating and distinguishing sounds.

We calculated the vocal tract resonance frequencies for all of the speakers. The numbers in Story *et al.* [1996] are from an adult male subject. Previous studies [Peterson and Barney 1952; Bennett 1981; Busby and Plant 1995] suggested that female speakers have higher formant frequencies compared to male speakers. The first three formants for all the six speakers and the speaker BS from [Story *et al.* 1996] are presented in Table 7.7. For comparison purposes, we have also included the set of formants generated from area functions of *generic* phonemes provided in VTAR. Note that the generic area functions are speaker-independent, and have been

either estimated using MR images [Zhou *et al.* 2004], or have been modelled based on Fant's area function models [Fant 1960]. To facilitate the analysis, we have visualised the values in Table 7.7 in form of scatter plots of first and second formants in

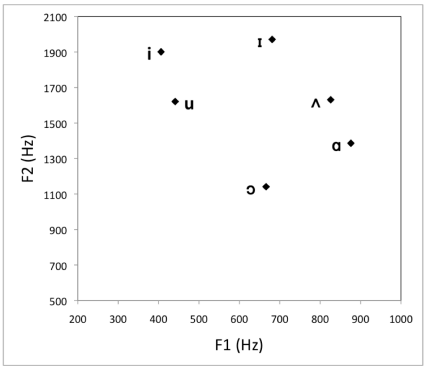
Table 7.7: First three formant frequencies based on the area functions of our speakers, and of speaker BS in Figure 7.14 (a) – (e). The generic area functions are available from VTAR software [Zhou *et al.* 2004]. Note that some of the vowels were not available for all speakers ("NA"). The values with † are the exceptions to the expected formant distribution pattern.

		vowels					
speaker	formant (Hz)	ɑ	ʌ	ɔ	ɪ	i	u
speaker A	F1	876	826	666	681	406	441
	F2	1386	1631	1141	1971	1901	1621
	F3	2811	2776	3026	2881	2871	2876
speaker L	F1	696	846	651	601	496	NA
	F2	1306	1641	1296	2226	2016	NA
	F3	2891	2911	2971	3506	2661	NA
speaker M	F1	731	711	421	421	NA	471
	F2	1576	1641	1061	2661	NA	2101†
	F3	2946	3081	2511	4381	NA	2586
speaker R	F1	851	1041†	821	441	601	NA
	F2	1596	2206†	1411	1646	2311	NA
	F3	2951	3011	2676	2971	3336	NA
speaker C	F1	551	621	516	446	NA	401
	F2	1271	1446	1121	1621	NA	2141†
	F3	2456	2616	2676	2416	NA	3236
speaker P	F1	721	656	626	451	291	NA
	F2	1451	1531	1466	2251	2546	NA
	F3	2591	2966	2711	2811	4161	NA
speaker BS	F1	826	706	641	516	301	326
	F2	1146	1301	1061	2021	2496	1121
	F3	2771	2581	2251	2571	3161	2461
generic vowel	F1	671	676	511	NA	241	301
	F2	1246	1306	1016	NA	2266	656
	F3	2461	2606	2536	NA	3131	2436

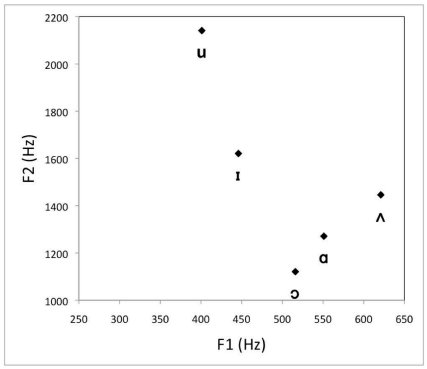
Figure 7.15. For most of the speakers, the obtained results are consistent with the expected articulation of the vowels. The first formant frequencies are consistently higher in low vowels such as /ɑ/ and /ʌ/ compared to high vowels such as /i/ and /ɪ/. The second formant follows a consistent pattern where it has lower frequencies for back vowels such as /ɑ/ and /ʌ/ and higher frequencies for front vowels /i/ and /ɪ/, in the majority of the speakers. The lip rounding in /ɔ/ leads to the lowest F2 among different vowels. The outliers to this patterns are marked by a † in the table.

The patterns discussed above can also be observed in the plots of Figure 7.15. In all of the plots, the low back vowels and high front vowels can be easily classified. The lower right triangle on the plots include the low back vowels /ɔ/, /ɑ/, and /ʌ/, while the upper left triangle includes front high vowels /i/ and /ɪ/. The exception to the overall patterns is the vowel /u/, that sometimes appears at unexpected locations on the plots. This may be because many young English speakers now pronounce /u/ as /ɨ/ [Hawkins and Midgley 2005].

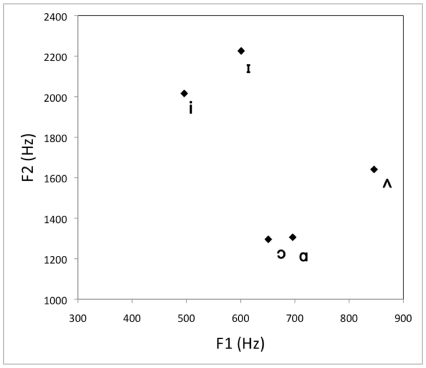
The formant frequencies of /ʌ/ for speaker P are examples of how the imprecision of length leads to incorrect formant estimates. The OASM segmentation returns a vocal tract that starts slightly above the glottis. The grid lines are consequently calculated from an area above the glottis. However, the format synthesis algorithm, in VTAR, assumes that the midsagittal distances start exactly at the glottis. Therefore, a big aperture that must be associated to the velum area is incorrectly assumed to be at the pharynx. Therefore, the vowel takes characteristics of both a low and a front vowel, resulting in high F2 frequency value.



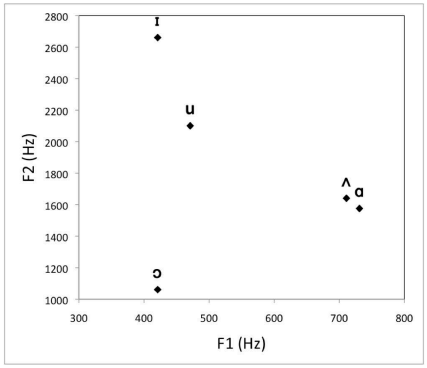
(a) speaker A



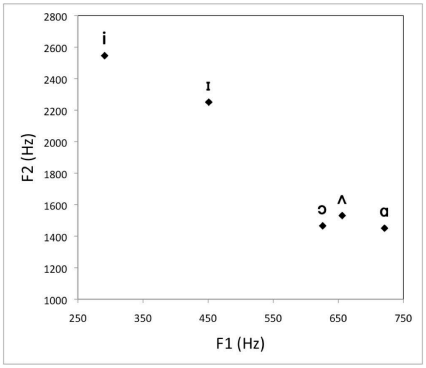
(b) speaker C



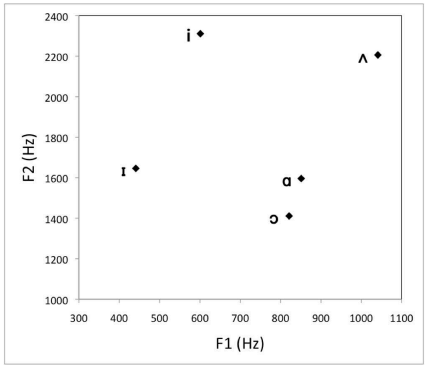
(c) speaker L



(d) speaker M



(e) speaker P



(f) speaker R

Figure 7.15: Formant plots for middle images of each phoneme, calculated using the area functions obtained from OASM segmentations.

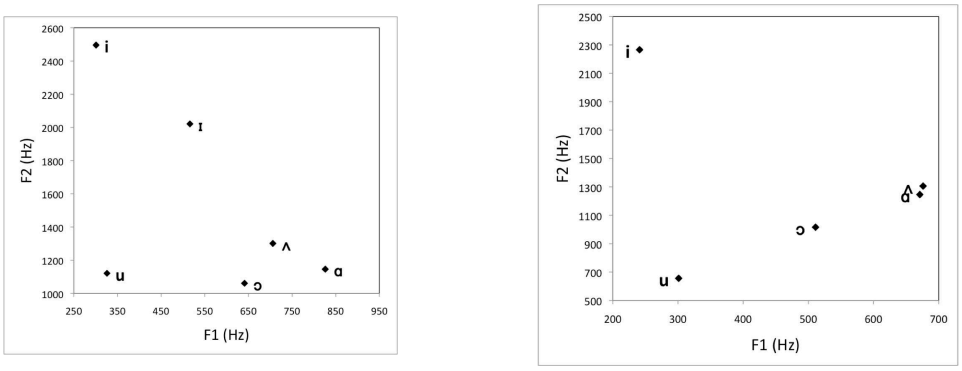
7.6 Discussion

As discussed in the above sections, the obtained area functions agree quite well with the expected pattern in previous work [Story *et al.* 1996]. The majority of

the area function plots are consistent in behaviour and rather similar in values in the middle of the vocal tract. By *middle* we are referring to all the regions along the vocal tract except for the beginning and end part of the tube. If we consider the dividing lines depicted in Figure 7.10, these regions approximately correspond to mid-pharynx, oropharynx, velum, and palate. Given the longer length of the former areas, this roughly corresponds to three-quarters of the vocal tract length. However, some inconsistencies are noticed at regions closer to the glottis, alveolar ridge and labial areas. These inconsistencies are in agreement with our quantitative evaluation results, where the maximum difference in terms of overlap between the gold standard and OASM segmentations varied between 13%–25%. In the rest of this section we explain some of the observed inconsistencies.

Coarticulation. Note that the area functions we obtained from OASM segmentations are from MR images of phonemes that are articulated in certain contexts. Therefore, the phoneme's articulation is always affected by coarticulation and does not necessarily have the characteristics of an isolated phoneme. This is important when comparing the area functions or formants with Story *et al.* [1996]'s suggested area functions. The area functions in Figure 7.14 (c) illustrate an example of this effect. The phoneme /ɔ/ in this example is in the context /dɔn/. The tongue tip is at a front position for dental constriction of /d/, and it moves only slightly back after release to take the posture for /ɔ/, and then moves towards the alveolar ridge again to take the constriction posture for /n/. In other words, the tongue tip is almost always somewhat raised.

In formant assessments also, we notice that although the general pattern is correct between the classes of vowels, there are some differences within the classified groups of patterns. In Figure 7.15, the placements of vowels /ɑ/ and /ʌ/, and /i/ and /ɪ/ show inconsistent patterns. This again can be related to the coarticulatory effects and the fact that they may not precisely follow the pattern of static speech articulations presented in Figure 7.16 (a) and (b). For instance, we examined the



(a) speaker BS

(b) Generic area functions

Figure 7.16: Formant plots for each phoneme calculated from area functions. Speaker BS's area functions are from [Story *et al.* 1996]. The generic area functions are available in VTAR software [Zhou *et al.* 2004].

case for vowel /u/ and compared the vocal tract shape and area functions with each other and with examples from [Story *et al.* 1996]. In our database the vowel /u/ is used in the triphone context of /sun/ (soon). The surrounding context makes the articulation of /u/ rather short, with the tongue in an up and front position instead of an up and back posture. In addition, as pointed out before, the contemporary pronunciation of /u/ is more similar to /ɨ/ among young British English speakers [Hawkins and Midgley 2005]. This can be observed in Figure 7.17. Consequently, vowel /u/ takes characteristics of a front vowel, with a higher F2 value than expected.

Glottis. In many of the area functions from OASM segmentations, the areas appear to be larger than expected at distances closer to the glottis. This is directly linked to the observation that the vocal tract lengths with OASM segmentation are occasionally shorter than the actual length of the vocal tract. Therefore, the region that is expected to be the glottis is actually the region slightly above the glottis. We illustrate this through Figure 7.18, where an example of a contour obtained by OASM segmentation in magenta is plotted over the original image, with the user-steered contours in green. This could be improved by manually adjusting the

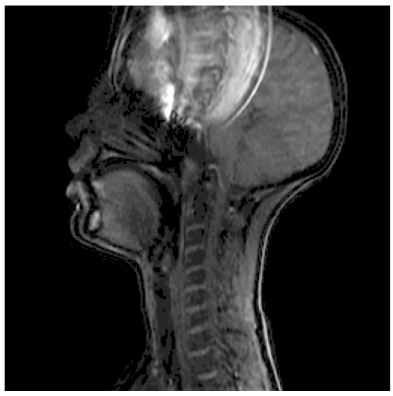


Figure 7.17: An MR image of articulation of phoneme /u/ in the context /sun/ (soon). The coarticulatory effects influence the typical articulation of /u/, forcing a more front vowel articulation.

landmarks around the glottis, so that more landmarks are tagged on the tiny airway that starts at the bottom of the epiglottis and ends at the glottis.

Lips, sublingual cavity and teeth. Another — expected — inconsistency observed is at the front of the cavity, roughly at the last 3–4 gridlines on the contour. The missing teeth information in this area affects the vocal tract midline and consequently the gridlines, hence area functions at the front of the cavity. This results in generating surprisingly large area function values at the front. This can be observed

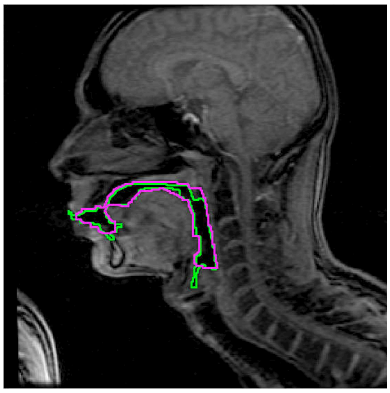


Figure 7.18: The OASM segmentation results (magenta) and user-steered contour (green). The length of the tract is estimated to be shorter in the OASM segmentation. The user-steered contour ends at the glottis, while the lowest point on the OASM contour is slightly above the glottis.

in the area function plots in Figures 7.13 and 7.14. Providing the dentition information, either by using casts of speakers' teeth, or by covering them with materials so that they appear in the MR images during imaging, could resolve this issue.

7.7 Conclusion

In this chapter, we evaluated the obtained vocal tract shapes from OASM segmentations. The results were evaluated separately in two parts: from an image processing perspective and from articulatory modelling perspective. In image-based evaluations, the results were compared with the set of user-steered segmentations of the same images. The results suggested a Dice similarity of 76.51% to 87.71% between the OASM and user-steered segmentations.

We studied the nature of the dissimilarities between the OASM results and user-steered segmentations by analysing the results from an articulation modelling view, and comparing the midsagittal widths and area functions. The results revealed that the general pattern of midsagittal widths along the vocal tract length seems to agree with the user-steered segmentations, with a maximum average RMS error of 0.58 cm. The comparisons of area functions within speakers suggested that the area function plots mostly follow a similar trend for each phoneme. The obtained area functions, from middle images in the phoneme durations, mostly agree with the trend suggested in the literature.

Finally, to evaluate the OASM segmentations further, we estimated formants from the obtained tube models for the middle images of phonemes. The outcome proved that generated formants are consistent with prior formants frequencies measurements of phoneme categories.

In general, the proposed framework proved to be a promising approach for extracting the vocal tract shape from MR images by capturing the features of articulation that have vital roles in generating sounds and distinguishing between them.

Chapter 8

Conclusions

The overall aim of this thesis was to investigate and develop a range of automatic methods for studying human articulation using large MRI databases. We achieved this goal by addressing the several objectives of this research, as presented in section 1.2. In the remainder of this section, we provide a summary of the work accomplished in this dissertation, and we discuss how our initial objectives were addressed.

We started by studying articulation using the basis of MR images: grey-level pixel intensities. The outcome of this step was deriving "typical" articulatory configurations of each phoneme, for each speaker. The results suggested that the typical articulations derived from continuous speech is consistent with descriptions of static articulations reported in previous literature, but the dynamics of context are reflected in the articulation through the gradation of grey-levels at the boundaries of the vocal tract. We further studied the dynamics by measuring the amount of variation in the articulation of each phoneme, and estimating the correlation between the articulatory movements and the energy in the generated speech. We concluded that to be able to study articulation more comprehensively, a method that facilitates a parametric description of the shape is necessary. This inspired us to investigate statistical methods of feature extraction from images. Our initial experiments using fuzzy connectedness demonstrated that a purely region-based segmentation method cannot be autonomously used to extract the correct shape from images. The inad-

equacy of the region-based approach was mainly due to the discontinuities in the vocal tract boundaries, connecting the airway to other areas with similar intensities to the vocal tract. A framework for automatic segmentation was proposed based on the OASM, which can address the vocal tract boundary segmentation challenges more effectively than purely region-based segmentation methods, by using both image intensity and shape model information in finding and delineating the shape.

Through the steps above, we addressed the objectives of this thesis (stated in section 1.2), as explained below.

- Studying the dynamics of articulation and general vocal tract shape using MR images of articulation.

  We investigated the dynamics of articulation using the pixel-based approaches proposed in chapter 4. The articulators' movements and vocal tract shape variations in a typical articulation approach were analysed by examining the average shape and gradation of grey-levels around the edges. The articulators' dynamics were further studied and measured by parametrising the amount of movements in the vocal tract involved in production of different sounds.

- Devising methods to extract articulatory features from images, with minimal or zero human supervision.

  We introduced new methods for obtaining articulatory information such as articulators' movements and displacements and the degree of articulators' dynamics. These features are obtained automatically with computer scripts once the data is prepared. We also introduced a segmentation method for extracting the vocal tract contours from the images. In the OASM framework, following the initial data preparation, the training and segmentation are performed automatically, without requiring further human supervision.

- Utilising spatial and temporal information of the images to investigate articulatory movements.

  We employed pixel intensity attributes, distributions and variances in se-

quences of images to study the nature and extent of articulatory movements. The average articulation research involved spatial analysis of the pixel's attributes (i.e. grey-levels), as we were looking for areas of more stability compared to the areas of frequent/extreme distortions. Further, the magnitude of change in pixel intensities over computed time intervals was utilised to explore the dynamics of articulation.

- Devising automatic methods for segmenting new images for parametric modelling of the vocal tract.

  We investigated the techniques applied in the literature for vocal tract segmentation, and demonstrated that traditional methods are limited and not sufficient for automatic segmentation of new images without human supervision (chapter 5). We introduced a new framework in chapter 6 for automatic recognition and delineation of vocal tract boundaries. The results, presented in form of contours or shapes, can be used for parametric vocal tract shape description.

- Assessing the applicability and efficacy of the proposed methods.

  We extensively evaluated the performance and results of the OASM framework from image-processing aspects as well as articulation modelling aspects. The evaluation outcomes suggested that the OASM framework is indeed a valid approach for accurate shape extraction from vocal tract MR images.

8.1 Limitations

The MRI database we used in this thesis was collected for another purpose, a project for investigating pharynx and velum movements. Thus, dentition information was not collected before or during scanning. It would be valuable to evaluate the performance of the proposed framework on a database with available dentition information and compare the results with those reported in the literature. Also, the speech acoustics recorded were not qualitatively and quantitatively sufficient to be used for

studying the relation of speech acoustics and articulation, as they were recorded with a very narrow bandwidth microphone to minimise the scanner noise in the recordings. Although the acoustics recorded were the best that can be obtained inside a scanner, in future we would like to collect or use a database with better quality, for example by recording the acoustics outside the scanner separately as well as inside the scanner. A more generous quantity of acoustic signals to study relationships between the articulation and generated acoustics in a more comprehensive way, e.g. by using data-driven approaches, is also desirable. We would like to use a dynamic or real-time MRI database to create a new baseline system for studying the acoustic and articulations from magnetic resonance images of vocal tract.

The automatic segmentation framework presented in this thesis was applied to 2D images. Ideally, we would like to apply the automatic segmentation to 3D images. Since, both RBS automatic landmark tagging and OASM can easily be adapted to 3D images, this should not be very difficult. It would be interesting to examine the performance of the method on 3D images as well.

In future work, we would like to improve the performance of the OASM segmentation framework by using other methods of automatic landmark tagging, or manually placing some of the landmarks that are "points of interest" but which the landmark tagging approach failed to recognise. In addition, it would be interesting to investigate how the framework can be modelled on multiple speakers to generate one OASM and to recognise images of new speakers.

Finally, we would like to apply the segmentation technique to study the process of articulation in more detail. For example, typical articulation can be modelled with a procrustes analysis of the obtained models for each phoneme, and the amount of variation from the typical shape could be measured quantitatively, in each context. Another interesting research direction would be using procrustes analysis on all images of each phoneme to obtain a general parametric shape structure for modelling the vocal tract and its variations. This could be applied to articulatory speech synthesis.

8.2 Contributions

The main contribution of this thesis is introducing new methods for speech production analysis that facilitate automatic feature extraction from large databases of MR images of articulation. Through these methods, novel approaches for investigating articulation in dynamic MRI have been proposed. We generalise our research by providing a framework for automatic shape extraction that is applicable to other methods of studying speech production, such as area function analysis. The most important contributions of this thesis are summarised below.

8.2.1 Pixel-based image analysis

We first explored articulatory configuration and articulation dynamics using only image-based properties instead of using shape priors and detecting air–tissue boundaries. The suggested approach can be applied to dynamic MRI databases where substantial variation is present and the articulation of phonemes is affected by surrounding contexts.

Speaker-specific typical vocal tract shapes. We carried out a new study in observing the articulation in running speech. Instead of looking at single instances of articulation of phonemes, we took all the coarticulatory information into account. We averaged the images that were captured during the articulation of a phoneme for each speaker. The range of images for each phoneme was determined using the duration of each phoneme in the recorded speech, obtained by means of forced alignment. The images inevitably included all the coarticulatory effects caused by surrounding context. Averaging the images enabled us to distinguish between the pixels which had roughly a constant value of intensity and those that change in intensity. Further, typical articulation provided us with information about the position and movements of the articulators during speaking. The regions of consistent higher intensity values depicted less movements, while other regions engaged in movements had lower intensity values, hence more blurred presentation, e.g. around the edges of the artic-

ulators where movements occurred. The areas with higher intensity values therefore were less affected by extreme movements and represented the typical articulatory posture of each phoneme. The typical articulatory configurations were consistent with observations reported about the static articulatory postures of phonemes in the literature. The bonus in the obtained images was the information it provided regarding the variance in the articulation resulting from coarticulation.

Correlating speech acoustics and articulation. To further use the information obtained from intensity features of the images, we quantified the articulators' movements in the articulation of a phoneme. The intensity values of the pixels in the *difference* images were calculated to estimate the extent of movements involved in articulation of a phoneme. We calculated the amount of energy in the sound generated for each phoneme. The correlation between the variance estimated from the images and the amount of energy present in the generated acoustics were investigated. As expected, a positive correlation was observed, meaning that there was a direct relationship between the amount of articulatory movements and the amount of energy embedded in the generated speech. The latter observation confirms that intensity information of the images can indeed provide us with useful information about the articulation of speech.

8.2.2 Automatic shape extraction

Region-based methods: fuzzy connectedness. We demonstrated that traditional region-based segmentation methods for extracting the shape are not adequate when a substantial number and range of images are to be processed. This is mostly due to the fact that vocal tract is not a closed boundary object and is connected to other areas with similar properties to the tract airway, such as the nasal cavity and the surrounding air. Thus, boundaries may leak to other areas if there is a discontinuity in the boundary of the vocal tract. Traditionally, the application of these methods has been accompanied by manual corrections or artificial blocking of

the leaks. None of these are applicable in our work as we aim for minimum manual supervision and some of the areas where leakage happens are dynamic and cannot be easily tracked and blocked.

OASM segmentation framework. We introduced a novel framework for automatic extraction of the vocal tract shape using existing image segmentation techniques. The framework uses an algorithm that involves both model-based and boundary-based segmentation techniques, and consequently considers both the generic vocal tract shape and intensity properties of the image in segmenting new images. To provide the initial landmarks we applied an automatic approach for landmark tagging. Incorporating all these methods makes the model automatic and less dependent on human experts than current alternatives, hence less tedious and error-prone. We demonstrated that if the models are trained per speaker specifically, the method can successfully recognise and delineate the vocal tract in new images. The framework is easy to apply since most of the steps involved in developing the model and applying the model for segmentation are automatic, except for the initial segmentation of training images by human expert.

Evaluation. We carried out image-based and phonetic-based assessments to analyse the performance and efficiency of the framework both from an image processing and articulatory modelling perspectives. These evaluations demonstrated that the new framework proposed in this thesis is very successful in extracting the vocal tract shapes from MR images.

8.3 Final remarks

Understanding human speech production is an important step in speech communication research. Articulatory information can be applied to find solutions for speaking disorders, in learning foreign languages, and to improve automatic speech recognition and synthesis.

In this thesis, we investigated speech articulation using MR images of different speakers' vocal tracts during running speech. More specifically, we focused on approaches for automatic extraction of useful articulatory information from images. We introduced novel methods for studying articulation by considering the temporal and spatial information of the images in a sequence. We proposed a new framework based on the existing image segmentation techniques for extracting parametric information from large scale MR databases of articulation. We evaluated the success of the proposed framework by investigating the outcome from image processing and phonetics aspects.

Overall, our novel contributions provide new tools and techniques for investigating speech production, and facilitate the data-driven analysis and modelling of articulation.

Bibliography

The British National Corpus. `http://www.natcorp.ox.ac.uk/`, accessed November 2011.

The CMU pronunciation dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudic`, accessed November 2011.

R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.

C. Alvey, C. Orphanidou, J. Coleman, A. McIntyre, S. Golding, and G. Kochanski. Image quality in non-gated versus gated reconstruction of tongue motion using magnetic resonance imaging: A comparison using automated image processing. *International Journal of Computer Assisted Radiology and Surgery*, 3(5):68–73, 2008.

A. Alwan, S.S. Narayanan, and K. Haker. Toward articulatory-acoustic models for liquid consonants based on MRI and EPG data. Part II: The rhotics. *Journal of the Acoustical Society of America*, 101(2):1078–1089, 1997.

J.K.F. Anthony. Replica of the vocal tract. *UCLA Working Papers in Phonetics*, 10:10–14, 1964.

M.S. Avila-García. *Automatic 3D vocal tract shape extraction from Magnetic Resonance Image sequences*. PhD Thesis, University of Southampton, Southampton, UK, 2007.

M.S. Avila-García, J.N. Carter, and R.I. Damper. Extracting tongue shape dynamics from magnetic resonance image sequences. *Transactions on Engineering, Computing and Technology*, 2:216–219, 2005.

P. Badin, G. Bailly, M. Raybaudi, and C. Segebarth. A three-dimensional linear articulatory model based on MRI data. In *Proceedings of the 3rd ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, pages 249–254, Blue Mountains, Australia, 1998.

P. Badin, G. Bailly, L. Reveret, M. Baciu, C. Segebarth, and C. Savariaux. Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics*, 30(3):533–553, 2002.

P. Badin, I. Makarov, and V. Sorokin. Algorithm for calculating the cross-section areas of the vocal tract. *Acoustical Physics*, 51(1):38–43, 2005.

T. Baer. Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *Journal of the Acoustical Society of America*, 90(2):799–828, 1991.

T. Baer, J. Gore, S. Boyce, and P. Nye. Application of MRI to the analysis of speech production. *Magnetic Resonance Imaging*, 5(1):1–7, 1987.

T. Baer, J. Gore, L. Gracco, and P. Nye. Vocal tract dimensions obtained from magnetic resonance images. *Journal of the Acoustical Society of America*, 84: S125, 1988.

L. Baghai-Ravary. Automatic differentiation between accents of native and non-native English, and the significance of prosody. In *Proceedings of the 5th International Conference of Speech Prosody*, pages 100204:1–4, Chicago, IL, 2010.

L. Baghai-Ravary, S. Grau, and G. Kochanski. Detecting gross alignment errors in the Spoken British National Corpus. *The Computing Research Repository (CoRR)*, abs/1101.1682, 2011.

L.E. Baum. An inequality and an associated maximisation technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities*, 3(1):1–8, 1972.

D. Beautemps, P. Badin, G. Bailly, A. Galván, and R. Laboissière. Evaluation of an articulatory–acoustic model based on a reference subject. In *Proceedings of the 4th Speech Production Seminar/ETRW*, pages 45–48, Autrans, France, 1996.

S. Bennett. Vowel formant frequency characteristics of preadolescent males and females. *Journal of the Acoustical Society of America*, 69(1):231–238, 1981.

E. Bresch and S.S. Narayanan. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE Transactions on Medical Imaging*, 28(3):323–338, 2009.

E. Bresch, J. Adams, A. Pouzet, S. Lee, D. Byrd, and S.S. Narayanan. Semi-automatic processing of real-time MR image sequences for speech production studies. In *Proceedings of the 7th International Seminar on Speech Production (ISSP)*, pages 427–434, Ubatuba, Brazil, 2006.

P.A. Busby and G.L. Plant. Formant frequency values of vowels produced by preadolescent boys and girls. *Journal of the Acoustical Society of America*, 97(4):2603–2606, 1995.

J. Busset and Y. Laprie. Adaptation of cepstral coefficients for acoustic-to-articulatory inversion. In *Proceedings of the 9th International Seminar on Speech Production (ISSP)*, pages 95–102, Montréal, Canada, 2011.

I. Carbone, P. Martins, A. Teixeira, and A. Silva. A vocal tract segmentation and analysis over a European Portuguese MRI database. *Electrónica e Telecomunicações*, 4(9):1050–1053, 2013.

T. Chiba and M. Kajiyama. *The Vowel: Its Nature and Structure.* Tokyo-Kaiseikan Publishers Ltd., Tokyo, Japan, 1941.

K.C. Ciesielski and J.K. Udupa. Affinity functions in fuzzy connectedness based image segmentation II: Defining and recognizing truly novel affinities. *Computer Vision and Image Understanding*, 114(1):155–166, 2010.

J. Coleman, M. Liberman, G. Kochanski, L. Burnard, and J. Yuan. Mining a year of speech. New tools and methods for very-large-scale phonetics workshop, Philadelphia, `http://www.phon.ox.ac.uk/jcoleman/MiningVLSP.pdf`, 2011, accessed March 2012.

T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models–their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

T.F. Cootes, C.J. Twining, K.O. Babalola, and C.J. Taylor. Diffeomorphic statistical shape models. *Image and Vision Computing*, 26(3):326–332, 2008.

M. A. Crary, I. M. Kotzur, J. Gauger, M. Gorham, and S. Burton. Dynamic magnetic resonance imaging in the study of vocal tract configuration. *Journal of Voice*, 10 (4):378–388, 1996.

J. Dang, K. Honda, and H. Suzuki. MRI measurements and acoustic investigation of the nasal and paranasal cavities. *Journal of the Acoustical Society of America*, 94(3):1765–1765, 1993.

L. Davidson. Ultrasound as a tool for speech research. *The Oxford Handbook of Laboratory Phonology*, pages 484–496, 2012.

R.H. Davies, C.J. Twining, T.F. Cootes, J.C. Waterton, and C.J. Taylor. A minimum description length approach to statistical shape modelling. *IEEE Transactions on Medical Imaging*, 21(5):525–537, 2001.

D. Demolin, T. Metens, and A. Soquet. Three-dimensional measurement of the vocal tract by MRI. In *Proceedings of the 4th International Conference on Spoken Language (ICSLP)*, pages 272–275, Philadelphia, PA, 1996.

D. Demolin, M. George, V. Lecuit, T. Metens, A Soquet, and H. Raeymaekers. Coarticulation and articulatory compensations studied by dynamic MRI. In *Proceedings of 5th European Conference on Speech Communication and Technology (Eurospeech)*, pages 31–34, Rhodes, Greece, 1997.

D. Demolin, T. Metens, and A. Soquet. Real time MRI and articulatory coordinations in vowels speech production. In *Proceedings of the 5th Speech Production Seminar (ISSP)*, pages 86–93, Kloster Seeon, Germany, 2000.

D. Demolin, S. Hassid, T. Metens, and A. Soquet. Real-time MRI and articulatory coordination in speech. *Comptes Rendus Biologies*, 325(4):547–556, 2002.

S.R. Eddy. What is dynamic programming? *Nature Biotechnology*, 22(7):909–910, 2004.

W. Engelke, T. Bruns, M. Striebeck, and G. Hoch. Midsagittal velar kinematics during production of VCV sequences. *The Cleft Palate Craniofac Journal*, 33(3): 236–244, 1996.

O. Engwall. Are statical MRI data representative of dynamic speech? Results from a comparative study using MRI, EMA and EPG. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, volume 3, pages 17–20, Beijing, China, 2000.

O. Engwall. Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. *Speech Communication*, 41(2–3):303–329, 2003.

O. Engwall and P. Badin. Collecting and analysing two and three-dimensional MRI data for Swedish. *Department of Speech, Music and Hearing Quarterly Progress and Status Report - KTH (TMH-QPSR)*, 40(3–4):11–38, 1999.

A.X. Falcão, J.K. Udupa, S. Samarasekera, S. Sharma, B.E. Hirsch, and R.A. Lotufo. User–steered image segmentation paradigms: Live wire and live lane. *Graphical Models and Image Processing*, 60(4):233–260, 1998.

G. Fant. *Acoustic Theory of Speech Production with Calculations based on X-Ray studies of Russian Articulations*. The Hague: Mouton, 1960.

A. Foldvik, U. Kristiansen, J. Kvaerness, A. Torp, and H. Torp. Three-dimensional ultrasound and magnetic resonance imaging: a new dimension in phonetic research. In *Proceedings of the 13th International Congress of Phonetic Sciences (ICPhS)*, volume 4, pages 46–49, Stockholm, Sweden, 1995.

A.F. Frangi, W.J. Niessenand D. Rueckert, and J.A. Schnabel. Automatic 3D ASM construction via atlas-based landmarking and volumetric elastic registration. In *Proceedings of the 17th International Conference on Information Processing in Medical Imaging (IPMI)*, pages 78–91, London, UK, 2001.

J. Frankel, K. Richmond, S. King, and P. Taylor. Continuous speech recognition using articulatory data. In *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP)*, pages 254–257, Beijing, China, 2000.

O. Fujimura, S. Kiritani, and H. Ishida. Computer controlled radiography for observation of movements of articulatory and other human organs. *Computers in Biology and Medicine*, 3(4):371–384, 1973.

B. Gick. The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association*, 32(2):113–121, 2002.

H. Gray Jr and J. D. Markel. Distance measures for speech processing. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(5):380–391, 1976.

A.R. Greenwood, C.C. Goodyear, and P.A. Martin. Measurements of vocal tract shapes using magnetic resonance imaging. *IEEE Proceedings in Communications, Speech and Vision*, 139(6):553–560, 1992.

G. Hamarneh. Accessed january 2013. Livewire, 2005. URL `http://www.cs.sfu.ca/~hamarneh/software/livewire/index.html`.

W.J. Hardcastle. The use of electropalatography in phonetic research. *Phonetica*, 25:197–215, 1972.

R. Harshman, P. Ladefoged, and L. Goldstein. Factor analysis of tongue shapes. *Journal of the Acoustical Society of America*, 62(3):693–707, 1977.

S. Hawkins and J. Midgley. Formant frequencies of RP monophthongs in four age groups of speakers. *Journal of the International Phonetic Association*, 35:183–199, 2005.

T. Heimann, B. Van Ginneken, M.A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, F. Bello, G. Binnig, H. Bischof, A. Bornik, P. Cashman, Ying Chi, A. Cordova, B.M. Dawant, M. Fidrich, J.D. Furst, D. Furukawa, L. Grenacher, J. Hornegger, D. Kainmuller, R.I. Kitney, H. Kobatake, H. Lamecker, T. Lange, J. Lee, B. Lennon, Rui Li, Senhu Li, H.-P. Meinzer, G. Nemeth, D.S. Raicu, A.-M. Rau, E.M. van Rikxoort, M. Rousson, L. Rusko, K.A. Saddi, G. Schmidt, D. Seghers, A. Shimizu, P. Slagmolen, E. Sorantin, G. Soza, R. Susomboon, J.M. Waite, A. Wimmer, and I. Wolf. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Transactions on Medical Imaging*, 28(8):1251–1265, 2009.

J.M. Heinz and K.N. Stevens. On the derivation of area functions and acoustic spectra from cineradiographic films of speech (A). *Journal of the Acoustical Society of America*, 36:1037, 1964.

P. Hoole, A. Wismüller, G. Leinsinger, C. Kroos, A. Geumann, and M. Inoue. Analysis of tongue configuration in multi-speaker, multi-volume MRI data. In *Proceedings of the 5th Speech Production Seminar (ISSP)*, pages 157–160, Kloster Seeon, Germany, 2000.

P. Hough. Method and Means for Recognizing Complex Patterns. U.S. Patent 3.069.654, December 1962.

D.P. Huttenlocher, G.A. Klanderman, G.A. Kl, and W.J. Rucklidge. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:850–863, 1993.

F. Itakura and S. Saito. A statistical method for estimation of speech spectral density and formant frequencies. *Electronics and Communication in Japan*, 52-A: 36–43, 1970.

C. Johansson, J. Sundberg, H. Wilbrand, and C . Ytterbergh. From sagittal distance to area: A study of transverse, cross-sectional area in the pharynx by means of computed tomography. *Department of Speech, Music and Hearing Quarterly Progress and Status Report - KTH (KTH STL-QPSR)*, 4:39–49, 1983.

D. Jurafsky and J.H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.* Prentice Hall, second edition, 2008.

M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

C.A. Kelsey, R.J. Woodhouse, and F.D. Minifie. Ultrasonic observations of coarticulation in the pharynx. *Journal of the Acoustical Society of America*, 46(4B): 1016–1018, 1969.

S.A. King and R.E. Parent. Creating speech-synchronized animation. *IEEE Transactions on Visualization and Computer Graphics*, 11(3):341–352, May 2005.

S. Kiritani, Y. Tateno, T. Iinuma, and M. Sawashima. Computed tomography of the vocal tract. *Dynamic Aspects of Speech Production, current results, emerging problems, and new instrumentation*, pages 203–206, 1977.

D.H. Klatt and L.C. Klatt. Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857, 1990.

G. Kochanski and J. Coleman. Accessed december 2012. articulation and coarticulation in lower vocal tract, 2008. URL `http://www.esrc.ac.uk/my-esrc/grants/RES-000-23-1094/read`.

B.J. Kröger, R. Winkler, C. Mooshammer, and B. Pompino-Marschall. Estimation of vocal tract area function from magnetic resonance imaging: Preliminary results. In *Proceedings of the 5th Speech Production Seminar (ISSP)*, pages 333–336, Kloster Seeon, Germany, 2000.

P. Ladefoged, J. Anthony, and C. Riley. Direct measurement of the vocal tract. *UCLA Working Papers in Phonetics*, 19:4–13, 1971.

A. Lammert, M.I. Proctor, and S.S. Narayanan. Data-driven analysis of real-time vocal tract MRI using correlated image regions. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1572–1575, Makuhari, Japan, 2010.

A. Lammert, V. Ramanarayanan, M.I. Proctor, and S.S. Narayanan. Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 959–962, Lyon, France, 2013.

Y. Laprie, M. Loosvelt, S. Maeda, R. Sock, and F. Hirsch. Articulatory copy synthesis from cine X-ray films. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech)*, Lyon, France, 2013.

S.U. Lee, S.Y. Chung, and R.H. Park. A comparative performance study of several global thresholding techniques for segmentation. *Computer Vision, Graphics, and Image Processing*, 52(2):171–190, 1990.

J. Liu and J.K. Udupa. Oriented active shape models. *IEEE Transactions on Medical Imaging*, 28(4):571–584, 2009.

A. Loukina, G. Kochanski, C. Shih, and E. Keane. Rhythm Measures with Language Independent Segmentation. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1531–1534, Brighton, UK, 2009.

J. Ma and R. Cole. Animating visible speech and facial expressions. *The Visual Computer*, 20(2-3):86–105, 2004.

M. Mády, R. Sader, A. Zimmermann, P. Hoole, A. Beer, H. Zeilhofer, and C. Hanning. Use of real-time MRI in assessment of consonant articulation before and after tongue surgery and tongue reconstruction. In *Proceedings of the 4th International Speech Motor Conference*, pages 142–145, Nijmegen, Netherlands, 2001.

S. Maeda. *On the conversion of vocal tract X-ray data into formant frequencies.* Bell Laboratories, Murray Hill, NJ, 1972.

S. Maeda. Improved articulatory models. *Journal of the Acoustical Society of America*, 84(S1):S146, 1988.

S. Marsland and C. J. Twining. Constructing data-driven optimal representations for iterative pairwise non-rigid registration. *Biomedical Image Registration*, 2717: 50–60, 2003.

S. Masaki, M. K. Tiede, K. Honda, Y. Shimada, I. Fujimoto, Y. Nakamura, and N. Ninomiya. MRI observation of dynamic articulatory movements using a synchronized sampling method. *Journal of the Acoustical Society of America*, 102(5): 3166, 1997.

I. Matthews, T.F. Cootes, J. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):198–213, 2002.

P. Mermelstein. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53(4):1070–1082, 1973.

P.H. Milenkovic, S. Yaddanapudi, H.K. Vorperian, and R.D. Kent. Effects of a curved vocal tract with grid-generated tongue profile on low-order formants. *Journal of the Acoustical Society of America*, 127:1002, 2010.

M. Mohammad, E. Moore, J.N. Carter, C.H. Shadle, and S.R. Gunn. Using mri to image the moving vocal tract during speech. In *Proceedings of 5th European Conference on Speech Communication and Technology (Eurospeech)*, volume 4, pages 2027—2030, Rhodes, Greece, 1997.

C.A. Moore. The correspondence of vocal tract resonance with volumes obtained from magnetic resonance images. *Journal of Speech and Hearing Research*, 35(5): 1009–1023, 1992.

M. Mrayati, R. Carré, and B. Guerin. Distinctive regions and modes: a new theory of speech production. *Speech Communication*, 7(3):257–286, 1988.

K.G. Munhall, E. Vatikiotis-Bateson, and Y. Tohkura. X-ray film database for speech research. *Journal of the Acoustical Society of America*, 95(5):2822, 1998.

S.S. Narayanan, A. Alwan, and K. Haker. An articulatory study of fricative consonants using Magnetic Resonance Imaging. *Journal of the Acoustical Society of America*, 98(3):1325–1347, 1995.

S.S. Narayanan, A. Alwan, and K. Haker. Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals. *Journal of the Acoustical Society of America*, 101(2):1064–1077, 1997.

S.S. Narayanan, D. Byrd, and A. Kaun. Geometry, kinematics, and acoustics of Tamil liquid consonants. *Journal of the Acoustical Society of America*, 106(4): 1993–2007, 1999.

S.S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd. An approach to real-time magnetic resonance imaging for speech production. *Journal of the Acoustical Society of America*, 115(4):1771–1776, 2004.

L.G. Nyúl and J.K. Udupa. On standardizing the MR image intensity scale. *Magnetic Resonance in Medicine*, 42(6):1072–1081, 1999.

D. Ong and M. Stone. Three-dimensional vocal tract shapes in /r/ and /l/: A study of MRI, ultrasound, electropalatography, and acoustics. *Phonoscope*, 106(1):1–13, 1998.

J.S. Perkell. *Physiology of Speach Production: Results and Implications of a Quantitative Cineradiographic Study*. The MIT Press, 1969.

J.S. Perkell, M.H. Cohen, M.A. Svirsky, M.L. Matthies, I. Garabieta, and M.T. Jackson. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America*, 92(6): 3078–3096, 1992.

P. Perrier, L. Boë, and R. Sock. Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: Modeling the transition with two sets of coefficients. *Journal of Speech and Hearing Research*, 35:53–87, 1992.

G.E. Peterson and H.L. Barney. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2):175–184, 1952.

M.I. Proctor, D. Bone, A. Katsamanis, and S.S. Narayanan. Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1576–1579, Makuhari, Japan, 2010.

M.I. Proctor, A.C. Lammert, A. Katsamanis, L. Goldstein, C. Hagedorn, and S.S. Narayanan. Direct estimation of articulatory kinematics from real-time magnetic

resonance image sequences. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 281–284, Florence, Italy, 2011.

K. Rohr. Extraction of 3D anatomical point landmarks based on invariance principles. *Pattern Recognition*, 32(1):3–15, 1999.

M. Rokkaku, K. Hashimoto, S. Imalzumli, S. Nilmi, and S. Kiritani. Measurements of the three dimensional shape of the vocal tract based on the magnetic resonance imaging technique. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, 20:47–54, 1986.

D. Rueckert, A.F. Frangi, and J.A. Schnabel. Automatic construction of 3D statistical deformation models using non-rigid registration. In *Medical Image Computing and Computer-Assisted Intervention, MICCAI 2001*, volume 2208 of *Lecture Notes in Computer Science*, pages 77–84. Springer Berlin–Heidelberg, 2001.

S. Rueda and J. K. Udupa. Global-to-local, shape-based, real and virtual landmarks for shape modelling by recursive boundary subdivision. In *Proceedingd of SPIE Medical Imaging 2011: Image Processing*, volume 7962, pages 796247–1–796247–13, 2011.

S. Rueda and J.K. Udupa. nD statistical shape model building via recursive boundary subdivision. In *Proceedings of SPIE Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modelling*, volume 7261, pages 72611I–1–72611I–11, 2009.

G.O. Russell. *The Vowel: Some X-ray and Photo Laryngoperiskopik Evidence*. Ohio state University Press, Columbus, Ohio, 1928.

P. Saha and J.K. Udupa. Optimum image thresholding via class uncertainty and region homogeneity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(7):689–706, 2001.

P.W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad. Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, 31(1):26–35, 1987.

J. Schroeter and M. Sondhi. Speech coding based on physiological models of speech production. *Advances in Speech Signal Processing*, pages 231–267, 1991.

C.H. Shadle, M. Mohammad, J.N. Carter, and P.J.B. Jackson. Dynamic magnetic resonance imaging: New tools for speech research. In *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS)*, pages 623–626, San Francisco, CA, 1999.

M. Sondhi. Articulatory modeling: a possible role in concatenative text-to-speech synthesis. In *Proceedings of IEEE Workshop on Speech Synthesis*, pages 73–78, Santa Monica, CA, 2002.

B.C. Sonies, T.H. Shawker, T.E. Hall, L.H. Gerber, and S.B. Leighton. Ultrasonic visualization of tongue motion during speech. *Journal of the Acoustical Society of America*, 70(3):683–686, 1981.

A. Soquet, V. Lecuit, T. Metens, B. Nazarian, and D. Demolin. Segmentation of the airway from the surrounding tissues in magnetic resonance images: a comparative study. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 3083–3086, Sydney, Australia, 1998.

A. Soquet, V. Lecuit, T. Metens, and D. Demolin. Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI. *Speech Communication*, 36(3–4):169–180, 2002.

V. Sorokin. *Speech synthesis*. Science, Moscow [in Russian], 1992.

A. Souza and J.K. Udupa. Automatic landmark selection for active shape models. In *Proceedings of SPIE Medical Imaging 2005: Image Processing*, pages 1377–1383, 2005.

M. Stone. A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data. *Journal of the Acoustical Society of America*, 87(5): 2207–2217, 1990.

M. Stone and A. Lundberg. Three-dimensional tongue surface shapes of English consonants and vowels. *Journal of the Acoustical Society of America*, 99(6):3728–3737, 1996.

M. Stone, M. Goldstein Jr, and Y. Zhang. Principal component analysis of cross sections of tongue shapes in vowel production. *Speech Communication*, 22(2–3): 173–184, 1997.

B.H. Story. Speech synthesis by mapping articulator movement patterns to a shape-based area function model of the vocal tract. *Journal of the Acoustical Society of America*, 109(5):2444–2445, 2001.

B.H. Story, I.R. Titze, and E.A. Hoffman. Vocal tract area functions from Magnetic Resonance Imaging. *Journal of the Acoustical Society of America*, 100(1):537–554, 1996.

A.M. Sulter, D.G. Miller, R.F. Wolf, H.K. Schutte, H.P. Wit, and E.L. Mooyaart. On the relation between the dimensions and resonance characteristics of the vocal tract: A study with MRI. *Magnetic Resonance Imaging*, 10(3):365–373, 1992.

J. Sundberg. On the problem of obtaining area functions from laterl x-ray pictures of the vocal tract. *Department of Speech, Music and Hearing Quarterly Progress and Status Report - KTH (KTH STL-QPSR)*, 1:43–45, 1969.

J. Sundberg, C. Johansson, H. Wilbrand, and C. Ytterbergh. From sagittal distance to area: A study of transverse vocal tract cross-sectional area. *Phonetica*, 44: 76–90, 1987.

H. Takemoto, K. Honda, S. Masaki, Y. Shimada, and I. Fujimoto. Measurement of temporal changes in vocal tract area function from 3d cine-MRI data. *Journal of the Acoustical Society of America*, 119(2):1037–1049, 2006.

H.H. Thodberg. Minimum description length shape and appearance models. In *Proceedings of Image Processing Medical Imaging (IPMI)*, pages 51–62, Ambleside, UK, 2003.

H.H. Thodberg and H. Olafsdottir. Adding curvature to minimum description length shape models. In *Proceedings of the British Machine Vision Conference (BMVC))*, pages 251–260, Norwich, UK, 2003.

M. K. Tiede and H. Yehia. A shape–based approach to vocal tract area function estimation. *Journal of the Acoustical Society of America*, 100(4):2658, 1996.

C.J. Twining, S. Marsland, and C.J. Taylor. Measuring geodesic distances on the space of bounded diffeomorphisms. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 2, pages 847–856, Cardiff, UK, 2002.

J.K. Udupa and S. Samarasekera. Fuzzy connectedness and object definition: Theory, algorithms, and applications in image segmentation. *Graphical Models and Image Processing*, 58(3):246–261, 1996.

J.K. Udupa, R.J. Goncalves, K. Iyer, S. Narendula, D. Odhner, S. Samarasekera, and S. Sharma. 3DVIEWNIX: an open, transportable software system for the visualization and analysis of multidimensional, multimodality, multiparametric images. *SPIE, Medical Imaging: Image Capture, Formatting, and Display*, 1897: 47–58, 1993.

J.K. Udupa, V.R. Leblanc, Y. Zhuge, C. Imielinska, H. Schmidt, L.M. Currie, B.E. Hirsch, and J. Woodburn. A framework for evaluating image segmentation algorithms. *Computerized Medical Imaging and Graphics*, 30(2):75–87, 2006.

M. Vaillant, M.I. Miller, L. Younes, and A. Trouv. Statistics on diffeomorphisms via tangent space representations. *NeuroImage*, 23:S161–S169, 2004.

M.J.M. Vasconcelos, S.M.R. Ventura, D.R.S. Freitas, and J.M.R.S. Tavares. Using statistical deformable models to reconstruct vocal tract shape from magnetic res-

onance images. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, 224(10):1153–1163, 2010.

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

S.M.R. Ventura, D.R.S. Freitas, and J.M.R.S. Tavares. Magnetic resonance imaging of the vocal tract: techniques and applications. In *Proceedings of International Conference on Imaging Theory and Applications (IMAGAPP)*, pages 105–110, Lisbon, Portugal, 2009.

S.M.R. Ventura, M.J.M. Vasconcelos, D.R.S. Freitas, I.M.A.P. Ramos, and J.M.R.S. Tavares. Speaker-specific articulatory assessment and measurements during Portuguese speech production based on magnetic resonance images. *Language Acquisition [Chapter 4]*, pages 117–138, 2012.

K.N. Walker, T.F. Cootes, and C.J. Taylor. Automatically building appearance models from image sequences using salient features. *Image and Vision Computing*, 20(5):435–440, 2002.

C. Westbrook, C. Roth, and J. Talbot. *MRI in Practice*. Blackwell Science, second edition, 1998.

J. Westbury, P. Milenkovic, G. Weismer, and R. Kent. X-ray microbeam speech production database. *Journal of the Acoustical Society of America*, 88:S56, 1990.

S. Wood. A radiographic analysis of constriction locations for vowels. *Journal of Phonetics*, 7(1):25–43, 1979.

A.A. Wrench and K. Richmond. Continuous speech recognition using articulatory data. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 145–148, Beijing,China, 2000.

A.A. Wrench and J.M. Scobbie. Very high frame rate ultrasound tongue imaging. In *Proceedings of 9th International Seminar on Speech Production (ISSP)*, pages 155–162, Montréal, Canada, 2011.

S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. The HTK Book. `http://htk.eng.cam.ac.uk/`, 2006, accessed April 2010.

X. Zhou, Z. Zhang, and C. Espy-Wilson. VTAR: A matlab-based computer program for vocal tract acoustic modelling. *Journal of the Acoustical Society of America*, 115(5):2543–2543, 2004.