*Aalto University*
*School of Science*
*Degree Programme in Engineering Physics and Mathematics*

Atle Kivelä

# Acoustics of the Vocal Tract
## MR image segmentation for modelling

*Master's thesis*
*Helsinki, 25 May 2015*

*Supervisor:* *Prof. Rolf Stenberg*
*Advisor:* *D.Sc. (Tech.) Jarmo Malinen*

Aalto-yliopisto
Perustieteiden
korkeakoulu

*DIPLOMITYÖN TIIVISTELMÄ*

**Tekijä:** Atle Kivelä

**Työn nimi:** Ääntöväylän akustiikka: MR kuvan segmentointi mallinnusta varten

**Päivämäärä:** 25.5.2015 **Kieli:** englanti **Sivumäärä:** 51

**Koulutusohjelma:** Teknillinen fysiikka ja matematiikka

**Vastuuopettaja:** Prof. Rolf Stenberg

**Ohjaaja:** TkT Jarmo Malinen

Tässä työssä esitetään menetelmä ihmisen ääntöväylää kuvaavien geometrioiden purkamiseksi magneettikuvauslaitteella kerätystä datasta. Lisäksi purettujen ääntöväylägeometrioiden akustiikkaa mallinnetaan laskennallisella mallilla.

Käytetty data koostuu koehenkilöstä vokaaliäännön aikana otetuista magneettikuvista. Ääntöväylän magneettikuvaus ei ole suoraviivaista, koska magneettikuvauskone ei erota luukudosta ilmasta siinä olevien vetyatomien vähäisen määrän takia. Tästä johtuen hampaat sekä ylä- ja alaleuka yhdistyvät ääntöväylän ilmapatsaaseen magneettikuvassa. Tässä työssä kehitetään menetelmä, jota voidaan käyttää näiden ylimääräisten artefaktien poistamiseen magneettikuvasta. Menetelmä perustuu erikseen ylä- ja alaleualle rakennettujen mallien kohdentamiseen otetun MRI-kuvan kanssa. Kohdentamisesta saatua informaatiota käytetään poistamaan leuoista johtuvat artefaktit. Menetelmällä pyritään minimoimaan ääntöväylän purkamiseen tarvittavaa manuaalista työtä, mikä on välttämätöntä käytössä olevan tutkimusdatan suuren määrän takia.

Työssä kuvattua metodia käytetään purkamaan ääntöväylägeometrioita suomen kielen vokaaleille. Ääntöväylägeometrioiden akustisia resonanssiominaisuuksia tutkitaan kahdella akustisella mallilla. Malleina käytetään kolmiulotteista Helmholtz-yhtälöä ja yksiulotteista Websterin yhtälöön perustuvaa resonanssimallia. Resonansseja myös verrataan magneettikuvauksen aikana samanaikaisesti nauhoitetuista ääninäytteistä purettuihin formantteihin.

**Avainsanat:** ääntöväylä, akustiikka, magneettikuvaus, resonanssi, formantti, elementtimenetelmä

Aalto University
School of Science

*ABSTRACT OF*
*MASTER'S THESIS*

**Author:** Atle Kivelä

**Title:** Acoustics of the Vocal Tract: MR image segmentation for modelling

**Date:** 25.5.2015 **Language:** English **Number of pages:** 51

**Degree programme:** Engineering Physics and Mathematics

**Supervisor:** Prof. Rolf Stenberg

**Advisor:** D.Sc. (Tech.) Jarmo Malinen

This work presents a method for extracting vocal tract (VT) geometries from MRI data. Computational models are used to model VT acoustics in the geometries obtained.

The data consists of MR images of a single test subject pronouncing Finnish vowels. MR imaging presents a challenge for VT extraction since osseous tissue, such as the maxillae and mandible, are indistinguishable from air in the image data. This limitation has been overcome by producing external models for the maxillae and mandible, and registering them with the image data. The registration information can then be used to mask the maxillae and mandible in the original data. The presented method aims to extract the VT geometry by minimizing required manual work, which is necessary due to large number of research data.

The described method is used to automatically extract geometries for Finnish vowels. The extracted geometries are used to computationally model VT acoustic resonance properties during pronounciation of vowels. For this purpose, Helmholtz and Webster resonance models are used. Computed resonances are also compared to formants that have been extracted from sound recordings carried out simultaneously with the MR imaging.

**Keywords:** vocal tract, acoustics, MRI, resonance, formant, FEM

# Contents

# Chapter 1

# Introduction

The motivation for this work stems from mathematical modelling of human speech, see Hannukainen et al. (2007); Aalto et al. (2015b). In this work, methods for automatically extracting the human *vocal tract* from magnetic resonance (MR) images are presented. The obtained anatomical geometries are then used in a computational study of the acoustic properties of the vocal tract.

## 1.1 Human speech production

A schematic diagram of the human vocal mechanism is shown in Figure 1.1. The lungs produce an airflow that passes through the trachea and the vocal tract (VT). The air will exit the VT through the mouth opening or the *nasal tract*, both of which can independently be open or closed depending on the situation. The nasal tract begins at the velum and terminates at the nostrils. Airflow through the nasal tract is regulated by the velum opening.

The trachea is separated from the VT by the larynx which is a narrowing containing the *vocal folds*. The orifice of the vocal folds is called the *glottis*. In ordinary phonation, the accelerating airflow through the glottis – according to Bernoulli's principle – causes a pressure drop, which then exerts an aerodynamic force that pulls the vocal folds together and closes the glottis. The resulting compression of the vocal fold tissues and the pressure difference between the trachea and the VT open the glottis again. Thus, the vocal folds vibrate, acting like a valve that periodically opens and closes the glottis.

The oscillation may start after the airflow exceeds a limit known as the phonation threshold. The periodic opening and closing of the glottis produces flow pulses that pass through the vocal tract. These flow pulses can be considered as an

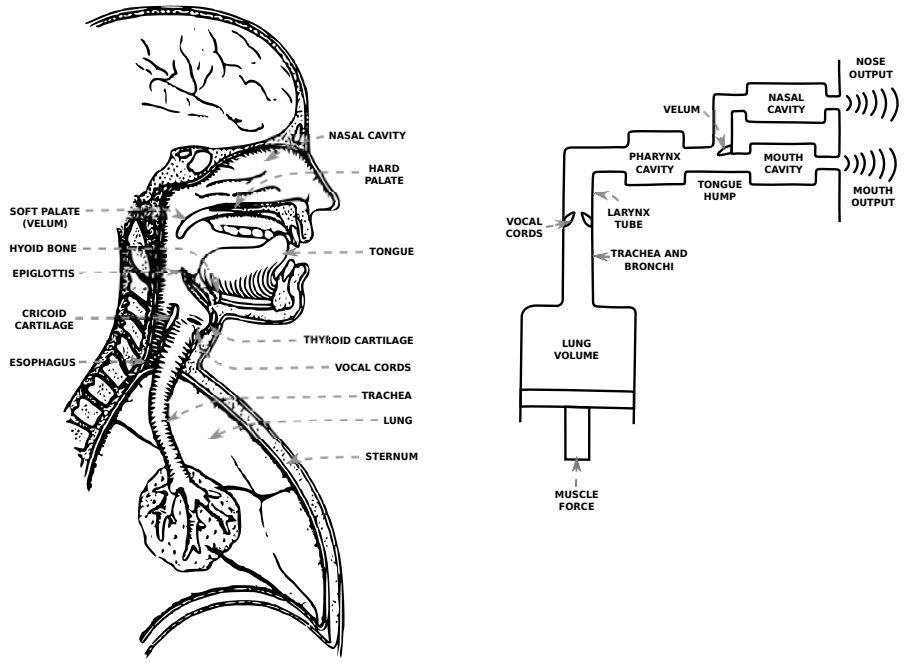



**Figure 1.1:** Schematic diagram of the human voice production mechanism (Flanagan, 1972).

acoustic sound signal that is filtered by the vocal tract before it is transmitted to the exterior space through the mouth and/or nostrils.

The geometry, i.e., the shape, of the VT varies during phonation, allowing different kinds of speech sounds to be produced. In sustained vowel production, the vocal tract geometry remains more or less static. The VT is usually open at every point and the sound is mainly transmitted through the mouth.

Different vowels can be classified by the *acoustic resonance frequencies* of the corresponding VT geometry. In phonetics, these resonance frequencies are often referred to as *formant frequencies*; the ambiguities between different but related definitions are discussed later in this work. Other speech sounds, such as fricatives and consonants, involve temporarily constricting or closing the VT during phonation, in which case the primary sound source is not the glottis. This complicates modelling phonation significantly since the geometry is inherently time dependent. (Flanagan, 1972)

Only static VT geometries corresponding to Finnish vowels are examined in this work.

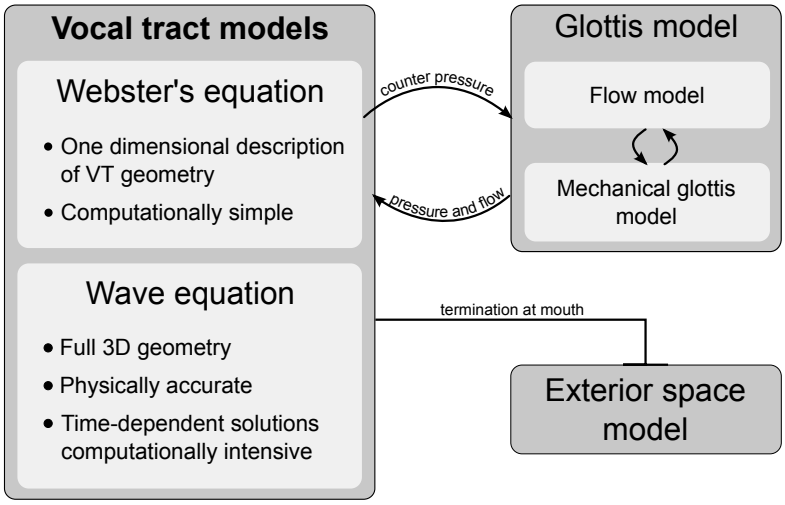



**Figure 1.2:** Schematic diagram of a model for vowel production.

## 1.2 Modelling speech production

Speech production models are used in telephony, speech synthesis, and increasingly more often in medical applications (see Svancara and Horacek, 2006; Flanagan, 1972). Increased computational capacity has made it possible to model the physics of speech production more accurately. Speech synthesis from true VT geometries is becoming a possibility. The precise relation between the vocal tract geometry and the produced speech sound is especially relevant in medical applications. In speech synthesis, only the perceived quality of the produced signal matters and simplified acoustic models are adequate. However, the human speech production system as a whole is very complicated. Simplifications will have to be made even if most parts of the system could be accurately modelled. For example, the neuronal control, and thus neural diseases, are known to be closely related to speech production, but the mechanisms are not yet very well understood (Dromey et al., 1995; Skodda et al., 2011).

Most speech production models are based on the Source-Filter theory that has been attributed to Fant (1971), even though some of the ideas can be found in earlier work by von Helmholtz (1912). In the Source-Filter theory, the source produces a signal that is then filtered. The source is usually assumed to be independent of the filter, and the filter is assumed to be independent of the termination. When modelling vowel production, the glottis, the VT, and the mouth are the source, the filter, and the termination, respectively. If the filter is assumed to be linear and time invariant, inverse filtering techniques can be used to estimate the source signal from a measured output (see, e.g Fant, 1971; Alku et al., 2006).

Figure 1.2 shows a schematic diagram of the speech model treated in this work. A mechanical model for producing the glottis signal has been presented by Aalto (2009) and refined by Murtola (2014). The glottis model uses Webster's-resonator-based VT model and a subglottal model to produce a counterpressure for the glottis model. However, Webster's resonator is not able to capture, e.g., cross-modes at frequencies above 4 kHz, which can be done using a full 3D acoustic model of the VT.

All speech models take the VT configuration into account in some way. The physically accurate models used in this work depend on a high quality representation of the VT geometry in 3D. The main part of this work focuses on constructing reliable methods for extracting VT geometries of different vowel configurations from medical images of variable quality.

In addition to the speech production mechanisms, the exterior space plays a significant role in forming the final acoustic signal. At least for time-dependent models, constructing a physically fully accurate exterior space model is computationally unattainable. However, a separate, simplified exterior space model can be used (see Vampola et al., 2013; Hannukainen et al., 2014). Modelling the exterior space is outside the scope of this work.

## 1.3 Medical imaging

Since the introduction of the X-ray machine by Wilhelm Röntgen in 1895, medical imaging has grown to be a routine part of medicine. Medical imaging technology has improved considerably since then, and nowadays there are various imaging methods such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Ultrasound. Different imaging methods rely on different physical phenomena, and they can also be used to capture non-anatomical information. The data for this work has been obtained using MRI. Compared to CT scanning, the MRI has the additional benefit of not using ionizing radiation, which would — due to research-ethical reasons related to test subject health — limit the amount of images that can be taken.

Raw medical images are usually analyzed by trained medical professionals. Image processing and analysis techniques can, however, be applied to aid the analysis. Computational models of organs can also be used together with image analysis software to give additional information to a user.

Medical image analysis and medical image computing combine many fields, including computer vision, data science, physics, and mathematics. As presented in

Angenent et al. (2006), image analysis usually tackles four key problems:

**Segmentation** - automated methods to locate relevant anatomical objects from images,

**Registration** - methods to align multiple datasets that usually present 2D or 3D images,

**Visualization** - technology to give a user the possibility to control the process and inspect the results,

**Simulation** - models that can be used in planning procedures and simulation of treatment results.

In general, all medical image processing applications require custom procedures to cope with specific situations. In this work, all of the problems above have been considered, maintaining focus on the computational modelling of human speech. Specifically, we have searched for methods to cope with the dental structures missing from the MR image due to the poor contrast between osseous tissue and air in MRI. The missing information can be added by registering external dental models with the VT models that have been extracted from the MRI data.

Extracted geometric models can be used in computational simulations of the acoustics and fluid mechanics of the VT anatomy. Figure 1.3 illustrates pressure distribution on the surface of the VT – for a single resonance frequency – during sustained pronunciation of a single vowel. A 3D print of the VT is also shown in Figure 1.3.

**Figure 1.3:** A surface mode of a pressure distribution for the 5th resonance of vowel [e] and a 3D print of the geometry for vowel [a].

## 1.4 Outline of this work

This work consists of three main parts. First, physical models for VT acoustics are presented in Chapter 2. These models are part of the full phonation model discussed in Section 1.2. The acoustic wave equation is initially derived from first principles, and the Helmholtz equation and the generalized Webster's equation, as shown in Lukkari and Malinen (2013), are discussed. Both of these equations can be derived from the wave equation. The Helmholtz equation is the time-independent version of the wave equation whereas Webster's equation is a simplification that approximates the 3D wave equation in a tubular domain. These models are later used in Chapter 5 to computationally model VT acoustics.

Second, an introduction to the methods and algorithms used for feature extraction and registration is given in Chapter 3. These tools are then used in Chapter 4 to build an algorithm for extracting a 3D geometric representation of the VT from MR images. This VT geometry is required for computations that use the models presented in Chapter 2. The goal has been to design algorithms and software that can be used to extract VT geometries from MR images without extensive manual work. This is important since statistically relevant studies in speech medicine will require thousands of images to be processed. MRI does not distinguish osseous tissue from air, and thus the main part of the algorithm makes use of 3D image registration techniques to add missing information to the anatomical VT model.

Third, the presented mathematical models and the extracted geometries are used in Chapter 5 to investigate the acoustics of the VT. Resonances of the VT geometry for Finnish vowels are obtained computationally using the Finite Element Method (FEM) for solving the models presented in Chapter 2. The computed resonances are compared with formant frequencies that have been extracted from sound samples recorded simultaneously to MR imaging. All data for the experiments has been obtained as described in Aalto et al. (2011a).

In addition to validating the extracted geometries by comparing the computationally obtained results to the recorded sound samples, we are interested in how the Helmholtz model compares to the time-independent Webster's horn model in different VT geometries. These two models, although mathematically rather simple, already provide interesting insights into VT acoustics. The obtained results could help understand the effects of oral and maxillofacial surgeries to speech.

# Chapter 2

# Differential equations for acoustic modelling

Next, we consider the mathematical background of the models used in the experimental part of this work. The derivation of the 3D acoustic wave equation will be outlined. Simplified versions of the full wave equation, i.e., the *Helmholtz equation* and the time-harmonic *Webster's equation*, will also be introduced. The Helmholtz equation is obtained by assuming the solution of the wave equation is time-harmonic. The derivation of the generalized Webster's equation is more involved (Lukkari and Malinen, 2013; Aalto et al., 2015a; Lukkari and Malinen, 2015). Webster's equation is an essential part of the glottis model shown in Fig. 1.2.

## 2.1 Wave equation and Helmholtz equation

The wave equation is a second-order hyperbolic partial differential equation that governs the propagation of waves in fluid. The fluid can be considered as a continuous medium specified by the mass density $\rho(\mathbf{x}, t)$, the velocity $\mathbf{v}(\mathbf{x}, t)$, and the pressure $p(\mathbf{x}, t)$. To derive the wave equation, we first present fundamental fluid mechanical equations and then impose further restrictions to capture the compressional characteristic of waves in fluids.

### Fundamental equations in fluid mechanics

**Euler's equation**: Taking into account the forces caused by the pressure $p$ and an external force $\mathbf{f}_{ext}$ – which can be zero –, one can arrive at *Newton's second law*

*for an ideal fluid*, known as *Euler's equation*

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\frac{1}{\rho}\nabla p + \mathbf{f}_{ext}, \tag{2.1}$$

**Conservation of mass:** Assuming *mass conservation* one can obtain the *continuity equation*

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \tag{2.2}$$

which is a very general conservation law seen in many other applications as well.

**Conservation of energy:** If the underlying thermodynamic process is assumed locally *isentropic*, i.e., reversible, then one can relate the integral energy $\epsilon$ of a small enough volume element to the work done to the element using the following definition

$$\epsilon(s, \rho) = \int_{\rho} \frac{p(s, \rho')}{\rho'^2} \, \mathrm{d}\rho', \tag{2.3}$$

where $s$ is the *entropy* per unit mass. Here the volume elements are assumed to be small enough for $p, s,\ \rho$, and $\epsilon$ to be independent of position, i.e., there is a local thermodynamic equilibrium. Differentiating Eq. (2.3) we obtain the *isentropic relation*

$$\frac{\partial \epsilon}{\partial t} = \left(\frac{\partial \epsilon}{\partial \rho}\right)_s \frac{\partial \rho}{\partial t} = \frac{p}{\rho^2}\frac{\partial \rho}{\partial t}. \tag{2.4}$$

In addition, we can define the *enthalpy* per unit mass as $\epsilon + \frac{p}{\rho}$ which has the property

$$\nabla\left(\epsilon + \frac{p}{\rho}\right) = \frac{1}{\rho}\nabla p \tag{2.5}$$

**Bernoulli's theorem:** Combining the vector identity

$$(\mathbf{v} \cdot \nabla)\mathbf{v} = \nabla(\frac{1}{2}|\mathbf{v}|^2) - \mathbf{v} \times (\nabla \times \mathbf{v})$$

with Euler's equation (2.1) and the property (2.5) results in *Bernoullis's theorem for isentropic irrotational flow*

$$\epsilon + \frac{p}{\rho} + U + \frac{1}{2}v^2 - \frac{\partial \Phi}{\partial t} = 0, \tag{2.6}$$

where we have used the fact that for an irrotational flow $\nabla \times \mathbf{v} = 0$.

## Wave equation

To derive the acoustic wave equation, we consider a stationary fluid at constant density $\rho_0$ and pressure $p_0$. We then define the acoustic waves in the fluid as small perturbations in density and pressure. This gives

$$\rho = \rho_0 + \rho' \quad \mathbf{v} = \mathbf{v}_0 + \mathbf{v}' \quad p = p_0 + p', \tag{2.7}$$

where $\rho'$, $\mathbf{v}'$, and $p'$ present small first-order perturbations in the corresponding quantities, and $\rho_0$, $p_0$, and $\mathbf{v}_0$ are constants. During the derivation, the fluid is considered to be at rest and thus $\mathbf{v}_0 = 0$ and $\mathbf{v} = \mathbf{v}'$. We continue by linearizing Euler's equation Eq.(2.1) respect to the perturbations to obtain

$$\frac{\partial \mathbf{v}'}{\partial t} = -\frac{1}{\rho_0}\nabla p'. \tag{2.8}$$

Doing the same for the continuity equation Eq. (2.2) gives

$$\frac{1}{\rho_0}\frac{\partial \rho'}{\partial t} + \nabla \cdot \mathbf{v}' = 0. \tag{2.9}$$

We further consider the process to be isentropic and thus reversible. There exists an isentropic equation of state for pressure $p = p(s, \rho)$ where $s$ is entropy and $\rho$ is density. Linearization of this equation of state yields

$$p' = p(s, \rho_0 + \rho') - p(s, \rho_0) \approx \left(\frac{\partial p}{\partial \rho}\right)_s \rho'. \tag{2.10}$$

Now defining the constant $c > 0$ as $c^2 = \left(\frac{\partial p}{\partial \rho}\right)_s$ we obtain the identity

$$p' = c^2 \rho'. \tag{2.11}$$

Next we assume that the flow is irrotational ($\nabla \times \mathbf{v} = 0$) and it has a velocity potential $\Phi(\mathbf{x}, t)$ with

$$\mathbf{v}' = -\nabla \Phi. \tag{2.12}$$

After some manipulation, the first-order expansion of the isentropic Bernoulli's equation (2.6) and the relation $\frac{\partial \epsilon}{\partial \rho} = \frac{p}{\rho^2}$ from (2.4) yields

$$p' = \rho_0 \frac{\partial \Phi}{\partial t} \quad \Leftrightarrow \quad c^2 \rho' = \rho_0 \frac{\partial \Phi}{\partial t} \tag{2.13}$$

When Eq. (2.12) and the time derivative of Eq. (2.13) is plugged to the linearized continuity equation (2.9), we obtain the wave equation for the velocity potential

$$\nabla^2 \Phi = \frac{1}{c^2} \frac{\partial^2 \Phi}{\partial t^2} \tag{2.14}$$

From Eq. (2.13) we also see that there is a direct relationship between the velocity potential $\Phi$ and the pressure $p$, thus solving the velocity potential also gives us the acoustic pressure distribution inside the volume.

### The Helmholtz equation

If the velocity potential in equation (2.14) is assumed to be time-harmonic, i.e., $\Phi_\lambda(\mathbf{r}, t) = \operatorname{Re} \Phi_\lambda(\mathbf{r}) e^{-i\lambda t}$, then we obtain the Helmholtz equation

$$\nabla^2 \Phi_\lambda(\mathbf{r}) + \frac{\lambda^2}{c^2} \Phi_\lambda(\mathbf{r}) = 0. \tag{2.15}$$

This is an eigenvalue problem that is mathematically considerably easier than the wave equation. Solutions of the Helmholtz equation correspond to the resonance frequencies of the full wave equation.

## 2.2 Webster's equation

Webster's equation is a simplification of the wave equation (2.14). It models waves in a tubular domain $\Omega \subset \mathbb{R}^3$ and assumes that the waves move mainly in the longitudinal direction of the tube. This is done by considering only the averages of the solution in tube cross-sections. A complete derivation of the generalized Webster's equation can be found from Lukkari and Malinen (2013); Aalto et al. (2015a); Lukkari and Malinen (2015). Only an overview of the equation will be presented here.

The tubular domain $\Omega$ is assumed to have a centerline presented by a smooth curve $\gamma(s)$ parametrized by arc length. The curvature of the centerline at a point $\gamma(s)$ is defined to be $\kappa(s) = \|\gamma''(s)\|$. We can fix an orthonormal coordinate system, known as the *Frenet frame*, to every point of $\gamma$. The unit vectors of the coordinate system are defined by

$$\mathbf{t}(s) = \gamma'(s), \quad \mathbf{n}(s) = \frac{\mathbf{t}'(s)}{\kappa(s)} \quad \text{and} \quad \mathbf{b}(s) = \mathbf{t}(s) \times \mathbf{n}(s).$$

These vectors are called the tangent vector, the normal vector and the binormal vector, respectively. In the derivation of Webster's equation it is assumed that $\kappa(s) > 0$, which results in a right-handed coordinate system at all points of the curve.

The domain $\Omega$ can be described using cross-sections $\Gamma(s)$. The cross section $\Gamma(s)$, at a point $\gamma(s)$, is perpendicular to $\mathbf{t}(s)$, i.e., the tangent is a normal for the cross-section. The boundary of $\Omega$ consists of the ends of the tube $\Gamma(0)$ and $\Gamma(L)$ and the tube wall $\Gamma = \cup_{s\in[0,L]}\partial\Gamma(s)$. In computations and in the derivation each $\Gamma(s)$ is assumed to be a circular disk, the corresponding hydrodynamic radius of which is denoted by $R(s)$.

If $\Phi$ is assumed to be the solution of (2.14), the average solution can be written as

$$\overline{\Phi}(s,t) = \frac{1}{A(s)}\int_{\Gamma(s)} \Phi \,\mathrm{d}A, \tag{2.16}$$

where $A(s) = \pi R(s)^2$. After a lengthy derivation in Lukkari and Malinen (2013), Webster's equation can be written as

$$\frac{1}{c^2\Sigma(s)^2}\frac{\partial^2\Phi}{\partial t^2} + \frac{2\pi\alpha W(s)}{A(s)}\frac{\partial\Phi}{\partial t} - \frac{1}{A(s)}\frac{\partial}{\partial s}\left(A(s)\frac{\partial\Phi}{\partial s}\right) = 0 \tag{2.17}$$

where $\alpha$ is the dissipation coefficient at the VT walls, and the *sound speed correction factor*, $\Sigma(s)$, and the stretching factor, $W(s)$, are defined as

$$\Sigma(s) = \left(1 + \frac{1}{4}\eta^2(s)\right)^{-1/2}, \text{ with } \eta(s) = R(s)\kappa(s), \text{ and} \tag{2.18}$$

$$W(s) = R(s)\sqrt{R'(s)^2 + (\kappa(s)-1)^2}. \tag{2.19}$$

.

As with the Helmholtz equation, making the time-harmonic assumption $\Phi_\lambda(s,t) = \mathrm{Re}\,\Phi_\lambda(\mathbf{r})e^{-i\lambda t}$ yields the *time-harmonic Webster's equation*

$$\frac{\lambda^2}{c^2}\frac{1}{\Sigma(s)^2}\Phi_\lambda + \lambda\frac{2\pi\alpha W(s)}{A(s)}\Phi_\lambda - \frac{1}{A(s)}\frac{\partial}{\partial s}\left(A(s)\frac{\partial\Phi_\lambda}{\partial s}\right) = 0. \tag{2.20}$$

The main difference between Eq. (2.15) and Eq. (2.20) is that Webster's equation only captures resonances in the longitudinal direction. If the cross-sectional area of the tube is large enough, there can be mixed resonances, i.e., resonances having transversal components. The Helmholtz equation will also capture the transversal resonances.

## 2.3 Boundary conditions

All of the models in Sections 2.1 and 2.2 require boundary conditions to attain a unique solution. In this work, we will only be using the Helmholtz equation and the time-harmonic version of Webster's equation. Thus, none of the problems will be time-dependent.

We define $\Omega$ to be the VT air volume. The boundary $\partial\Omega = \Gamma_g \cup \Gamma_w \cup \Gamma_m$ consists of the *glottis boundary* $\Gamma_g$, the *VT wall boundary* $\Gamma_w$, and the *mouth boundary* $\Gamma_m$.

First, at the glottis boundary we use the radiating *Robin boundary condition* of the form

$$\lambda\Phi + c\frac{\partial\Phi}{\partial\nu} = 0, \quad \text{at } \Gamma_g. \tag{2.21}$$

This condition presents the signal coming from the glottis, and its time variant form is $\frac{\partial\Phi}{\partial t} + c\frac{\partial\Phi}{\partial\nu}$. Second, at the VT walls we have the boundary condition

$$\alpha\lambda\Phi + \frac{\partial\Phi}{\partial\nu} = 0, \quad \text{at } \Gamma_w. \tag{2.22}$$

This boundary condition accounts for the energy dissipation to the tissue wall. The boundary condition is used in the derivation of Webster's equation (Eq. 2.20) and it is not explicitly defined when numerically solving it.

Finally, at the mouth boundary we use a *Dirichlet boundary condition*

$$\Phi = 0, \quad \text{at } \Gamma_m. \tag{2.23}$$

This boundary condition regards the mouth boundary as an idealised open end of an acoustic tube. As such, it is a very radical simplification of the exterior space acoustics, the efficient modelling of which is a very complicated subject.

# Chapter 3

# Segmentation and registration methods

As mentioned in Chapter 1, image analysis tries to tackle two basic problems, namely *segmentation* and *registration*. Segmentation usually means dividing the dataset (e.g., the image) to multiple segments and/or using *feature extraction* methods to identify relevant parts of the dataset. Registration methods find correspondences between two datasets, which might be obtained using different methods and/or at different times. The registration information can be used, e.g., to aid segmentation or to combine multiple datasets. This chapter gives an introduction to the feature extraction and registration methods used in this work.

## 3.1 Representation of 3D data

Algorithms presented in this chapter treat three different presentations of 3D data. These three presentations are *the voxel image*, *point cloud*, and *a triangular mesh*. Naturally, conversions between these presentations can be carried out, but they do not necessarily preserve all information.

In general, the image data can be thought of as a mapping $I : D \to C$, where $I$ maps a given point $\mathbf{x} \in D$ to a color $I(\mathbf{x}) \in C$. The color space can consist of one or multiple color values, or other data. In this chapter – unless otherwise mentioned – all image data is considered to be 3D with one-dimensional color space, i.e., grayscale. In image analysis, the spaces $D$ and $C$ and the mapping $I$ are usually considered to be continuous. However, the most common way to present images in computer memory is an array of voxels (pixels in 2D) consisting of a regular 3D lattice of points with their corresponding color values. This presentation is called a *voxel image*.

Whereas a voxel image represents the whole space, individual objects or features in the space can be presented using a 3D *triangle mesh* consisting of a set of triangles in 3D space. Most applications impose additional requirements for the mesh. For example, tetrahedral mesh generation requires the mesh to be interconnected, closed, and a manifold. The last condition means that the neighborhood of every point resembles an euclidean space, i.e., the surface does not fold through itself. A triangular mesh consists of a set of vertices and a triangulation defined on them.

*Point clouds* are 3D sets of points, each of which usually have additional information, e.g., a color value, associated with them. Thus a voxel image can also be presented as a point cloud. However, there is a semantic distinction as a voxel image aims to represent every point in the whole space whereas the point cloud usually consist of – not necessarily regularly spaced – points that represent important features or measurements.

It is possible to turn a point cloud representing a surface to a triangle mesh by approximating surface normals and triangulating the points. A triangle mesh can trivially be converted into a point cloud by ignoring the triangulation information, which constraints what points can be considered neighbors. Point clouds and triangle meshes can be converted into voxel images by choosing the values of the voxels depending on how the mesh or point cloud would intersect them. It is possible to lose information in some of these transformations.

In the context of this work, medical imaging devices are the most typical source of voxel images. Triangulations and point clouds are usually obtained by using physical measurement devices or by extracting them from (voxel) image data. Applications that use 3D scanning techniques usually rely heavily on point cloud processing.

## Data structures and basic problems

Practical implementations of image processing and feature extraction algorithms rely on an ability to efficiently obtain basic information about the underlying data. Thus it is important to choose data structures that support the implementation of efficient algorithms.

A voxel image of dimensions $m \times n \times k$ has $mnk$ elements and is usually stored in computer memory as a continuous array. This representation is a basic data structure in many programming languages, and it is widely used, e.g., for dense matrices. It is possible to access and modify voxel values in constant time. Modification of the dimensions, i.e., adding or removing voxels, might be expensive but

it is rarely required in applications.

The presentation of point clouds needs careful consideration. As the points are not regularly spaced, the data structure needs to support efficient search of points according to their spatial coordinates. Most point cloud algorithms also rely heavily on finding *nearest neighbors* of points. The $k$ nearest neighbors of a point are called the $k$-neighborhood of the point. A visualization of the $k$-neighborhood can be seen in Figure 3.3a.

Point clouds are usually stored in *k-d trees* or *octrees*. A k-d tree is a binary tree structure where every node stores a $k$-dimensional point. Each non-leaf node defines a hyperplane that partitions the space into two subspaces respect to some of the $k$ dimensions. An octree is a tree where each node has exactly eight children. The octree stores the points by recursively subdividing the 3D space into eight octants, corresponding to the eight children. Both of these data structures can be used to search for points and nearest neighbors in $O(\log n)$ time. As mentioned before, triangle meshes are essentially point clouds with associated neighbouring information.

## 3.2 Algorithms for feature extraction

Feature extraction aims to detect and/or isolate predetermined shapes or objects from image data. For example, different forms of edge detection, isosurface extraction, and template matching can all be considered to be forms of feature extraction. This section gives a general overview of the *Marching Cubes algorithm*, which is a well known method for isosurface extraction by Lorensen and Cline (1987). We also present *parametric deformable models*, which are also used to isolate regions of an image.

### The Marching Cubes algorithm

The Marching Cubes algorithm is used to extract *isosurfaces* from image data. An isosurface is a surface inside the image corresponding to a pre-defined color of the image. For example, using the notation adopted earlier, the isosurface of a gray level $l$ in a grayscale image corresponds to the set $S = \{\mathbf{x} \in D | I(\mathbf{x}) = l\}$.

When $I$ is a continuous function, the resulting surface will be closed. Naturally, this is not actually true for the image data in computer memory. The Marching Cubes algorithm constructs a polygonal surface from a scalar field that has been sampled on a regular 2D or 3D grid. The algorithm was first proposed by Lorensen and Cline

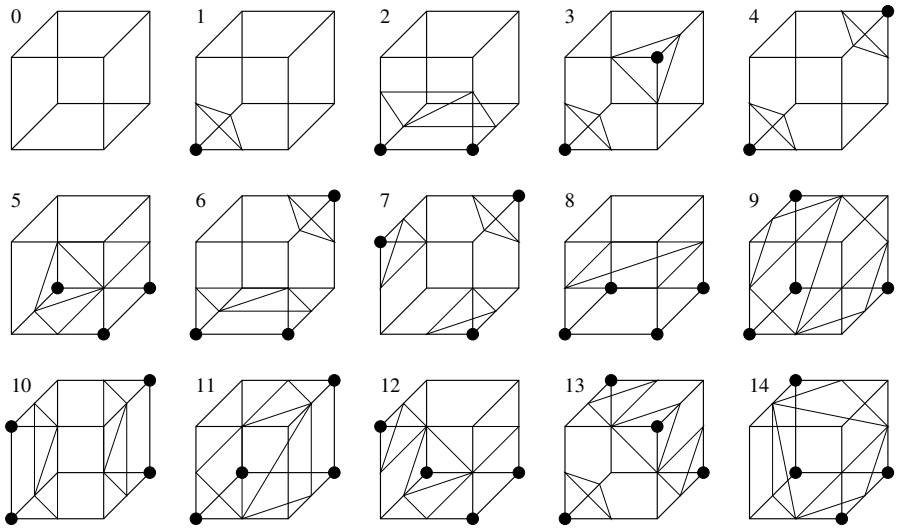



**Figure 3.1:** Cases for marching cubes (Antiga, 2002).

(1987). Many improvements and variations of the algorithm have been proposed, but new algorithms are still being compared against the original Marching Cubes algorithm.

The basic idea of the algorithm is to approximate the scalar field as a continuous function by linearly interpolating between neighboring vertices. In order to do this efficiently, it is noted that, inside an imaginary cube formed by 8 neighboring points in the scalar grid, there is only a limited number of possible triangulations which can be tabulated beforehand.

If the 8 grid points are labeled as being *above* or *below* the threshold value the cube vertices have $2^8 = 256$ different combinations, the number of which can be reduced to 15 when rotational and reflective symmetries are taken into account. Figure 3.1 shows a predetermined way to triangulate the volume inside the cube for each of these configurations. Some of the cases have complementary cases as there are many ways to triangulate the volume. These cases are shown in Figure 3.2. The complementary cases have to be correctly selected, as a wrong choice can change surface topology, resulting in holes. The exact location of the triangle vertices at the cube edge is determined by linearly interpolating between the vertex scalar values.

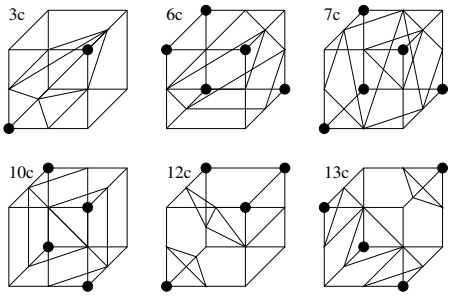



**Figure 3.2:** Complementary cases for marching cubes (Antiga, 2002).

## Parametric deformable models

One problem with isosurface extraction is that it is not really based on *image features* as it just simply extracts a surface corresponding to a given color value. To take image features in account we can use *parametric deformable models* to extract features from the image.

Parametric deformable models are models, e.g., curves or surfaces, that evolve in the image space. The deformations of the models are described from the Lagrangian point of view. Compared to the isosurface extraction this approach copes better with noise, artefacts, and low contrast in the original images. All of the parametric deformable models used in this work have been implemented in the VMTK library. Thus only a general overview will be given here.

### Snakes

A *snake* is a closed curve used in 2D shape detection. The curve can be described in the Lagrangian frame as a time dependent parametrised function

$$C : \underbrace{U}_{s} \times \underbrace{\mathbb{R}^+}_{t} \longrightarrow \mathbb{R}^2.$$

This curve then evolves in time so that it eventually minimizes the energy functional

$$E_{snake}(C) = E_{smooth}(C) + E_{image}(C), \tag{3.1}$$

where

$$E_{smooth}(C) = \int_0^1 w_1 |C_s|^2 + w_2 |C_{ss}|^2 \, \mathrm{d}s$$

and

$$E_{image}(C) = \int_0^1 w_3 P(s) \, \mathrm{d}s,$$

which is the "driving energy" that takes image features into account using the scalar potential $P$ (e.g. $P(\mathbf{x}) = -|\nabla I(\mathbf{x})|$).

The energy functional (3.1) can be minimized using gradient descent minimization, i.e.,

$$C_{t+1} = C_t - \gamma \nabla E_{snake}(C).$$

In practice, the curve $C$ is presented as a set of discrete points $p_i$ and its energy functional can be written in terms of the points as

$$E^*_{snake} \approx \sum E_{snake}(p_i),$$

after which the gradient descent iteration for the discrete points $p_i$ describing the curve $C$ becomes

$$p_{i,t+1} = p_{i,t} + \gamma \nabla E^*_{snake}(p_{i,t}).$$

This is the simplest and most straightforward way to implement the algorithm. The derivatives can be approximated using finite differences. The initial curve can be put inside or outside the feature to be extracted, and the parameters will be chosen so that the curve either expands or shrinks depending on the position of the initial curve.

In theory, it is possible to do 3D feature extraction by combining the results of multiple extracted 2D features. However, with complicated shapes such as the VT, problems are to be expected. For example, if the alignment of the 2D slice of the 3D image is not consistent with the medial axis of the object to be extracted (which may not be known), it might be difficult to combine the extracted features reliably to form a 3D image. This is further discussed later in Section 4.3.

### Balloons

A *balloon* is a closed 3D surface that is used similarly to *snakes*. Moving from a 2D model to a 3D one complicates the model somewhat. For example, if a surface is shrunk according to its curvature, it doesn't necessarily evolve into a sphere before collapsing to one point, whereas a 2D curve always evolves into a circle when shrinking according to its curvature. For this reason, balloons are usually only inflated instead of shrinking them.

The parametric deformable surface is defined by a function

$$S : \underbrace{U}_{r} \times \underbrace{U}_{s} \times \underbrace{\mathbb{R}^+}_{t} \longrightarrow \mathbb{R}^3.$$

It is also assumed that for the partial derivatives $|S_r| = |S_s| = 1$ and $S_r \cdot S_s = 0$. The energy functional for this surface is

$$E_{balloon}(S) = E_{infl}(S) + E_{smooth}(S) + E_{image}(S). \tag{3.2}$$

The inflation term $E_{infl}$ is a function describing the inflation energy of the balloon, and it can be constant or dependent on the surface features.

In practice, problems arise when two snakes or balloons collide and force a re-parametrization. For this reason, these algorithms are actually implemented using *scalar field level set* to avoid re-parametrizations. The algorithm describes the surface as an isosurface of a scalar field that is modified during the iteration of the algorithm. The downside of using scalar field description of the curve or surface is that the complexity of the algorithm will be dependent on the size of the image. More extensive description of scalar field level sets can be found in Antiga (2002).

## 3.3 Algorithms for registration

Registration aims to find a transformation that aligns two separate datasets. Here the data is considered to be a *point cloud* consisting of a set of points in 3D space. In addition to their coordinates, the points can have other data associated with them. Since point clouds tend to be fairly large and complicated data structures – thus making a brute force search impossible – their registration is usually done by estimating features that can aid the alignment. Additional information, such as the neighboring information from a polygonal mesh, can also be incorporated. All the algorithms presented here are implemented in the Point Cloud Library (PCL) (Rusu and Cousins, 2011).

First, the surface normals of the point clouds are estimated, and the normal information is then used to compute point features that are independent of the *6D orientation* ($x$, $y$, $z$, roll, yaw, pitch) of the point cloud. These orientation independent descriptors are then compared in order to find matching points in the two point clouds. When enough matching points are known, an affine transformation can be estimated and the quality of the alignment can be measured, e.g., in terms of squared euclidean distance of closest points in the two clouds.

## Surface Normal Estimation

One of the simplest ways to estimate surface normals is to try to find the normal of the plane tangent to the surface. Finding the tangent surface is a least-square plane fitting estimation problem. The problem can be reduced to doing Principal Component Analysis (PCA) for a covariance matrix created from the nearest neighbors of the point where the normal is to be estimated. For a set of points $\mathbf{p}_i$ in the neighborhood – including the point itself – the covariance matrix is assembled as

$$\mathbf{C} = \frac{1}{k}\sum_{i=1}^{k}(\mathbf{p}_i - \overline{\mathbf{p}}) \cdot (\mathbf{p}_i - \overline{\mathbf{p}})^T, \tag{3.3}$$

where $\overline{\mathbf{p}}$ is the centroid of the points $\mathbf{p}_i$ and $k$ is the number of points. In practice the PCA of $\mathbf{C}$ finds the directions in which the set of points $\mathbf{p}_i$ has the greatest variance, and thus two linearly independent tangent vectors of the plane are most likely to point to the directions of the eigenvectors corresponding to the two largest eigenvalues of $\mathbf{C}$.

## Point Feature Histograms

The Point Feature Histogram (PFH) tries to capture the $k$-neighborhood mean curvature of a point using a multidimensional histogram value that is independent of the 6D orientation of the data. Estimations of surface normals and surface curvature give a local description of the data, but they are usually too general, i.e., there are many points with the same normals and curvature, to effectively align two point clouds. Point Feature Histograms were developed to mitigate this problem.

Figure 3.3a shows the $k$-neighborhood of a point. The PFH is based on the relationship of these points and their estimated surface normals, which is why the quality of the PFH presentation is also dependent on the quality of the estimated normals. PFH estimates the relative difference of a pair of points $\mathbf{p}_t$ and $\mathbf{p}_s$ and their associated normals $\mathbf{n}_t$ and $\mathbf{n}_s$ by defining a coordinate frame (see Fig. 3.3b)

$$\begin{cases} \mathbf{u} = \mathbf{n}_s \\ \mathbf{v} = \mathbf{u} \times \frac{(\mathbf{p}_t - \mathbf{p}_s)}{\|\mathbf{p}_t - \mathbf{p}_s\|_2} \\ \mathbf{w} = \mathbf{u} \times \mathbf{v} \end{cases} .$$

Using this **uvw**-frame the difference between the normals $\mathbf{p}_t$ and $\mathbf{p}_s$ can be ex-

pressed using angular features as

$$\begin{cases} \alpha = \arccos(\mathbf{v} \cdot \mathbf{n}_t) \\ \phi = \arccos(\mathbf{u} \cdot \frac{(\mathbf{p}_t - \mathbf{p}_s)}{d}) \\ \theta = \arctan(\mathbf{w} \cdot \mathbf{n}_t, \mathbf{u} \cdot \mathbf{n}_t) \end{cases},$$

where $d$ is the Euclidean distance between the points $\mathbf{p}_t$ and $\mathbf{p}_s$. The quadruplet $\langle \alpha, \phi, \theta, d \rangle$ is computed for each pair of two points in the $k$-neighborhood. These quadruplets are then binned into a histogram to form the PFH of a given point.

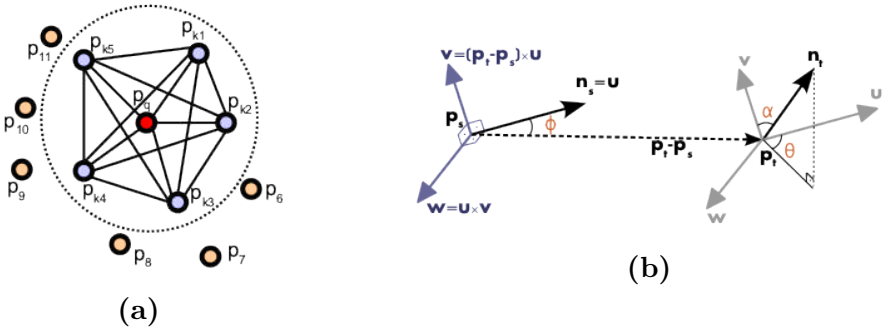



**Figure 3.3:** (a) $k$-neighborhood and (b) PFH frame (Rusu, 2010).

The computational complexity of the full PFH algorithm is $\mathcal{O}(nk^2)$ for a set of $n$ points with $k$ neighbors each. To reduce the complexity, a variant known as the Fast Point Feature Histogram (FPFH) has been developed by Rusu (2010).

FPFH avoids estimating all the possible combinations inside the neighborhood by only computing the estimators for the query point's neighbours (see Fig. 3.3a and Fig. 3.4). This estimator is called Simplified Point Feature Histogram (SPFH). The full FPFH descriptor is then calculated as a weighted combination of the SPFH descriptors in the $k$-neighborhood. Figure 3.4 shows a visualization of the FPFH neighborhood. The FPFH descriptor for a point $\mathbf{p}_q$ with neighbors $\mathbf{p}_i$ can be written as

$$FPFH(\mathbf{p}_q) = SPFH(\mathbf{p}_q) + \frac{1}{k} \sum_{i=1}^{k} \frac{1}{w_k} SPFH(\mathbf{p}_i). \tag{3.4}$$

By avoiding the computation of all possible $k$-neighborhood combinations, the FPFH algorithm reaches a computational complexity of $\mathcal{O}(nk)$, while still providing results comparable to those of the PFH descriptor. This speed difference allows FPFH descriptors to be used in real-time applications.

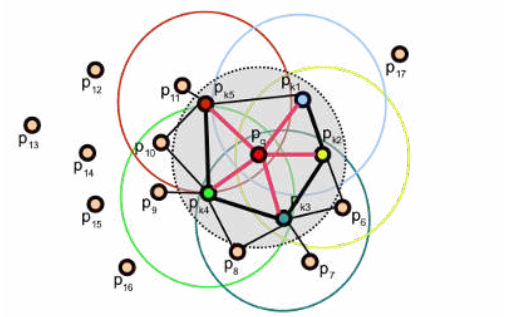



**Figure 3.4:** Visualization of the FPFH neighborhood (Rusu, 2010).

## Random Sample Consensus

Once the 6D orientation-independent feature descriptors have been determined, they can be used to register the point clouds. The well known Random Sample Consensus (RANSAC) algorithm, first presented by Fischler and Bolles (1981), is used for computing the initial alignment.

In general, we usually have a large *target* dataset and a much smaller *source* dataset and we want to find a subset of the target dataset that best corresponds with the source dataset – with respect to some measure. The RANSAC algorithm finds pairs of matching subsets of the two datasets. After a pairing has been determined, its quality is examined by finding a transformation that aligns the matching points by minimizing their squared euclidean distance. All this is done multiple times during the execution of the algorithm.

The RANSAC algorithm iteratively chooses a small random subset of the target data and checks its correspondence with the source data points. RANSAC is only guaranteed to converge to a correct solution when a sufficient number of iterations is used. The algorithm depends on three parameters which are

- *error tolerance*, determining a threshold value for compatibility of two point sets,
- *number of iterations*, determining the number of randomly selected subsets to try out, and
- *threshold value*, being the number of compatible points required to imply that the correct model has been found.

The error tolerance and threshold values are usually determined experimentally. Since absolute correspondence between the target and the model can not be guaranteed, both of these parameters should be large enough to accept some varia-

tion.

The number of required RANSAC iterations can be estimated. If the number of points in the target dataset is $N$ and the number of inliers, i.e., points corresponding to the source, is $m$, then the probability that a randomly selected point is an inlier is $w = \frac{m}{N}$. If we select a subset of $n$ random points, then all of them are inliers with probability $w^n$, and at least one is an outlier with probability $1 - w^n$. If we denote by $p$ the probability that all selected points are inliers and take $k$ iterations, we get the equality

$$1 - p = (1 - w^n)^k.$$

By taking logarithms, we arrive at

$$k = \frac{\log(1 - p)}{\log(1 - w^n)}.$$

## Iterative Closest Point

After initial alignment, the alignment can be further refined using the Iterative Closest Point (ICP) algorithm. ICP is a fairly simple algorithm that consists of the following steps executed iteratively

1. Pair points from the target cloud with the closest points from the model cloud.
2. Estimate a transformation that minimizes the least squares error between the pairs.
3. Apply the transformation to the model cloud.
4. Check if convergence criteria is met.

Since the source and model datasets are of different size, the ICP algorithm is very sensitive for the initial alignment and, as such, is not suitable for registration by itself.

# Chapter 4

# Extracting 3D models of the vocal tract

Until recently, the research in medical imaging methods has concentrated on producing visualizations meant for evaluation by trained medical professionals. Computational methods require more precise description of the geometry. As a consequence, many medical imaging methods are not applicable as such. This chapter presents methods for extracting descriptions of the 3D VT geometry from Magnetic Resonance (MR) images. These methods rely on the algorithms presented in Chapter 3. The extraction is ultimately done for a dataset consisting of several thousands of images. Thus, avoiding manual work has been the major motivation in developing these methods.

## 4.1 VT geometry extraction problem

The interior of the VT consists of the volume starting from the trachea and ending to the mouth and, if the nasal tract is open, to the nostrils. The data for this work has been obtained using a Siemens Magnetom Avanto MRI device. The goal of the extraction is to produce geometries of the VT configurations during sustained pronunciation of Finnish vowels. An example of the resulting geometries can be seen in Figure 4.1.

To obtain a geometry, a 3D image of the subject is taken with the MRI machine. The machine takes a sequence of evenly spaced images – in some orientation – and they are then combined into a voxel image. The voxel values correspond to the measured tissue densities at the voxel locations. Methods described in Chapter 3 are then used to extract the VT geometry from the voxel image.

Unfortunately, the described procedure is not straightforward, and there are var-

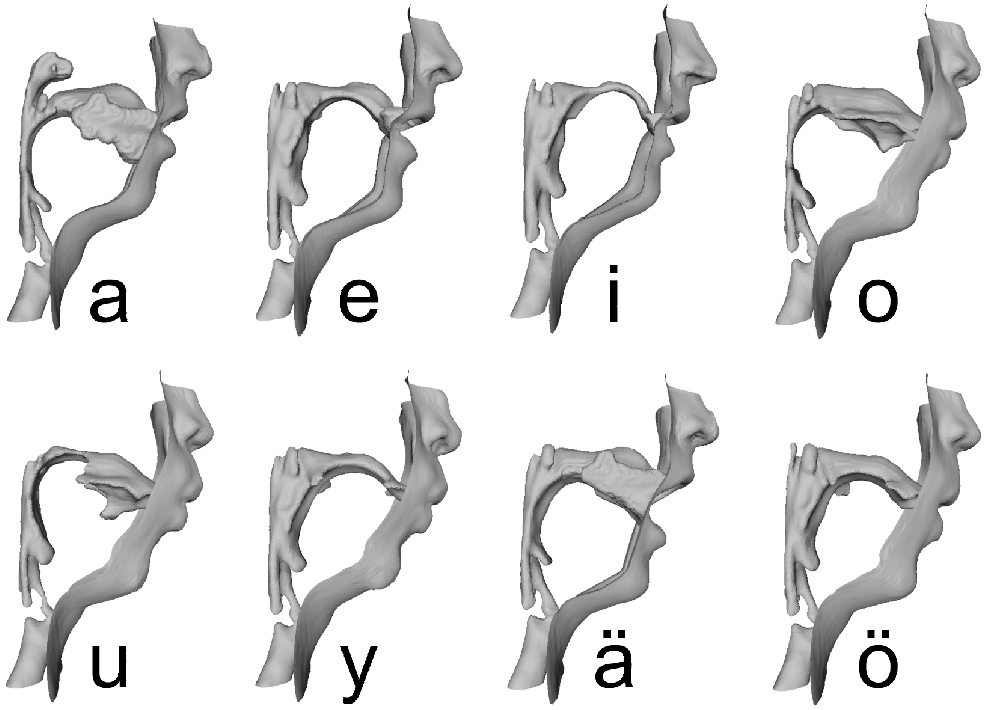



**Figure 4.1:** VT geometries of Finnish vowels from a male subject. The geometry for [a] shows part of the nasal tract.

ious issues that have to be tackled in the extraction process. For example, the tissue density reported by the MRI machine is based on the amount and state of the hydrogen atoms in the imaged tissue. As a consequence of this, osseous tissues, such as the maxillae and the mandible, are indistinguishable from air in the image data. This results in the merging of the maxillae and mandible with the VT air volume in a naive extraction, as shown in Figure 4.2.

Acoustic modelling of the VT geometry will not be possible if the maxillae and mandible are merged with the VT acoustic space. This problem is solved by *registering* externally constructed maxillae and mandible models with the image data. The registration information is then used to *mask* the unwanted volume from the data. The registration process will also account for variations between different measurements, i.e, VT configurations, of the same subject. In general, only one model for the maxillae and mandible is required for each subject.

In addition to the abovementioned challenges, all the usual medical image analysis problems apply here as well. The images are noisy, and MR images tend to have measurement artefacts resulting from the reconstruction algorithm of the MRI machine, movement of the subject, or presence of materials that cause disturbances

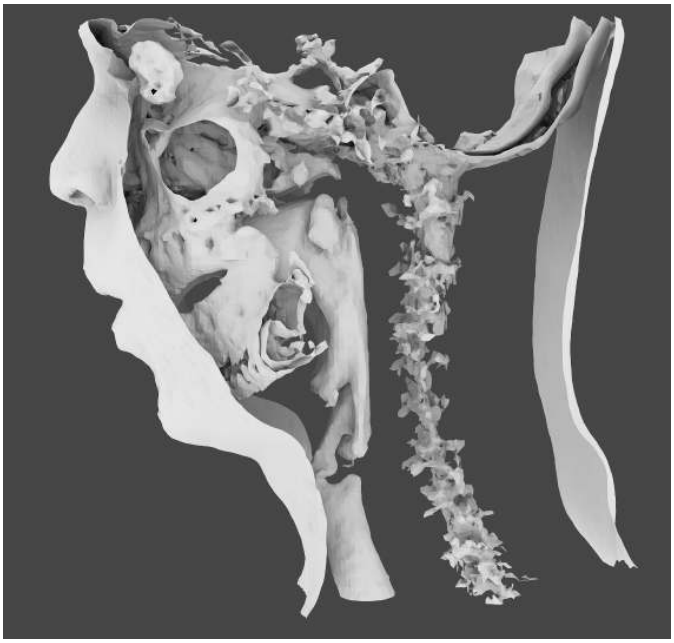


**Figure 4.2:** A naively extracted isosurface of a VT geometry.

in the magnetic resonances (Aalto et al., 2011a,b). In some cases, the resolution of the images is very close to the size of the imaged objects, e.g., vocal folds, leading to higher relative errors. Subject specific anatomical variations can also be significant, and this is particularly true for anatomically non-typical patients. Since the process is highly automatized, there should be a way, at least visually, to verify the quality of the result.

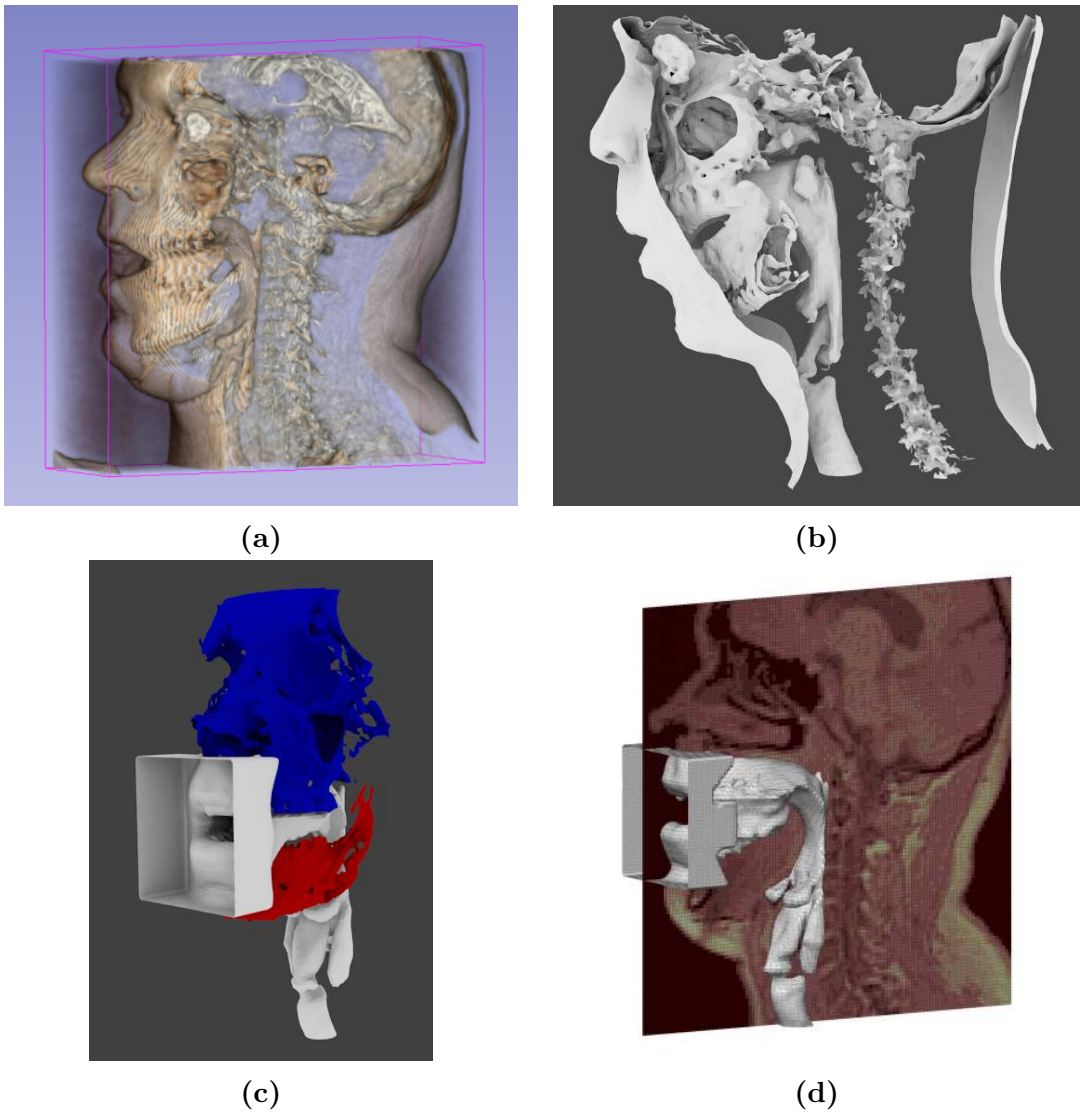
## 4.2 Overview of the algorithm

The procedure designed for the VT extraction consists of the following steps:

1. **Image Acquisition**
   The image is acquired using MRI and stored in DICOM format. A volume render of the voxel data inside the DICOM image can be seen in Figure 4.3a.
2. **Initial feature extraction**
   An isosurface corresponding to the tissue-air interface is extracted from the voxel image using the marching cubes algorithm described in Section 3.2. Figure 4.3b shows the resulting initial surface.
3. **Registration**
   Separate surface models for the maxillae and the mandible are registered with the initial surface using methods discussed in Section 3.3. Colored and aligned maxillae and mandible models can be seen in Figure 4.3c.
4. **Masking**
   The surface models are transformed using the registration information, voxelized, and used to mask out the maxillae and the mandible from the MRI voxel data.
5. **Final feature extraction**
   A new isosurface is extracted from the masked data. This isosurface can be further refined using the parametric deformable models described in Section 3.2.
6. **Mesh post-processing**
   The quality of the extracted triangular mesh can be improved using techniques described in Section 4.2.
7. **Verification**
   The resulting geometry is visually verified. Figure 4.3d shows a surface model plotted with a slice of the original voxel data from the DICOM image.

### Registration

The maxillae and mandible models for registration have been manually extracted for each subject from a single initial surface. The maxillae and mandible are rigid structures, and the registration algorithm can account for changes in their alignment between different images of the same subject. The used registration method does not attempt to scale the models. The approximate nature of the



**(a)** **(b)**

**(c)** **(d)**

**Figure 4.3:** (a) A volume render of DICOM data (Step 1). (b) A naively extracted isosurface. (c) The maxillae and mandible models aligned with the initial surface (Steps 3 and 4). (d) Verification against the original data (Step 7).

registration algorithm can account for small scale changes, but care has to be taken that the models have been extracted using the same isolevels and that they have the same scale as the target surface.

Medical imaging devices present the images in their own coordinate systems. Usually the image files do not contain accurate spatial information for anatomical features within the image data. Registration can be seen as a way to obtain such spatial information for the specified anatomic features. An alternative to matching predefined models is to let a human operator annotate anatomic features from the images.

Markers can also be added to the data to aid automatic registration algorithms. For example, MRI visible markers can be attached to patient's teeth, and in vascular modelling the patient can be injected with a chemical that increases contrast in the MR images (Antiga, 2002).

### Benefits of parametric deformable models

For most geometries, the naive extraction of the isosurface is adequate, considering the mathematical models used. However, if the response of the MRI machine has not been stable across the VT or the isolevel is chosen incorrectly, the final result can have crude errors. Test subjects may also have constricted airways whose dimensions are near the resolution of the MRI machine. All of these challenges can, to some extent, be countered by using parametric deformable models or similar approaches.

### Mesh post-processing

The quality of the extracted surface mesh can be improved by re-meshing it. The purpose of the re-meshing is to improve the quality of mesh faces, e.g., by removing small and degenerate faces (such as triangles with a very sharp angles). For physical models, it is usually necessary to maintain some features of the mesh, such as the surface area and/or the volume. Many algorithms for mesh refinement exist, and most Finite Element mesh generators do mesh refinement as well (Si, 2006; Antiga, 2002).

## 4.3 Area functions for Webster's equation

The Webster's equation presented in Chapter 2 makes use of a one-dimensional representation of the VT. We call this presentation an *area function*. The area function is a function $A(s)$ that gives the areas of cross-sectional slices of the VT taken along a centerline curve $\gamma(s)$ (see Fig. 4.4).

Construction of the area function requires a parametrised centerline $\gamma$ for the domain $\Omega$. The parameter $s$ is usually taken as the curve length of $\gamma$. The centerline algorithm of the VMTK toolkit is used for obtaining the centerline. The algorithm is based on finding the Voronoi sheets for the surface $\partial\Omega$ and then tracking a path defined by the center points of the maximal inscribed spheres. However, some approximations are required since the Voronoi sheets are not manifold surfaces, and they are sensitive to small surface disturbances (Styner et al., 2003).

After the centerline $\gamma(s)$ has been obtained, the area functions can be produced by slicing the VT surface mesh with planes perpendicular to the centerline. The area function produces an approximate representation of the VT domain. A single area function is plotted in Figure 4.4.

It should be noted that there are several ambiguities associated with finding centerlines and area functions. First, the centerline, or the topological skeleton, can be defined in many ways. The centerline used in this work is the one obtained from the Voronoi-sheet-based algorithm of the VMTK library.

Second, since the area functions are obtained by only considering the part of the cross-sectional slice that is connected to the centerline, some information potentially left out. This can be seen by comparing the area function in Figure 4.4 with the surface meshes seen in Figure 4.3. In particular, valleculae and fossae piriformes do not show up in the area function. Webster's model assumes a strictly tubular domain with no branching, and, therefore, there are no trivial ways to compensate for these errors.

Third, once a centerline is chosen and the cross-sectional slices have been fixed, there are actually many possible lines of different length within the VT domain that would correspond to the same cross-sections (with different parametrisation). Figure 4.4 shows several centerlines that are all perpendicular to the shown set of slices.

As Webster's model is sensitive to the length of the centerline, one may obtain different results by choosing different centerlines. In reality, the centerline in Webster's model corresponds to an *acoustic centerline*. For example, in a straight pipe the acoustic centerline would naturally be the real centerline of the pipe, but in an

annulus-shaped domain the acoustic centerline will be frequency-dependent and closer to the inner curve of the annulus. In addition, the acoustic centerline will depend globally on the geometry, i.e., pieces of curved pipes joined will have a different acoustic centerline than the individual pieces have.

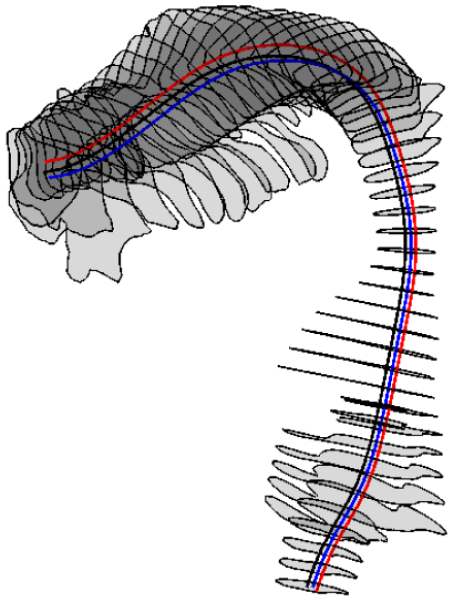


**Figure 4.4:** An area function and multiple matching centerlines.

## 4.4 Shortcomings, alternatives, and future work

Patient-specific variation and artefacts in MR images can still cause problems in the processing of some images. Features whose size is close to the resolution of the MR image are naturally harder to extract. Some of these challenges can be countered with parametric deformable models. However, they have not been extensively used in producing the geometries that are used in the computational part of this work.

The nasal tract is an example of a small cavity that is difficult to extract. For most subjects, its size is close to the resolution of the MR image. In addition, the manually extracted maxillae models are so crude that they tend to completely mask the nasal tract. Reliable extraction of the nasal tract requires higher resolution images. The measurement setup for this work has been designed to produce images

within a certain time frame which limits the attainable resolution of our image data. In terms of acoustics modelling, the impact of the nasal tract is minor, and most vowels don't involve the nasal tract at all. Webster's equation as a model for speech acoustics does not involve the nasal tract.

We have also chosen an approach where we first construct the surface geometry of the VT and then generate a tetrahedral mesh inside that surface. It is possible to directly generate a tetrahedral mesh in the whole image volume and then partition it in order to extract wanted features (see, e.g., Alliez et al., 2014). However, even with this approach one needs a way to get rid of the artefacts caused by the maxillae and the mandible.

All the challenges discussed here are common in imaging applications, and there exists a wide array of algorithms suitable for implementing different parts of the overall extraction pipeline. As mentioned before, computational modelling and geometric analysis of the resulting VT geometry imposes additional requirements for the results when compared to medical visualization, and in recent years, the use of computational models in medical imaging has become more common (see, e.g. Samani et al., 2001; Antiga, 2002).

The image processing algorithms presented here have been implemented in various open source libraries. The most widely used libraries include the Visualization Toolkit (VTK) and the Insight Segmentation and Registration Toolkit (ITK) (Schroeder et al., 2001; Ibanez et al., 2005). Some parts of the implementation make use of the Vascular Modeling Toolkit (VMTK) by Antiga et al. (2008), which itself relies heavily on the VTK and ITK libraries. Vascular Modelling Toolkit has been developed for the purposes of medical-image-based modelling of blood vessels. In addition, the registration algorithms in the Point Cloud Library (PCL) are used. PCL is a collaborative effort to develop efficient algorithms for point cloud processing, which is widely used in, e.g., machine learning and robotics applications (Rusu and Cousins, 2011).

# Chapter 5

# Vocal tract acoustics

In this chapter, VT acoustics are studied by computing the acoustic resonances with the Helmholtz model and Webster's resonance model described in Chapter 2. The geometry data consists of all the Finnish vowels, imaged from a Finnish 26-year-old male. All vowels have multiple measurements. The geometries have been obtained using the procedures described in Chapter 4. The resonance equations are solved using the Finite Element Method (FEM). We also compare the obtained results to the vowel formants extracted from sound samples recorded simultaneously with the image acquisition.

## 5.1 Data acquisition

The data has been obtained using a Siemens Magnetom Avanto 1.5 T scanner (Siemens Medical Solutions, Erlangen, Germany). The maximum gradient field strength of the system is 33 mT/m ($x$,$y$,$z$ directions) and the maximum slew rate is 125 T/m/s. A 12-element Head Matrix Coil and a 4-element Neck Matrix Coil are used to cover the vocal and nasal tracts from the lips and nostrils to the beginning of the trachea. The coil configuration allows the use of Generalize Auto-calibrating Partially Parallel Acquisition (GRAPPA) technique to accelerate acquisition.

3D VIBE (Volumetric Interpolated Breath-hold Examination) MRI sequence (Rofsky et al., 1999) is used to allow the rapid 3D acquisition required for the experiments. Sequence parameters are optimized in order to maximize resolution and minimize the acquisition time as explained in Aalto et al. (2014). The data used in this work was produced in 2012 from a Finnish 26-year-old male subject. An experimental arrangement has been developed (see Aalto et al., 2011a, 2014) to collect a combined data set of MRI and speech samples.

The MRI sequence output is stored as a DICOM file that comprises of 44 sagittal

plane images; each image has a resolution of $128 \times 128$ pixels of size $d = 1.9\,\text{mm}$. These form an array of voxel data consisting of $44 \times 128 \times 128$ voxels, each having a value (between 0 and 800) describing the response of the MRI machine at the location of the voxel. The alignment of the voxels is carried out using the location data produced by the MRI machine.

## 5.2 Finite element modelling

We will use the *Finite Element Method* (FEM) to solve the mathematical acoustic models presented in Chapter 2. A short description of FEM is given here, but for general error analysis and convergence properties the reader is referred to, e.g., Braess (2007) or Johnson (2012).

### Preliminaries

In order to solve this equation using FEM we will derive its *weak formulation*. The weak formulations are obtained by multiplying the equation with a test function

$$\varphi \in V = \{f \in H^1(\Omega) : f(\mathbf{r}) = 0 \text{ for } \mathbf{r} \in \Gamma_m\}$$

and integrating over the geometric domain $\Omega$. Dirichlet boundary conditions are taken into account in the choice of the function space. Other boundary conditions are embedded into the variational formulation, this is usually done by applying Green's formula.

It can be shown that finding a $\Phi_\lambda \in V$ that satisfies the weak formulation for all $\varphi \in V$ is equivalent with the following energy minimization problem

$$\text{Find } \Phi_\lambda \text{ such that } F(\Phi_\lambda) \leq F(\varphi) \text{ for all } \varphi \in V, \tag{5.1}$$

where $F : V \to \mathbb{R}$ is an energy functional of the problem. Physically this corresponds with finding a solution that satisfies the principle of minimum potential energy when all possible solution functions are taken into account.

In practice, the space $V$ has infinite dimension and, thus, it is not possible to find a solution (i.e. function) that can be described with a finite number of parameters. To obtain a problem that can be solved using a computer, we construct a finite dimensional subset $V_h \subset V$ of the original function space. This is done by discretizing, i.e., meshing, the space and representing the functions using a finite number of parameters.

The discretization elements, the functions, and the degrees of freedom of the functions on the elements can be chosen in multiple ways. These three choices define the type of the *finite element* used. In this work tetrahedrale elements with first-order, i.e., linear, basis functions are used.

## Tetrahedral mesh generation

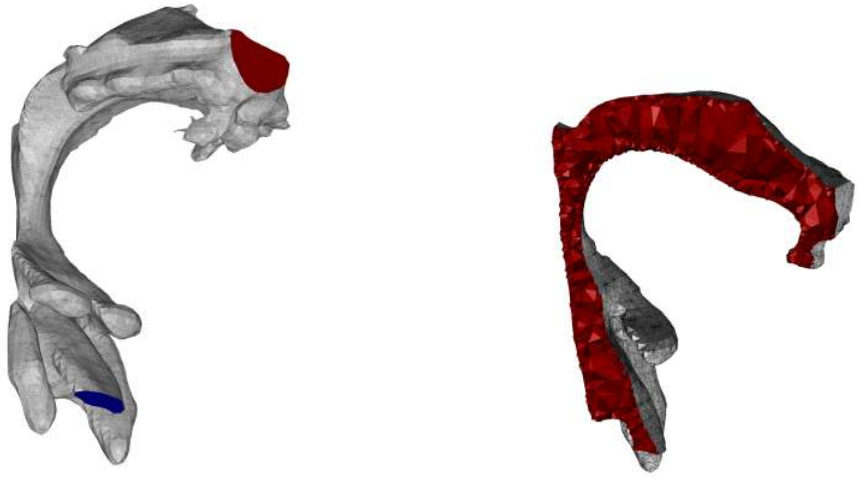
The Finite Element (FE) solver requires discretization of the domain. In the case of a 3D domain this discretization will be a tetrahedral mesh. This mesh is generated for each of the VT geometries obtained using the methods described in previous chapters.

Since FEM is widely used in engineering applications, there are many software packages available for mesh generation. We have used TetGen (Si, 2006) in this work.

The mesh generator takes a closed surface representing the VT volume where the glottis and mouth boundaries have been marked; this can be seen in Figure 5.1a. It produces a tetrahedral mesh that can be used as an input for the FEM solver. A cross-section of a tetrahedral mesh can be seen in Figure 5.1b.

The quality of the computational result is dependent on the quality of the tetrahedral mesh. The size and the shape of the elements affect the accuracy of the computational result. In general, using smaller elements produces better results. However, the use of smaller elements requires use of more elements, which will require more computational resources. Mesh quality can be inspected visually and by checking against different mesh quality criteria (see, e.g. Si, 2006). The degree to which the results are sensitive to the mesh quality also depends on the FEM model used. The FEM models used in this work are not very sensitive to the quality of the mesh.

Although mesh generation is a common engineering problem, most engineering applications are designed to create meshes for simple geometries. Obviously, the VT can have quite a complex shape, and there has been varying success with different mesh generators. However, the quality of the surface mesh is the most important factor in successful mesh generation.



**(a)** **(b)**

**Figure 5.1:** An example of a tetrahedral mesh. (a) Boundary conditions colored. (b) A cross-cut of the mesh showing some of the tetrahedra inside.

## Helmholtz equation

The full Helmholtz equation with boundary conditions is

$$c^2\nabla^2\Phi_\lambda + \lambda^2\Phi_\lambda = 0 \text{ in } \Omega \tag{5.2}$$

$$\lambda\Phi_\lambda + c\frac{\partial\Phi_\lambda}{\partial\nu} = 0 \text{ on } \Gamma_g \tag{5.3}$$

$$\alpha\lambda\Phi_\lambda + \frac{\partial\Phi_\lambda}{\partial\nu} = 0 \text{ on } \Gamma_w \tag{5.4}$$

$$\Phi_\lambda = 0 \text{ on } \Gamma_m. \tag{5.5}$$

Multiplying by a test function $\varphi \in H^1_{\Gamma_m}$ and integrating over $\Omega$ we obtain

$$c^2\int_\Omega \nabla^2\Phi_\lambda\varphi\,\mathrm{d}\mathbf{r} + \lambda^2\int_\Omega \Phi_\lambda\varphi\,\mathrm{d}\mathbf{r} = 0. \tag{5.6}$$

Using the Green's formula yields

$$c^2\int_\Omega \nabla\Phi_\lambda\nabla\varphi\,\mathrm{d}\mathbf{r} - c^2\int_\Gamma \frac{\partial\Phi_\lambda}{\partial\nu}\varphi\,\mathrm{d}\mathbf{s} + \lambda^2\int_\Omega \Phi_\lambda\varphi\,\mathrm{d}\mathbf{r} = 0, \tag{5.7}$$

where the boundary $\Gamma = \Gamma_g \cup \Gamma_w \cup \Gamma_m$. Due to the Dirichlet condition (5.5) the integral over $\Gamma_m$ vanishes since the test function $\varphi = 0$. For the boundary integrals

over $\Gamma_g$ and $\Gamma_w$, the partial derivatives are transformed according to (5.3) and (5.4), respectively. Thus we get

$$c^2 \int_\Omega \nabla\Phi_\lambda \nabla\varphi \, \mathrm{d}\mathbf{r} + \lambda^2 \int_\Omega \Phi_\lambda \varphi \, \mathrm{d}\mathbf{r} + \lambda \left[ c \int_{\Gamma_g} \Phi_\lambda \varphi \, \mathrm{d}\mathbf{s} + \alpha c^2 \int_{\Gamma_w} \Phi_\lambda \varphi \, \mathrm{d}\mathbf{s} \right] = 0. \quad (5.8)$$

FEM assembly of the variational forms in Equation (5.8) results in the corresponding matrix equation

$$\lambda^2 \mathbf{M}\mathbf{x}(\lambda) + \lambda \left[ c\mathbf{P}_g + \alpha c^2 \mathbf{P}_w \right] \mathbf{x}(\lambda) + c^2 \mathbf{K}\mathbf{x}(\lambda) = 0, \quad (5.9)$$

where we have the *mass matrix* $\mathbf{M}$, *stiffness matrix* $\mathbf{K}$, and matrices $\mathbf{P}_g$ and $\mathbf{P}_w$ presenting the boundary conditions at glottis $\Gamma_g$ and VT walls $\Gamma_w$, respectively. These matrices are symmetric square matrices and their dimensions are dependent on the discretization of the domain, i.e., the FEM mesh.

Equation (5.9) is an eigenvalue problem that can be solved by writing it in a form of generalized eigenvalue problem

$$\mathbf{A}\mathbf{y}(\lambda) = \lambda \mathbf{B}\mathbf{y}(\lambda), \quad (5.10)$$

where

$$\mathbf{A} = \begin{bmatrix} -\left[c\mathbf{P}_g + \alpha c^2 \mathbf{P}_w\right] & -c^2\mathbf{K} \\ \mathbf{I} & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{M} & 0 \\ 0 & \mathbf{I} \end{bmatrix} \quad \text{and} \quad \mathbf{y}(\lambda) = \begin{bmatrix} \lambda\mathbf{x}(\lambda) \\ \mathbf{x}(\lambda) \end{bmatrix}.$$

This and other methods for formulating eigenvalue problems are discussed, for example, in Saad (1992).

If the quality of the discretization is high enough, the eigenvalues $\lambda_i$ of the problem (5.10) correspond to the $\lambda$ in Eq. (5.2). The lowest resonance frequencies $R_1, R_2, \ldots,$ correspond to the eigenvalues $\lambda_i$ ordered with increasing imaginary part. Furthermore, the eigenvectors $\mathbf{x}(\lambda_i)$ give an approximation of the velocity potential $\Phi_{\lambda_i}$ at mesh nodes.

## Time-harmonic Webster's equation

The variational formulation for the time-harmonic Webster's equation is obtained similarly to the Helmholtz equation. Leaving the parameter $s$ implicit for $\Phi_\lambda$, $\varphi$, $A$, $W$, and $\Sigma$ we start with

$$\frac{\lambda^2}{c^2}\frac{1}{\Sigma^2}\Phi_\lambda + \lambda\frac{2\pi\alpha W}{A}\Phi_\lambda - \frac{1}{A}\frac{\partial}{\partial s}\left(A\frac{\partial\Phi_\lambda}{\partial s}\right) = 0 \text{ on } [0, L] \quad (5.11)$$

$$\lambda\Phi_\lambda - c\Phi_\lambda' = 0 \text{ at } s = 0 \quad (5.12)$$

$$\Phi_\lambda = 0 \text{ at } s = L \quad (5.13)$$

which is multiplied by $A(s)$ and $c^2$ before applying the test function and integrating to obtain

$$\lambda^2 \int_L \Phi_\lambda \varphi \frac{A}{\Sigma^2} \,\mathrm{d}s + \lambda c^2 2\pi\alpha \int_L \Phi_\lambda \varphi W \,\mathrm{d}s - c^2 \int_L \frac{\partial}{\partial s}\left(A \frac{\partial \Phi_\lambda}{\partial s}\right) \varphi \,\mathrm{d}s = 0. \tag{5.14}$$

The integration by parts formula then yields

$$\lambda^2 \int_L \Phi_\lambda \varphi \frac{A}{\Sigma^2} \,\mathrm{d}s + \lambda c^2 2\pi\alpha \int_L \Phi_\lambda \varphi W \,\mathrm{d}s - c^2 \Big|_0^L A \Phi_\lambda' \varphi + c^2 \int_L \Phi_\lambda' \varphi' A \,\mathrm{d}s = 0. \tag{5.15}$$

Which further simplifies to

$$\lambda^2 \int_L \Phi_\lambda \varphi \frac{A}{\Sigma^2} \,\mathrm{d}s + \lambda c^2 2\pi\alpha \int_L \Phi_\lambda \varphi W \,\mathrm{d}s + \lambda c \Phi_\lambda \varphi A + c^2 \int_L \Phi_\lambda' \varphi' A \,\mathrm{d}s = 0. \tag{5.16}$$

FEM assembly for this problem is done using linear 1D elements. The resulting eigenvalue problem is identical to (5.9).

## 5.3 Results

The results have been computed using a custom FEM solver implemented in MATLAB. The solver uses first order linear elements (e.g., Lagrange elements), in 1D for Webster's equation and in 3D for the Helmholtz equation. The solvers and their results have been verified against artificial objects and objects representing simplified vocal tracts. Simplified geometries and an area function inside the surface of the corresponding tetrahedral mesh can be seen in Figure 5.3. We have used the sound speed $c = 350\,\mathrm{m/s}$ and the VT wall dissipation coefficient $\alpha = 7.6 \times 10^{-7}\,\mathrm{s/m}$.

The difference between resonance frequencies and formants can be measured in semitones using a *discrepancy* calculation defined as

$$D(f_1, f_2) = 12 \ln\left(\frac{f_1}{f_2}\right) / \ln 2.$$

We computed the first six resonances with both acoustics models. The resulting resonances and their discrepancies are given in Table 5.1. The first six Helmholtz

**Table 5.1:** (a) Mean Helmholtz resonances. (b) Mean Webster resonances. (c) Mean discrepancies between Helmholtz and Webster resonances.

**(a)**

| | $H_1$ | $H_2$ | $H_3$ | $H_4$ | $H_5$ | $H_6$ |
|---|---|---|---|---|---|---|
| [a] | 538 | 1410 | 2659 | 2970 | 3057 | 3110 |
| [e] | 334 | 1943 | 2704 | 2876 | 3413 | 3485 |
| [i] | 195 | 2042 | 2986 | 3126 | 3378 | 3626 |
| [o] | 467 | 1040 | 1925 | 2682 | 2917 | 3005 |
| [u] | 245 | 909 | 2132 | 2867 | 2943 | 3060 |
| [y] | 272 | 1519 | 1814 | 2202 | 2674 | 2861 |
| [ae] | 524 | 1590 | 2253 | 2848 | 2920 | 3033 |
| [oe] | 391 | 1326 | 1887 | 2035 | 2866 | 3094 |

**(b)**

| | $W_1$ | $W_2$ | $W_3$ | $W_4$ | $W_5$ | $W_6$ |
|---|---|---|---|---|---|---|
| [a] | 621 | 1501 | 2979 | 3328 | 4401 | 5264 |
| [e] | 346 | 2067 | 2882 | 3047 | 4294 | 5761 |
| [i] | 203 | 2228 | 3125 | 3455 | 4276 | 5759 |
| [o] | 544 | 1191 | 2582 | 3163 | 4138 | 4757 |
| [u] | 270 | 1060 | 2341 | 3151 | 3504 | 4331 |
| [y] | 281 | 1796 | 2489 | 3082 | 3784 | 4390 |
| [ae] | 573 | 1711 | 3022 | 3323 | 4420 | 5687 |
| [oe] | 420 | 1531 | 2569 | 3085 | 4075 | 5153 |

**(c)**

| | $D(H_1, W_1)$ | $D(H_2, W_2)$ | $D(H_3, W_3)$ | $D(H_4, W_4)$ | $D(H_5, W_5)$ | $D(H_6, W_6)$ |
|---|---|---|---|---|---|---|
| [a] | -2.58 | -3.01 | -4.11 | -2.31 | -6.83 | -8.93 |
| [e] | -0.61 | -1.07 | -1.11 | -1.00 | -3.97 | -8.70 |
| [i] | -0.63 | -1.53 | -0.70 | -1.65 | -4.35 | -8.17 |
| [o] | -2.62 | -2.30 | -5.83 | -2.89 | -6.04 | -7.95 |
| [u] | -1.68 | -2.61 | -1.62 | -1.63 | -3.03 | -6.03 |
| [y] | -0.56 | -2.91 | -5.48 | -5.82 | -6.03 | -7.41 |
| [ae] | -1.57 | -1.24 | -5.53 | -2.68 | -7.19 | -10.89 |
| [oe] | -1.25 | -2.62 | -6.09 | -8.16 | -6.27 | -8.94 |

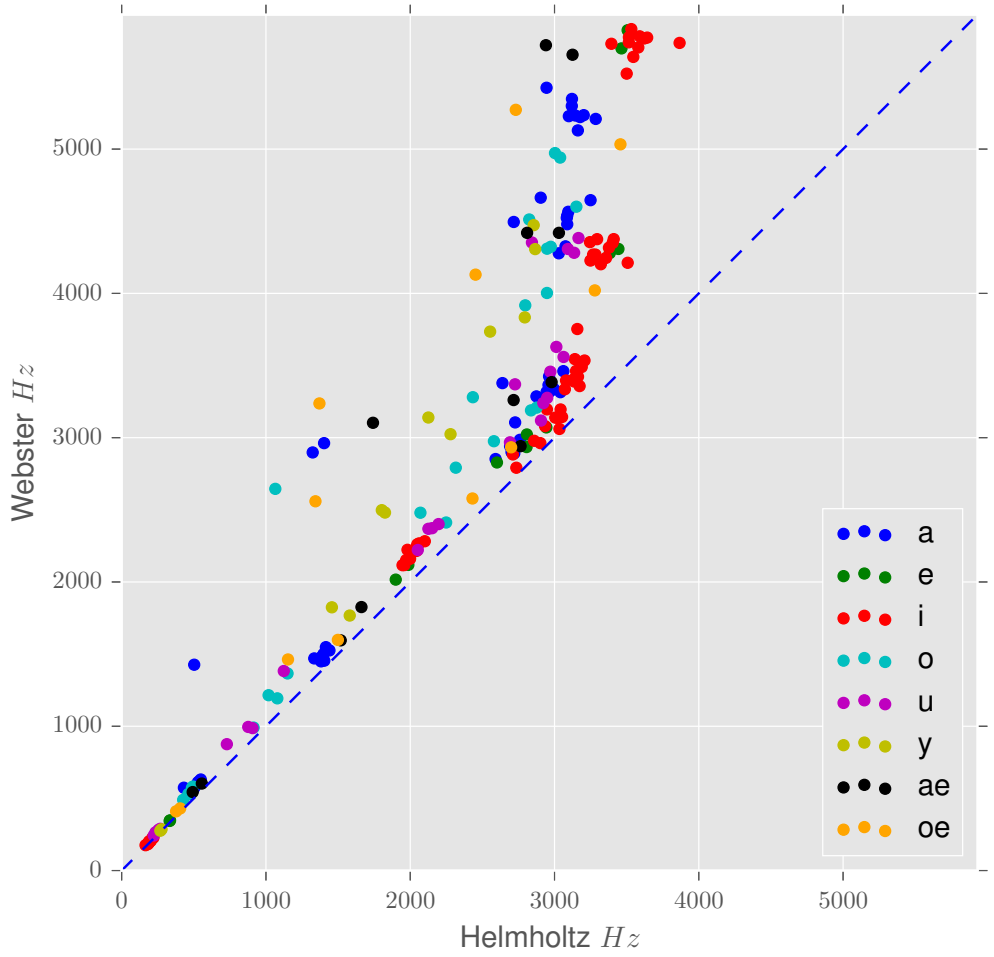



**Figure 5.2:** First six Helmholtz resonances plotted against Webster resonances.

and Webster resonances have been plotted against eachother in Figure 5.2. Figures 5.4 and 5.5 show the first four resonances plotted against formants extracted from sound samples.

The lower resonances for Webster's model and Helmholtz model match well for tube-like domains with no significant curvature, but neither of these assumptions is completely true for the VT geometries. The lowest four resonances, which are the most relevant ones, match well. Higher resonances start to mismatch because Webster's model only captures resonances in the longitudinal direction. Most of the differences in the lowest four resonances for the same vowel are due to a naive way of automatically choosing the location of the glottis and mouth boundaries. Each resonance of a single vowel should form same-colored group of dots in Figure 5.2.

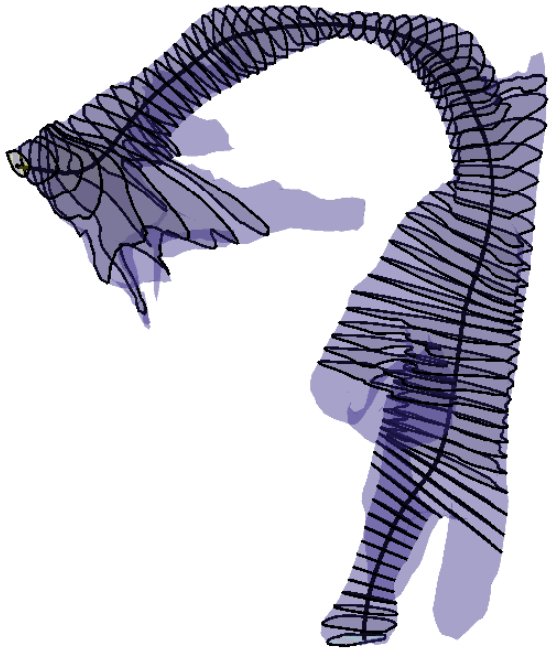


**Figure 5.3:** Surface of the tetrahedral mesh with the area slices for vowel [u]. The area function cuts out the piriform sinuses. The geometry also shows appendages left by the geometry extraction with a crude mandible model.

As an example of naive choice of boundary, if the teeth structures are not completely cleaned from the surface, narrow appendages to the geometry will result; this can be seen in Figure 5.3. In some of these cases the boundary has been chosen so as to cut out the appendages. This results in higher resonances because the distance from the glottis to the mouth will be shorter.

For some geometries, the appendages have not been cut out. Leaving the appendages will have no significant effect on Webster resonances, but in the Helmholtz model they act as low pass filters for the signal and result in significantly lower, and arguably totally unrealistic, resonances. The effect can be seen for some geometries of [a], [o], [y], [ae], and [oe] in Figure 5.2. This error should be countered by adjusting the geometry extraction to properly remove the teeth structures.

As was discussed earlier, the results from Webster's model can also change if a different centerline is chosen. We have used a spline interpolant of the centerline

vertices returned by the VMTK library.

We have only extracted the lowest four formants from the sound samples corresponding to the geometries. The formants differ more compared to the resonances. The Helmholtz and Webster's models are, in some sense, idealized, whereas the formants have been extracted from actual sound samples. Extracting the formants requires cancellation of the MRI machine noise from the sound samples. A significant amount of signal processing is involved in the formant extraction and as a result the extraction can sometimes produce spurious results. (Kuortti and Malinen, 2015)

Aside from computational and geometrical considerations, there are several real life sources of error. During the measurement the subject is laying inside an MRI machine in supine position for a period of about 1 hour. It is difficult to control errors that might be caused by the subject moving inside the machine during imaging, etc. Even for a trained person, which the test subject is not, it would be very difficult to consistently produce exactly the same utterances.

For the purposes of speech research, we are mostly interested in frequencies below 5 kHz, where human hearing is most sensitive. Higher resonances will also be numerically more unstable and thus no longer contain significant information. In fact, in most languages the vowel sounds can be classified using just the two lowest resonances.

To give more statistical credibility for the results, studying a larger set of data from multiple subjects will be required. That will be left for future work. It is likely that advances in medical imaging technology will make this kind of analysis cost-effective and routine. In the future taking time-dependent 3D images, i.e. 3D video, will also be a possibility (Gamper et al., 2008; Kampf et al., 2010). This will allow studying even more complicated phenomena, provided that suitable methods are developed.

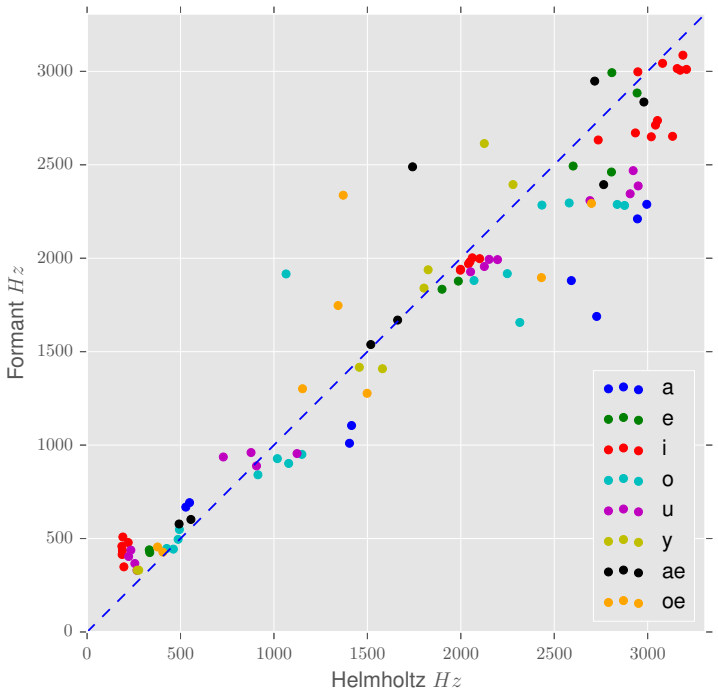



**Figure 5.4:** First four Helmholtz resonances plotted against Formants.

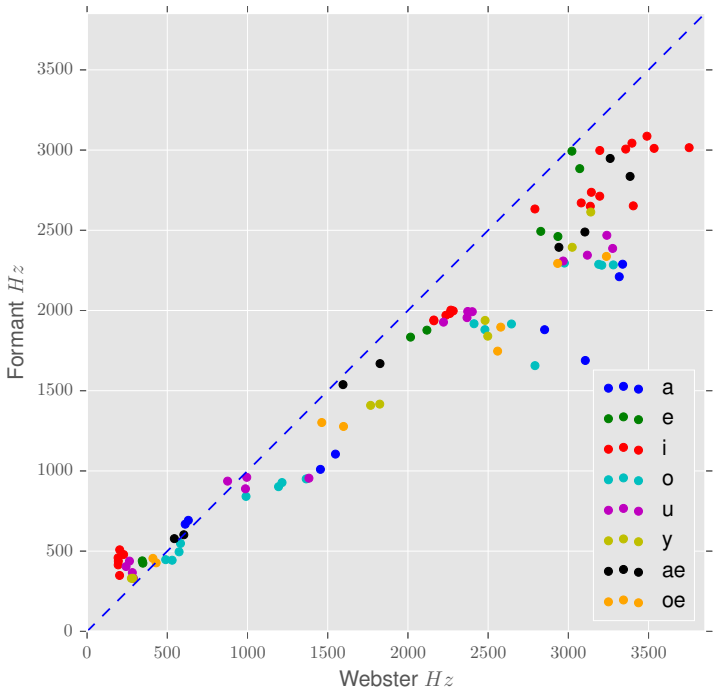



**Figure 5.5:** First four Webster resonances plotted against Formants.

# Chapter 6

# Discussion

The main part of this work considers methods for extracting vocal tract geometries (Chapter 3 and Chapter 4). These geometries were used to model vocal tract acoustics during vowel production. The computational results were presented in Chapter 5.

## Chapter 2

The acoustics models used in this work consisted of the Helmholtz and Webster's resonance equations. We used simple boundary conditions to present the glottis signal, VT walls, and mouth excitation. Especially the Dirichlet boundary condition at the mouth can be considered a rudimentary approximation for the exterior space, which is expected to have a significant effect on the acoustics. Nevertheless, this kind of models are widely used in speech research applications and they are adequate for many practical and theoretical applications. Due to the high computational requirements of large volume acoustic models, some kind of approximations are required for joining an exterior space model with the VT.

The presented mathematical models can also be used in time-dependent modelling. Modelling time-dependence will increase the computational cost of the models significantly, especially if the exterior space is to be modelled. Obtaining accurate time-dependent geometries is also problematic, although approximations, such as extrapolating the area function from a dynamic MR image of the mid-sagittal plane, are certainly possible (see, e.g., Baer et al., 1991). Some emerging MRI techniques also allow acquisition of dynamic 3D MR images of the whole speech apparatus.

## Chapter 3

The image processing methods presented here are all well known. They have many applications and they have been studied for a long time. They are also related to many other research areas, such as machine learning, robotics, and data analysis. Using multi-modal datasets effectively is also an important area of research. Due to the active and long continuing research, there is a significant amount of different methods available. Therefore, choosing the right ones and combining them to a working solution is not necessarily an easy task. In fact, model selection is a research area in itself. Simpler methods tend to be more effective, and this work mainly uses methods that are well known and widely used.

## Chapter 4

The primary goal of this chapter was to define a procedure for extracting VT geometries from MRI data. MRI is well suited for obtaining a large number of images since it does not produce ionizing radiation like CT procedures do. Ultrasound imaging has also been used in speech research applications, but the resolution is not adequate for our purposes. However, MR images contain no information about osseous tissue; this is a problem in imaging the VT since the air volume is directly connected to the teeth, maxillae, and mandible, causing them to merge with the interior of the VT in MR images.

A procedure for masking out the maxillae and mandible has been described. This allows the extraction of VT geometries suitable for computational modelling of VT acoustics. Since thousands of MR images need to be processed, minimization of manual work in the geometry extraction has been an important objective.

The maxillae and mandible were masked out from the MRI data by registering manually constructed models for the maxillae and mandible with the MR data. Since the maxillae and mandible are rigid structures, only one manually constructed model for each is required per one test subject. The maxillae and mandible models were constructed manually for this work. In principle, the maxillae and mandible models could also be obtained using different medical imaging methods such as a CT scan of the teeth.

The current procedure can not extract the nasal tract. In many subjects, the resolution of the MR image would also likely be inadequate for identifying the nasal tract. However, the nasal tract does not play a part in most vowel sounds. The used version of Webster's model is also not suitable for geometries with multiple

acoustic exits, although it can be extended to consider branching domains (Aalto et al., 2015a).

One problem in automatic geometry extraction for computational purposes is determining the position and shape of the glottis and mouth boundaries. This is currently done using the area function: the locations are heuristically determined, and the boundaries are assumed to be straight plane cuts. The location and shape of the boundaries may have to be more intelligently determined for higher accuracy modelling.

The procedure seems to perform fairly well for the restricted data set used in this work. However, variations in subject anatomy and artefacts in MR images may still cause problems. Many of these problems can be treated using the parametric deformable models and mesh post processing methods that have been discussed. Also, crude artefacts can not trivially be detected in the image data. Thus visual inspection of the result is of utmost importance. In the context of speech research, visual inspection of the computational geometry is not always adequate since it may be impossible, without additional information, to verify whether the geometry actually presents some specific utterance.

The presented procedure was successfully used to extract geometries for the computational experiments presented in Chapter 5. This can be considered as validation for the presented methods.

## Chapter 5

The acoustic resonance properties of the VT configurations extracted from MR images were studied computationally. The VT configurations were extracted from MR images taken during the pronunciation of Finnish vowels. The test subject was a 26-year-old male who is a native Finnish speaker.

The Finite Element Method (FEM) was used to solve the acoustic models described in Chapter 2. The obtained computational resonances were compared to formant resonances determined from sound samples that were recorded simultaneously with the MR imaging.

The results compare well with similar data found in literature. The 3D resonance model shows that it is possible to have significant transversal resonances inside the human VT. These transversal resonances are within the normal hearing range, although above 4 kHz. Transversal resonances may explain the spectral clustering, a phenomena that has been consistently observed in formant measurements from

phonations of trained singers. Some vowel formant variations between different subjects could also be explained by transversal resonances.

The simple acoustic models used in this work are quite forgiving in terms of the quality of detail in the VT geometry: small perturbations do not have significant effects on the results. As explained before, the most significant errors are a result of the naive boundary conditions. Computational modelling of the exterior space requires further research. Limits in computational capacity impose some restrictions, and it is likely that modelling of the exterior space requires a model of reduced dimensions.

Most existing speech models are used in telephony, speech synthesis, speech analysis, and medical applications for which they have been specifically developed. Such models are usually simplified. The presented model is far from a full speech model, but it still shows that accurate models of human speech can produce new information. For example, information about the pressure distribution of resonances inside the VT could be used as an aid in surgery planning. Simpler models will not be able to produce accurate enough information. There are other similar models, see e.g., Takemoto et al. (2010); Vampola et al. (2013); Arnela et al. (2013). The produced geometries could be used for many purposes, for example, in flow mechanical and time-dependent models, as well as in refinement and verification of existing speech models.

The main outcome of this work has been the verification that processing a significant amount of MRI-acquired VT data automatically is possible. More accurate modelling can potentially provide new information and insights in speech research and other fields.

# Bibliography


Aalto, A. 2009. *A low-order glottis model with nonturbulent flow and mechanically coupled acoustic load*. Masters thesis, Helsinki University of Technology.

Aalto, A., Lukkari, T., and Malinen, J. 2015a. Acoustic wave guides as infinite-dimensional dynamical systems. ESAIM: Control, Optimisation and Calculus of Variations, 21(2):324–347.

Aalto, A., Murtola, T., Malinen, J., Aalto, D., and Vainio, M. 2015b. Modal locking between vocal fold and vocal tract oscillations: Simulations in time domain. Technical report.

Aalto, D., Aaltonen, O., Happonen, R.-P., Jääsaari, P., Kivelä, A., Kuortti, J., Luukinen, J. M., Malinen, J., Murtola, T., Parkkola, R., Saunavaara, J., and Vainio, M. 2014. Large scale data acquisition of simultaneous MRI and speech. Applied Acoustics, 83(1):64–75.

Aalto, D., Aaltonen, O., Häppönen, R.-P., Malinen, J., Palo, P., Parkkola, R., Saunavaara, J., and Vainio, M. 2011a. Recording speech sound and articulation in MRI. In *Proceedings of BIODEVICES 2011*, 168–173, Rome.

Aalto, D., Malinen, J., Vainio, M., Saunavaara, J., and Palo, J. 2011b. Estimates for the measurement and articulatory error in MRI data from sustained vowel phonation. In *Proceedings of the International Congress of Phonetic Sciences*, 180–183.

Alku, P., Horacek, J., Airas, M., Griffond-Boitier, F., and Laukkanen, A.-M. 2006. Performance of Glottal Inverse Filtering as Tested by Aeroelastic Modelling of Phonation and FE Modelling of Vocal Tract. Acta Acustica united with Acustica, 92(5):717–724.

Alliez, P., Rineau, L., Tayeb, S., Tournois, J., and Yvinec, M. 2014. 3D Mesh Generation. In *CGAL User and Reference Manual*, CGAL Editorial Board, 4.5 edition.

Angenent, S., Pichon, E., and Tannenbaum, A. 2006. Mathematical methods in medical image processing. Bulletin of the American Mathematical Society, 43(3):365–396.

Antiga, L. 2002. Patient-specific modeling of geometry and blood flow in large arteries. Politecnico di Milano.

Antiga, L., Piccinelli, M., Botti, L., Ene-Iordache, B., Remuzzi, A., and Steinman, D. A. 2008. An image-based modeling framework for patient-specific computational hemodynamics. Medical & biological engineering & computing, 46(11):1097–1112.

Arnela, M., Guasch, O., and Alías, F. 2013. Effects of head geometry simplifications on acoustic radiation of vowel sounds based on time-domain finite-element simulations. The Journal of the Acoustical Society of America, 134(4):2946–2954.

Baer, T., Gore, J., Gracco, L., and Nye, P. 1991. Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. The Journal of the Acoustical Society of America, 90(2):799–828.

Braess, D. 2007. *Finite elements: Theory, fast solvers, and applications in solid mechanics*. Cambridge University Press.

Dromey, C., Ramig, L. O., and Johnson, A. B. 1995. Phonatory and Articulatory Changes Associated With Increased Vocal Intensity in Parkinson DiseaseA Case Study. Journal of Speech, Language, and Hearing Research, 38(4):751–764.

Fant, G. 1971. *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*. Walter de Gruyter.

Fischler, M. A. and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6):381–395.

Flanagan, J. L. 1972. *Speech analysis: Synthesis and perception. (2nd ed.)*, volume x. Springer-Verlag, Oxford, England.

Gamper, U., Boesiger, P., and Kozerke, S. 2008. Compressed sensing in dynamic MRI. Magnetic Resonance in Medicine, 59(2):365–373.

Hannukainen, A., Lukkari, T., Malinen, J., and Palo, P. 2007. Vowel formants from the wave equation. Journal of the Acoustical Society of America, 122(1):EL1–EL7.

Hannukainen, A., Ojalammi, A., and Malinen, J. 2014. Exterior space model for an Acoustic Eigenvalue Problem. In *Proceedings of 27th Nordic Seminar on Computational Mechanics*, 85–88.

von Helmholtz, H. 1912. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Longmans, Green.

Ibanez, L., Schroeder, W., Ng, L., and Cates, J. 2005. *The ITK Software Guide*. Kitware, Inc., second edition.

Johnson, C. 2012. *Numerical solution of partial differential equations by the finite element method*. Courier Corporation.

Kampf, T., Fischer, A., Basse-Lüsebrink, T., Ladewig, G., Breuer, F., Stoll, G., Jakob, P., and Bauer, W. 2010. Application of compressed sensing to in vivo 3D 19 F CSI. Journal of Magnetic Resonance, 207(2):262–273.

Kuortti, J. and Malinen, J. 2015. Post-processing speech recordings during MRI.

Lorensen, W. E. and Cline, H. E. 1987. Marching cubes: A high resolution 3D surface construction algorithm. In *ACM Siggraph Computer Graphics*, volume 21, 163–169, ACM.

Lukkari, T. and Malinen, J. 2013. Webster's equation with curvature and dissipation. ArXiv:1204.4075.

Lukkari, T. and Malinen, J. 2015. A posteriori error estimates for Webster's equation in wave propagation. Journal of Mathematical Analysis and Applications, 427(2):941–961.

Murtola, T. 2014. *Modelling vowel production*. Licentiate Thesis, Aalto University School of Science, Department of Mathematics and Systems Analysis.

Rofsky, N., Lee, V., Thomassen, D., and others. 1999. The volumetric interpolated breath hold examination (VIBE): a new approach to body MR imaging. Radiology, 212:876–884.

Rusu, R. B. 2010. Semantic 3D object maps for everyday manipulation in human living environments. KI-Künstliche Intelligenz, 24(4):345–348.

Rusu, R. B. and Cousins, S. 2011. 3d is here: Point cloud library (pcl). In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 1–4, IEEE.

Saad, Y. 1992. *Numerical methods for large eigenvalue problems*, volume 158. SIAM.

Samani, A., Bishop, J., Yaffe, M. J., and Plewes, D. B. 2001. Biomechanical 3-D finite element modeling of the human breast using MRI data. Medical Imaging, IEEE Transactions on, 20(4):271–279.

Schroeder, W., Martin, K., Avila, L. S., and Law, C. C. 2001. The Visualization Toolkit User's Guide, Version 4.0. Kitware, version, 4.

Si, H. 2006. A quality tetrahedral mesh generator and three-dimensional delaunay triangulator. Weierstrass Institute for Applied Analysis and Stochastic, Berlin, Germany.

Skodda, S., Visser, W., and Schlegel, U. 2011. Vowel articulation in Parkinson's disease. Journal of Voice, 25(4):467–472.

Styner, M., Gerig, G., Lieberman, J., Jones, D., and Weinberger, D. 2003. Statistical shape analysis of neuroanatomical structures based on medial models. Medical image analysis, 7(3):207–220.

Svancara, P. and Horacek, J. 2006. Numerical Modelling of Effect of Tonsillectomy on Production of Czech Vowels. Acta Acustica united with Acustica, 92(5):681–688.

Takemoto, H., Mokhtari, P., and Kitamura, T. 2010. Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method. The Journal of the Acoustical Society of America, 128(6):3724–3738.

Vampola, T., Horácek, J., Laukkanen, A.-M., and Švec, J. G. 2013. Human vocal tract resonances and the corresponding mode shapes investigated by three-dimensional finite-element modelling based on CT measurement. Logopedics Phoniatrics Vocology, 40(1):14–23.