Aalto University
School of Science

Tiina Murtola

Modelling Vowel Production

Thesis submitted in partial fulfilment of the requirements for the degree of Licentiate of Science in Technology

Espoo, 11.06.2014

Supervisor: Professor Timo Eirola

Instructor: Jarmo Malinen D.Sc. (Tech)

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076
AALTO
www.aalto.fi

Abstract of licentiate thesis

Author Tiina Murtola	
Title of thesis Modelling Vowel Production	
Department Department of Mathematics and Systems Analysis	
Field of research Mathematics	
Supervising professor Prof. Timo Eirola	**Code of professorship** F006Z
Thesis advisor(s) Jarmo Malinen, D.Sc. (Tech)	
Thesis examiner(s) Prof. Paavo Alku	
Number of pages 74	**Language** English
Date of submission for examination 11.06.2014	

Abstract

This thesis is focused on describing and testing a computationally light vowel synthesis model, which can be used to generate glottal flow pulses for more sophisticated acoustic simulators of the vocal tract. The core of the model consists of a low-order mass-spring system that represents the vocal folds, Bernoulli flow with viscous pressure loss in the glottis, and a Webster resonator that represent the vocal tract. The Webster resonator makes use of centreline and area function data which have been extracted from magnetic resonance images.

With the aim of producing a minimal model, new elements are added to the model one by one, and the impact of the added complexity is investigated. These additions include dissipation along the vocal tract, a horn-shaped Webster resonator to represent the subglottal tract, and losses caused by turbulence in the glottis. In addition, technical changes are also introduced which allow the model to be used with any vocal tract geometry and in a large number of simulations.

For such model to be of practical use, it must be able to produce glottal flow with a variety of fundamental frequencies and phonation types. This tunability is achieved by optimising four selected parameters. Solving the multi-objective optimisation problem directly is not practical due to the complicated dynamic behaviour of the model and long computing time of each simulation. Instead, a three-step procedure combining constrained single-objective optimisation, parameter space exploration, and manual pulse shape selection is introduced. Three well-known direct search optimisation algorithms, pattern search, simulated annealing, and genetic algorithm, are tested for the optimisation step. A pattern search-based algorithm is developed for pathwise parameter space exploration. Finally, the use of the closing quotient, a pulse shape parameter, as an aid for the final selection is tested.

Keywords Speech production, glottal pulse generator, glottal flow, mechano-acoustic model, parameter tuning

Aalto-yliopisto
Perustieteiden
korkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi

Lisensiaatin tutkimuksen tiivistelmä

Tekijä Tiina Murtola

Työn nimi Vokaalintuoton mallintaminen

Laitos Matematiikan ja systeemianalyysin laitos

Tutkimusala Matematiikka

Vastuuprofessori Prof. Timo Eirola **Professuurikoodi** F006Z

Työn ohjaajat Jarmo Malinen, TkT

Työn tarkastajat Prof. Paavo Alku

Jätetty tarkastettavaksi 11.06.2014 **Sivumäärä** 74 **Kieli** Englanti

Tiivistelmä

Tässä työssä kuvataan ja testataan laskennallisesti kevyt vokaalisynteesin malli, jolla voidaan tuottaa glottisvirtauspulsseja monimutkaisempien akustisten ääntöväylämallien tarpeisiin. Mallin ydin koostuu äänihuulia kuvaavasta matala-asteisesta massajousisysteemistä, Bernoullin lain mukaisesta, viskoosin painehäviön huomioivasta virtauksesta glottiksessa, sekä ääntöväylää kuvaavasta Websterin resonaattorista, joka käyttää hyväkseen magneettiresonanssikuvista erotettuja keskiviivoja ja pinta-alafunktioita.

Työn tavoite on kehittää minimaalinen malli. Tätä silmällä pitäen malliin lisätään uusia elementtejä yksi kerrallaan, ja lisääntyneen monimutkaisuuden vaikusta tarkastellaan. Näitä uusia elementtejä ovat kudoshäviöt ääntöväylässä, torvenmallinen Websterin resonaattori, joka edustaa glottiksen alapuolisia ääniväylän osia, sekä turbulenssin aiheuttamat häviöt glottiksessa. Lisäksi esitellään teknisiä muutoksia, jotka mahdollistavat mallin käytön minkä tahansa ääntöväylägeometrian kanssa suuressa määrässä simulaatioita.

Mallin hyödyntäminen käytännössä vaatii, että sillä pystytään tuottamaan glottispulsseja useilla eri perustaajuuksilla ja fonaatiotavoilla. Tämä viritettävyys saavutetaan optimoimalla neljän parametrin arvot. Monitavoiteoptimointiongelman ratkaisu suoraan ei ole käytännöllistä mallin monimutkaisen dynaamisen käytöksen ja pitkän laskenta-ajan vuoksi. Vaihtoehtona esitellään kolmiaskelinen menetelmä, jossa yhdistyvät rajoitettu yksitavoiteoptimointi, etsintä parametriavaruudessa sekä pulssimuodon manuaalinen valinta. Optimointiaskelta varten kokeillaan kolmen tunnetun gradienttivapaan algoritmin (pattern search, simulated annealing, genetic algorithm) toimivuutta ongelmassa. Parametriavaruudessa tapahtuvaa etsintää varten kehitetään pattern search-algoritmiin pohjautuva polkuetsintämetodi. Lopuksi testataan yhden pulssimuotoparametrin (closing quotient) soveltuvuus manuaalisen valinnan apuvälineeksi.

Avainsanat Puheen tuotto, glottispulssigeneraattori, glottisvirtaus, mekaanis-akustinen malli, parametrien viritys

Acknowledgments

I would like to thank my supervisor, Prof. Timo Eirola, and my instructor, Dr. Jarmo Malinen, for providing guidance when it was needed but allowing me to muddle through the project much as I wished otherwise. I am also grateful to Atte Aalto for access to and help with the original model and codes and for valuable comments on the manuscript, and to Atle Kivelä for MRI data processing. There are also a large number of other people who have been less directly involved in the production of this thesis - collaborators, colleagues, friends and family - who have my gratitude for their help, support and tolerance.

Espoo, 11 June 2014

Tiina Murtola

Contents

Abbreviations

ClQ closing quotient.

FEM finite element method.

MRI magnetic resonance imaging.

OQ open quotient.

SGT subglottal tract.

VT vocal tract.

Chapter 1

Introduction

Speech production is a complicated phenomenon which is of interest not only from the point of view of science but also of every day life. The ability of a person to function in human society depends on their capability to communicate efficiently and reliably with others. In this respect, producing and understanding speech is one of the most important social skills a person develops, and a devastating one to lose.

Against this background, it is not surprising that recent decades have seen numerous studies aiming at improving understanding and modelling of human voice production. Increasing knowledge of this phenomenon helps, for example, in diagnosing and treating speech disorders, and predicting and minimising the effect surgeries in the mouth and neck area have on an individual's speech. And of course, applications in communications and speech synthesis technologies should not be forgotten.

1.1 Human voice production

According to a fairly simplified view, human voice is produced using the vocal apparatus shown in Figure 1.1. The lungs are an air reservoir kept at constant pressure by muscle action. Air flows out of the lungs through the trachea and into the vocal tract (VT). Between these two channels, vocal folds form an orifice, called the glottis. Flow through the glottis induces vibrations in the vocal folds, causing the glottis to open and close periodically. Hence air flow enters the VT in the form of pulses. The VT, which filters this glottal flow, consists of the larynx, pharynx, and oral and nasal cavities. The position of the velum determines the extent to which the nasal cavity is coupled to the rest of the VT.

The vocal folds are two opposing tissue bodies which stretch across the larynx. Their positions, elongations and tensions can be controlled using phonatory mus-

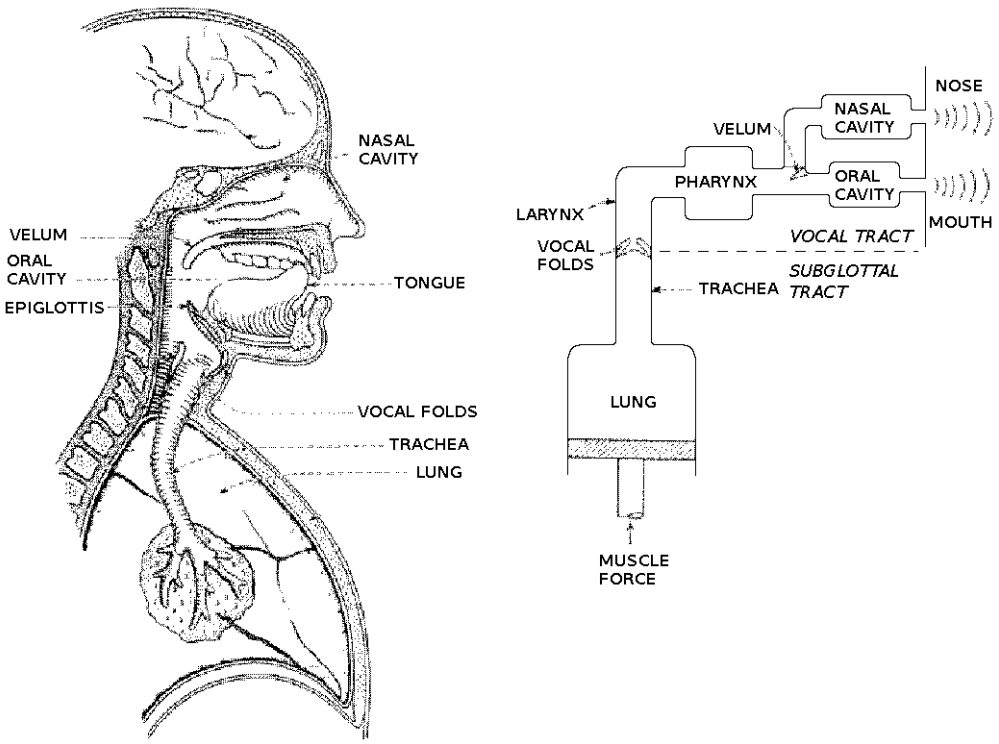


Figure 1.1: The physiological vocal apparatus and a schematic picture of the same modified from figure presented in Flanagan (1972).

cles in the larynx. The geometry of the VT can be altered for example by moving the tongue and jaws, and its termination condition can be changed by moving the lips. Pressure at the lungs is also controlled by the speaker. Such active controls make it possible to produce a wide variety of speech sounds at different pitches, i.e. perceived fundamental frequencies, and intensities.

This source-filter view of speech production, and models based on it, such as the one discussed in this work, are particularly well suited for vowel production. During the pronunciation of Finnish [ɑ, e, i, o, u, y, æ, œ], the VT is open at every point and sound is mostly transmitted to the external world through the mouth opening. Changes particularly in the position of the tongue and the degree of constriction at the mouth opening alter the eigenfrequencies of the air column in the VT. These frequencies, known in phonetics as formants, differentiate the vowels from each other.

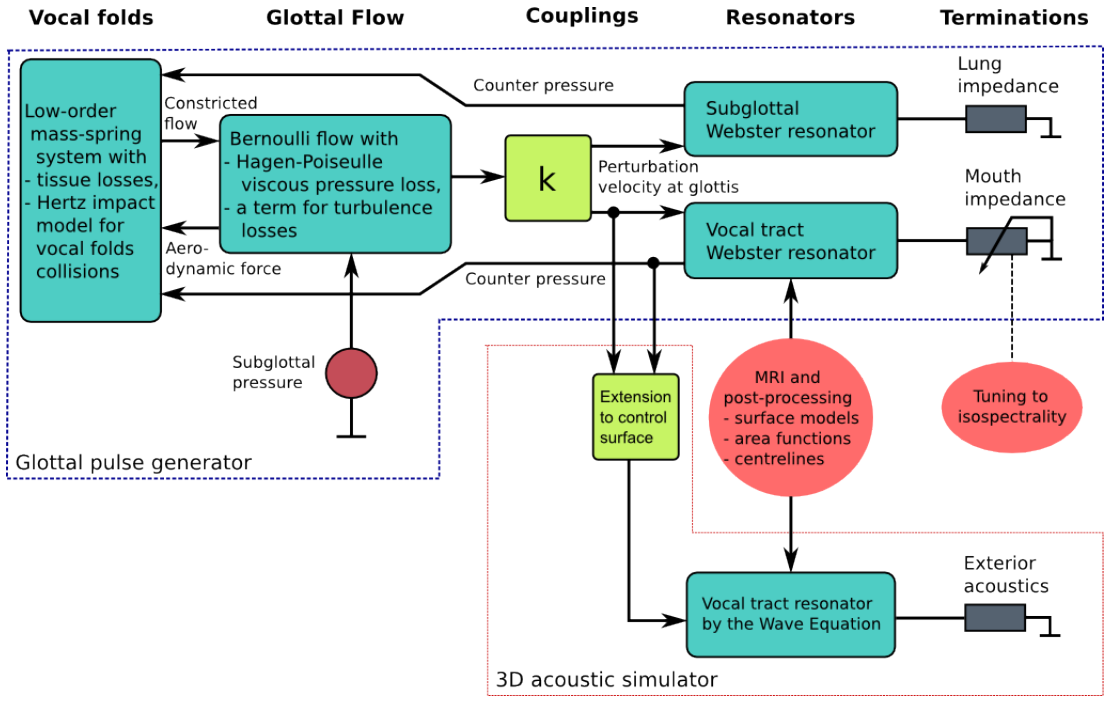


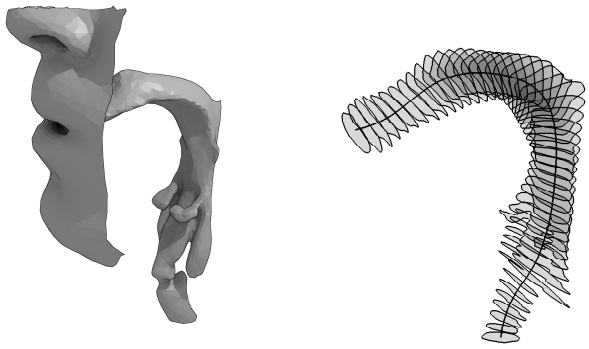
Figure 1.2: Complete model for vowel production (Comspeech@Aalto 2013).

1.2 Background and aim

This work is part of a wider effort to model the acoustics of vowel production. The full vowel production model consists of two parts: a 3-dimensional finite element method (FEM) -based VT resonator and a glottal pulse generator that consists of several submodels, including more simple (Webster) resonators for the vocal and subglottal tracts (Figure 1.2).

Both VT resonators make use of magnetic resonance imaging (MRI) data obtained during long phonation of Finnish vowels; the 3D model as surface models (Figure 1.3a) and the Webster resonator as centrelines and area functions (Figure 1.3b). The MRI data are a part of a coupled data set of sound and images collected keeping specifically in mind validation and parameter estimation needs of mathematical models (see Aalto et al. 2011; Aalto et al. 2014, and references therein for details of data collection and processing). At the time of writing, collection of the first clinically relevant data set is under way and is expected to produce over 2000 pairs of sound and VT geometries within the next few years (Aalto et al. 2014). In fact, a quarter of this data have already been collected.

This thesis focuses on the glottal pulse generator. The low-order glottal flow model introduced by Aalto (2009) is here developed further with the aim of producing a tunable pulse generator that can make use of the centrelines and area



(a) (b)

Figure 1.3: (a) A surface model of the air-tissue interface extracted from MRI data for vowel [œ] and (b) centreline and area function model of the same data.

functions described above. Model development is done using Occam's razor as a modelling paradigm: the simplest model that produces all essential features in the glottal flow is considered the best. Simplicity is desirable both because it improves tractability of the model and because it reduces the computational cost of using the model on large data sets such as the patient data being collected. In order to find the minimal model, the glottal pulse generator is built modular (Figure 1.2) so that the value added by each submodel can be investigated by turning them on and off separately.

For the glottal pulse generator to be of practical use, it must be possible control its output. The features of the glottal pulses that are of most interest in this work are the fundamental frequency of phonation, f_0, and the type of phonation (pressed, normal, or breathy). Hence, the aim of this work is to produce a model that can be tuned so that the glottal pulses it produces match any reasonable f_0 (both male and female) and phonation type targets.

1.3 Outline of this work

To start off with, Chapter 2 gives a brief overview of literature on vowel synthesis models, their control parameters, and how such parameters can be tuned. After this, the model and its tuning are treated separately.

First, in Chapter 3, the model used in this work is discussed. This includes a review of the model by Aalto (2009), addition of dissipation along the VT, addition of losses caused by turbulence in the the glottis, and presentation of a new subsystem to represent the subglottal tract (SGT). Changes required to make the model adaptable to different VT geometries and a large number of simulations

are also discussed here. Finally, at the end of this chapter, the model is tested and the impact of the changes made are investigated and conclusions are drawn on them.

In the second part of this work, Chapter 4, attention is turned to tuning, i.e. control of the model. First, the problem setup and its key features are discussed. Then, the methods used to find a combination of tuning parameter values which produces glottal flow with desired f_0 and phonation type are presented. Finally, results from numerical tuning experiments are presented and discussed.

In the final chapter, the tunable model is reviewed as a whole. Areas requiring future work are also identified and new questions arising from the results shown in the previous two chapters are considered.

Chapter 2

Literature review

This thesis consists of two interlinked parts: a vowel synthesis model and the tuning of its control parameters. In this chapter, a brief look is taken at literature available on each of these topics.

2.1 Vowel synthesis models

The source-filter theory of human voice production leads to a two part model of this phenomenon, one that consists of a source and a filter.

Since Flanagan and Landgraf (1968) published a one-mass model of the vocal folds and Ishizaka and Flanagan (1972) its successor, the classic two-mass model, lumped-parameter models have been used widely both as an object of research (see e.g. Lous et al. 1998; Horáček and Švec 2002; Horáček et al. 2005; Aalto 2009) and as a signal source (e.g. van den Doel and Ascher 2008; Ho et al. 2011). Lumped-parameter models are constructed based on varying degrees of simplifications concerning vocal fold geometry and mechanics as well as air flow through the glottal opening. They can be made computationally light and suitable for real-time simulations but this comes at the cost of reduced faithfulness to the real system making interpretation of results challenging.

More recently, thanks to advances in computer technology, more complex models of the laryngeal region have emerged. For example, multi-layered two- or three-dimensional finite element and finite volume models of vocal folds coupled with Navier-Stokes equations have the potential to capture speech production to a high degree of reality (e.g. Alipour et al. 2000; de Oliveira Rosa et al. 2003; Daily and Thomson 2013). The downside of such models is that they are still computationally cumbersome and hence not suitable for applications where speed is important but computer power limited.

When it comes to the filter part of speech production models, literature shows

again a variety of choices both in what to model and in how to model it. The simplest form of filter consist only of the vocal tract (VT) from the glottis to the lips (e.g. Ishizaka and Flanagan 1972). Better representation of reality can be achieved by including the trachea (Daily and Thomson 2013) or the entire subglottal tract (SGT) (e.g. Birkholz et al. 2007; Ho et al. 2011). Side branches can be added to the VT to represent the nasal tract with or without paranasal sinuses (see Ho et al. 2011; Birkholz et al. 2007, for examples of each, respectively) or to represent piriform fossa (Mokhtari et al. 2008).

Whichever parts of subglottal and supraglottal tracts are included, there are several choices for how to model them. If simplicity is desired, one approach is to build a transmission line to represent each part of the filter (Flanagan 1972; Mokhtari et al. 2008). The transmission line approach is computationally fairly light and has been successfully combined with a tree-like model of the SGT (Ho et al. 2011). However, the computational advantage is somewhat lost if all the transmission-line circuit elements need to be recalculated at each time step, for example due to time-varying cross-sectional area of the VT. Another simple approach, the Kelly-Lochbaum model (Kelly and Lochbaum 1962), suffers from similar problems. Third approach is to use Webster's equation, a simplification of the wave equation (Aalto 2009; van den Doel and Ascher 2008). This approach has not been used together with branching VT or SGT structures, although such constructions have been shown to be well posed (Aalto and Malinen 2013). If more simplicity is desired in the SGT model, the tree-like structure can be reduced to an expanding horn (Birkholz et al. 2007; Lous et al. 1998).

As with source models, increased computational power has recently lead to investigation and development of complex and hence more realistic VT models. For example, if high resolution is more important than computational speed, finite element method (FEM)-based solvers can be used compute the three-dimensional acoustics of realistic VT configurations (Hannukainen et al. 2007; Lu et al. 1993; Suzuki et al. 1993; Švancara et al. 2004; Vampola et al. 2008).

2.2 Control parameters and their tuning

Each of the models discussed above contains inbuilt parameters which need to either be estimated from physical or physiological considerations or be set as control parameters whose values are chosen to achieve some desired outcome. The more simplistic the model, the more likely it is that parameter values need to deviate from reality to compensate for the simplifications of the model. On the other hand, the more complicated the model, the larger the number of parameters, and estimating or controlling these can lead to problems.

In this work, parametric control focus is set on the source side of the model.

Therefore the focus here set on literature available on this topic.

The choice of control parameters is typically made to reflect some physiological mechanisms which are known to affect sound production. Subglottal (i.e. lung) pressure is a typical parameter which is expected to have an impact on speech production and is often straightforward to alter in models (e.g. Ishizaka and Flanagan 1972; Scimarella and d'Alessandro 2004; de Vries et al. 2002; Yang et al. 2011)

There are other often used mechanisms but their implementation depends on the model. For example, vocal fold stiffness, which is altered through muscle action, can be realised in models through altering spring constants directly (e.g. Ishizaka and Flanagan 1972; Scimarella and d'Alessandro 2004) or through choosing the eigenfrequencies of the vocal folds (Horáček et al. 2005). Vocal fold tension, another example, can be controlled through a special tension factor which scales down masses and scales up spring constants (e.g. Ishizaka and Flanagan 1972; Lous et al. 1998; Aalto et al. 2009). This has been observed to be a convenient way to alter the fundamental frequency of phonation, f_0, but the same effect can be achieved by altering the masses and constants separately (Scimarella and d'Alessandro 2004).

The choice of control parameters also depends on which properties of sound production are investigated. Considering the neutral glottal area sufficed for Ishizaka and Flanagan (1972) to make general observations about the limitations of their model. However, Horáček and Švec (2002), for example, broke the neutral glottal area down to vocal fold length and glottal gap in order to investigate the phonation threshold. Considering vocal fold length separately also allows modelling gender differences and adjustments a speaker might make during phonation (Scimarella and d'Alessandro 2004).

The above list is by no means complete. By far the most dominant factor in choosing control parameters is the choice of model and the details of its implementation. Assumptions about vocal fold geometry, symmetry, and fluid dynamics can increase or decrease the number of free parameters significantly and simultaneously lead to suggestions for a sensible choice of control parameters.

Another aspect of interest concerning control parameters is how their values are determined. Literature generally falls within one or both of two categories in this respect. In sensitivity studies, observations are made on how model output responds to changes in control parameter values (e.g. Ishizaka and Flanagan 1972; Horáček et al. 2005; Scimarella and d'Alessandro 2004; de Vries et al. 2002). In contrast, tuning studies focus on searching for sets of control parameter values which produce model output matching some criteria (e.g Aalto et al. 2009; Döllinger et al. 2002; Pinheiro et al. 2012; de Vries et al. 1999; Yang et al. 2011).

In both categories, model output is typically characterised in terms of the fundamental frequency of phonation and qualitative or quantitative description of

the glottal flow pulse shape (e.g Aalto et al. 2009; Ishizaka and Flanagan 1972; Scimarella and d'Alessandro 2004; de Vries et al. 2002). Less common features include vocal fold eigenfrequencies (Horáček et al. 2005; de Vries et al. 1999), glottal contact area (Scimarella and d'Alessandro 2004), phase differences between superior and inferior masses (Ishizaka and Flanagan 1972), and critical volume flows (Horáček and Švec 2002), to name a few. Many tuning studies that aim towards clinical applications, typically pathology detection, use time series of vocal fold deflections or glottal area as these can be extracted from patients using high-speed video (Pinheiro et al. 2012; Pinheiro and Kerschen 2013; Wurzbacher et al. 2006; Döllinger et al. 2002).

Tuning studies are of particular interest from the point of view of this work. Most approaches utilise single-objective optimisation procedures taking into account the non-convex nature of the objective function. This requires either a semi-analytic choice of initial values (Döllinger et al. 2002; Schwarz et al. 2006) or a combination of global and local optimisation (Pinheiro et al. 2012; Pinheiro and Kerschen 2013; Yang et al. 2011). Commonly used optimisation algorithms include Genetic Algorithms (Pinheiro et al. 2012; Pinheiro and Kerschen 2013; Schwarz et al. 2006), Particle Swarm Optimisation (Yang et al. 2011), Simulated Annealing (Wurzbacher et al. 2006; Yang et al. 2011), and Nelder-Mead Algorithm (Döllinger et al. 2002; Pinheiro and Kerschen 2013).

Thanks to the complexity of optimising control parameters, tuning studies are often carried out for modest data sets. The largest data set in the above mentioned studies was used by Schwarz et al. (2006) (430 synthesised and 30 measured signals) who had the benefit of a model simple enough (two-mass model similar to (Ishizaka and Flanagan 1972) but not coupled to acoustic loads) to be inverted semi-analytically. Another impressive effort was made by Yang et al. (2011) who matched the parameters of their 3-dimensional multi-mass model (with 531 degrees of freedom) to 50 synthetically generated samples and later to 24 *in vivo* measurements (Yang et al. 2012). The optimisation for the first task was reported to take 3-5 days.

Chapter 3

Vowel synthesis model

As mentioned in Chapter 1, the purpose of this work is to produce a minimal vowel synthesis model which can be used as a glottal pulse generator. In order to determine which features are necessary for producing essential features of glottal pulses, the model is built in a modular fashion so that the impact of each module can be investigated separately. The modelling approach taken is one of multi-physics; the (partial) differential equations governing each submodel are coupled, and the system is solved numerically without using transmission line or circuit presentations.

In accordance with these modelling principles, the starting point for this work is the low-order vowel synthesis model described by Aalto (2009) and Aalto et al. (2009), which itself relies heavily on the work of Horáček and Švec (2002) and Horáček et al. (2005). This model comprises three subsystems: a two-degree-of-freedom mass-spring model of vocal folds, an incompressible one-dimensional flow through the glottal opening, and a resonator based on Webster's equation, which represents the acoustic load of the vocal tract (VT). In this work, the model is developed further to include (i) a tissue losses in the VT model, (ii) losses caused by turbulence in the glottal flow model, (iii) a fourth subsystem representing the subglottal tract (SGT), and (iv) small adjustment mechanisms which make it easier to use the model with a variety of VT configurations and in large number of simulations.

In this chapter, the vowel synthesis model used in this work is described and the impact of the changes made are investigated. Section 3.1 describes the subsystems of model and the data used to represent the VT in Webster's equation. Next, in Section 3.2, the numerical methods used in simulations are described. Sections 3.3 and 3.4 show and discuss how the changes made impact the model, respectively.

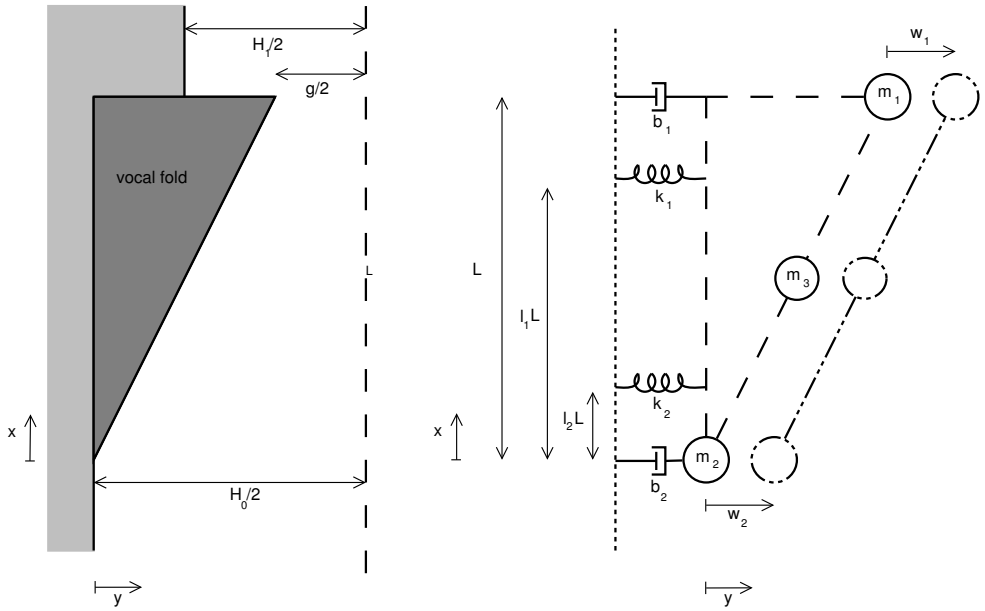


Figure 3.1: Geometry of one vocal fold (left) and the equivalent two-degree-of-freedom model (right).

3.1 The submodels

3.1.1 Vocal folds

Each vocal fold is modelled using a wedge shaped vibrating element of length L suspended from a rigid foundation with springs at $x = l_1 L$ and $x = l_2 L$ and dampers located at each end. The equivalent point mass model for the j^{th} vocal fold consists of three masses: m_{j1} at $x = L$, m_{j2} at $x = 0$, and m_{j3} at $x = L/2$ (Figure 3.1). The vocal fold elements are of constant depth, h, in the z-direction. Physiologically, h corresponds to the length of the vocal folds.

The mechanics of the two-degree-of-freedom (point mass) model of the vocal folds are described by equations of motion

$$\begin{cases} M_1 \ddot{W}_1(t) + B_1 \dot{W}_1(t) + K_1 W_1(t) = -F(t), \\ M_2 \ddot{W}_2(t) + B_2 \dot{W}_2(t) + K_2 W_2(t) = F(t), \end{cases} \tag{3.1}$$

where $W_j = (w_{j1}, w_{j2})^T$ contains the displacements of the ends of the j^{th} vocal fold for $j = 1, 2$, and M_j, B_j, and K_j are the mass, damping, and stiffness matrices for

each vocal fold, respectively. The matrices are given by

$$\begin{aligned} M_j &= \begin{bmatrix} m_{j1}+\frac{m_{j3}}{4} & \frac{m_{j3}}{4} \\ \frac{m_{j3}}{4} & m_{j1}+\frac{m_{j3}}{4} \end{bmatrix}, \\ B_j &= \begin{bmatrix} b_{j1} & 0 \\ 0 & b_{j2} \end{bmatrix}, \\ K_j &= \begin{bmatrix} l_1^2 k_{j1} + l_2^2 k_{j2} & l_1 l_2 (k_{j1}+k_{j2}) \\ l_1 l_2 (k_{j1}+k_{j2}) & l_1^2 k_{j1} + l_2^2 k_{j2} \end{bmatrix}. \end{aligned} \tag{3.2}$$

For this work, the model is simplified further by assuming symmetry of the left and right vocal folds so that $M_1 = M_2$, $B_1 = B_2$, and $K_1 = K_2$.

The load terms in Eq. (3.1), acting on m_1 and m_2, depend on whether the glottis is open or closed. During open state, assuming Bernoulli law for incompressible flow results in a force pair

$$F_A = \begin{bmatrix} \frac{1}{2}\rho v_o^2 h L \left(-\frac{H_1^2}{\Delta W_1 (\Delta W_2 - \Delta W_1)} + \frac{H_1^2}{(\Delta W_2 - \Delta W_1)^2} \ln\left(\frac{\Delta W_2}{\Delta W_1}\right)\right) - \frac{H_1(H_0 - H_1/2)}{4L} h\tilde{p}_c \\ \frac{1}{2}\rho v_o^2 h L \left(\frac{H_1^2}{\Delta W_2 (\Delta W_2 - \Delta W_1)} - \frac{H_1^2}{(\Delta W_2 - \Delta W_1)^2} \ln\left(\frac{\Delta W_2}{\Delta W_1}\right)\right) + \frac{H_1(H_0 - H_1/2)}{4L} h\tilde{p}_c \end{bmatrix}, \tag{3.3}$$

where $v_o(t)$ is velocity of the flow through the control area $H_1 h$ just above the glottis, $\Delta W_1(t) = g + w_{21}(t) - w_{11}(t)$ is the glottal opening at $x = L$ (superior end), $\Delta W_2(t) = H_0 + w_{22}(t) - w_{12}(t)$ is the glottal opening at $x = 0$ (inferior end), and $\tilde{p}_c(t)$ is the counter pressure the acoustic loads exerts on the vocal folds at $x = L$. The counter pressure, $p_c(t)$, given by the VT and SGT models (see Sections 3.1.3 and 3.1.4, especially Eq. (3.11)) is scaled by a factor Q_{pc} (i.e. $\tilde{p}_c(t) = Q_{pc} p_c(t)$) to account for the difficulty in estimating the area on which the counter pressure acts. Without this extra control parameter, overestimation of the acoustic load forces tends to lead to instability in the model.

When the glottis is closed, the only aerodynamic force exerted on the vocal folds is the counter pressure. Hence, when combined with the Hertz impact model, the load terms are given by

$$F_H = \begin{bmatrix} k_H |\Delta W_1|^{3/2} - \frac{H_1(H_0 - H_1/2)}{4L} h\tilde{p}_c \\ \frac{H_1(H_0 - H_1/2)}{4L} h\tilde{p}_c \end{bmatrix}, \tag{3.4}$$

where k_H is the spring constant of the impact. Hence, the load force is

$$F = \begin{cases} F_A, & \text{if } \Delta W_1 \geq 0, \\ F_H, & \text{if } \Delta W_1 < 0. \end{cases}$$

3.1.2 Glottal flow

As previously mentioned, the flow through the glottis is assumed to be one-dimensional and incompressible. Motivated by the Hagen-Poisseulle law, the flow

velocity, v_o, can be modelled with (see derivation in Aalto 2009: ch. 2 and ch. 5)

$$\dot{v}_o(t) = \frac{1}{C_{iner} h H_1} \left(p_{sub} - R_g(t) v_o(t) \right), \tag{3.5}$$

where p_{sub} is the subglottal (lung) pressure above ambient pressure, C_{iner} regulates flow inertia, and $R_g(t)$ represents pressure loss in the glottis. At its simplest form, the loss in the glottis is caused by viscous effects alone, so that

$$R_g(t) = R_v(t) = \frac{C_g}{\Delta W_1(t)^3}, \tag{3.6}$$

where C_g depends on the glottal geometry. A slightly more complex model includes losses caused by turbulence. Following van den Berg et al. (1957), glottal resistance in the model can be written as a sum of the viscous and turbulence terms

$$R_g(t) = R_v(t) + R_t(t) = \frac{C_g}{\Delta W_1(t)^3} + k_g \frac{\rho H_1^2 v_o(t)}{2 \Delta W_1(t)^2}. \tag{3.7}$$

The coefficient k_g in the turbulence term represents the difference between energy loss at the glottal inlet and pressure recovery at the outlet. This coefficient depends not only on the glottal geometry but also on the glottal opening, subglottal pressure, and flow through the glottis (Fulcher et al. 2011). However, in this model, the coefficient is taken to be a constant whose value is determined by trial and error.

3.1.3 Vocal tract

Webster's equation

Webster's horn model resonator is used as an acoustic load to represent the VT. This acoustic load is coupled mechanically to the glottis model through output velocity, v_o, and counter pressure, p_c.

Webster's horn equation with curvature and dissipation is derived and discussed in detail in Lukkari and Malinen (2013). A briefer discussion, which is directly related to the used model but excluding dissipation, can be found in Aalto (2009: chap. 3). Here, only the bare essentials necessary for this application are described.

Webster's equation provides an approximate solution to the wave equation

$$\begin{cases} \Phi_{tt} = c^2 \Delta \Phi, & \text{in} \quad \Omega, \\ \Phi_t + \theta c \frac{\partial \Phi}{\partial \nu} = 0, & \text{on} \quad \Gamma_1, \\ \alpha \Phi_t + \frac{\partial \Phi}{\partial \nu} = 0, & \text{on} \quad \Gamma_2, \\ \frac{\partial \Phi}{\partial \nu} = u, & \text{on} \quad \Gamma_3, \end{cases} \tag{3.8}$$

where Φ is a velocity potential function, Ω the VT, Γ_1 the mouth opening, Γ_2 the walls of the VT, and Γ_3 a control surface above the glottis. The coefficients are c, the speed of sound, θ, a normalised acoustic resistance at the mouth opening, and α, a dissipation coefficient at the VT walls.

Webster's equation is applicable if the VT is approximated as a curved tube of varying cross sectional area. The centreline $\gamma : [0, L_{VT}] \longrightarrow \mathbb{R}^3$ of the tube is parametrised using distance $s \in [0, L_{VT}]$ from the superior end of the glottis. At every s, the cross-sectional area of the tube perpendicular to the centreline is given by the area function, $A(s)$, and the radius of the tube by $R(s)$. The curvature of the tube is defined as $\kappa(s) := \|\gamma''(s)\|$, and the curvature ratio as $\eta(s) := R(s)\kappa(s)$. It is assumed that the tube does not fold on to itself, that is $\eta(s) < 1, \forall s \in [0, L_{VT}]$.

Webster's equation with curvature and dissipation is

$$\frac{1}{c^2\Sigma(s)^2}\frac{\partial^2\psi}{\partial t^2} + \frac{2\pi\alpha W(s)}{A(s)}\frac{\partial\psi}{\partial t} - \frac{1}{A(s)}\frac{\partial}{\partial s}\left(A(s)\frac{\partial\psi}{\partial s}\right) = 0, \tag{3.9}$$

where $\psi(s,t)$ is the velocity potential, and the stretching factor, $W(s)$, and the sound speed correction factor, $\Sigma(s)$, are defined as

$$\begin{aligned} W(s) &:= R(s)\sqrt{R'(s)^2 + (\eta(s)-1)^2}, \\ \Sigma(s) &:= \left(1 + \tfrac{1}{4}\eta^2(s)\right)^{-1/2}. \end{aligned}$$

The boundary conditions in Eq. (3.8) translate to

$$\begin{cases} \frac{\partial\psi}{\partial t}(L_{VT},t) + \theta c\frac{\partial\psi}{\partial s}(L_{VT},t) &= 0, \\ \frac{\partial\psi}{\partial s}(0,t) &= -c_1 v_0(t), \end{cases} \tag{3.10}$$

where the scaling variable c_1 has been added to the original model to extend the assumption of incompressibility from the control area above the glottis to the first VT area slice, i.e. $c_1 = \frac{H_1 h}{A(0)}$.

Coupling of the VT and vocal fold models is done through the counter pressure, p_c, given by Eq. (3.11) below. The pressure $\psi_t(0,t)$ needed for computing p_c is obtained from the solution of Eqs. (3.9)-(3.10).

Vocal tract data

Solving Webster's equation in the VT requires that the VT is represented with an area function and a centreline, from which curvature information can be computed. These are extracted from magnetic resonance imaging (MRI) data which were collected during long phonation of Finnish vowels. The extraction methods are discussed elsewhere (Aalto et al. 2013). For the purpose of this work, length and curvature of centrelines and area and radius functions are treated as input data.

In this and the following chapters, the model is developed using four different vowel geometries from a healthy male: [ɑ, i, u] and [œ]. The first three were produced at fundamental frequency $f_0 = 110$ Hz and the last at $f_0 = 137.5$ Hz. These data are from the first pilot study where the simultaneous sound collection and MR imaging technology was tested (Aalto et al. 2011).

3.1.4 Subglottal tract

Anatomically, the SGT consists of the airways below the larynx: trachea, bronchi, bronchioles, alveolar ducts, alveolar sacs and alveoli. Such system can be modelled either with a tree-like structure (see e.g. Ho et al. 2011) or with a single tube with an area that increases towards the lungs (e.g. Birkholz et al. 2007; Lous et al. 1998). The tree model has previously been used with transmission line models but the theoretical framework for combining it with Webster's equation exists: Aalto and Malinen (2013) showed that Webster's equation on any finite graph can be written as an impedance passive, internally well-posed, strong boundary node in the sense of Malinen and Staffans (2007). The single horn model, on the other hand, is a crude simplification but it is good enough for modelling the first order effects, namely the impact of the first subglottal formant. It also has the benefit that the Webster's equation solver for the VT (Section 3.1.3) can easily be adapted for the SGT.

Following the modelling paradigm used in this work, the simpler approach is taken. The SGT is modelled as a tube of varying cross-sectional area and radius, $A_s(s)$ and $R_s(s)$, $s \in [0, L_{SGT}]$ where L_{SGT} is the nominal length of the SGT. Further, the tube is assumed to be straight, i.e. $\eta(s) = 0$, so that $W_s(s) = R_s(s)\sqrt{R_s'(s)^2 + 1}$. Eqs. (3.9)-(3.10) then translate to

$$\begin{cases} \frac{1}{c^2}\frac{\partial^2\widetilde{\psi}}{\partial t^2} + \frac{2\pi\alpha W_s(s)}{A_s(s)}\frac{\partial\widetilde{\psi}}{\partial t} - \frac{1}{A_s(s)}\frac{\partial}{\partial s}\left(A_s(s)\frac{\partial\widetilde{\psi}}{\partial s}\right) &= 0, \\ \frac{\partial\widetilde{\psi}}{\partial t}(L_{SGT}, t) + \theta_s c\frac{\partial\widetilde{\psi}}{\partial s}(L_{SGT}, t) &= 0, \\ \frac{\partial\widetilde{\psi}}{\partial s}(0, t) &= c_2 v_0(t), \end{cases}$$

where $\widetilde{\psi}$ is the velocity potential in the SGT, θ_s is the normalised acoustic resistance at the lung end of the tube and $c_2 = \frac{H_1 h}{A_s(0)}$.

When the subglottal resonator is included in the model, it interacts with the other subsystems through the counter pressure,

$$p_c = \rho\psi_t(0, t) - \rho c_3\widetilde{\psi}_t(0, t), \tag{3.11}$$

where c_3 accounts for the differences in the areas and moment arms for the supra- and subglottal pressures in the vocal fold force equations. Notice that since $c_3 \neq 1$,

p_c is not simply the transglottal pressure difference but, rather, the equivalent counter pressure that can be used after scaling in Eqs. (3.3)-(3.4).

The MRI data that is used to model the VT does not cover the SGT. Instead, the subglottal area function (3.12) is constructed using an exponential horn, the beginning and end areas ($A_s(0)$ and $A_s(L_{SGT})$) used by Birkholz et al. (2007), and choosing L_{SGT} so that the first subformant frequency is approximately 500 Hz.

$$A_s(s) = A_s(0)e^{\beta s}, \quad \beta = \frac{1}{L_{SGT}} ln\left(\frac{A_s(L_{SGT})}{A_s(0)}\right). \tag{3.12}$$

3.2 Simulation methods and parameters

3.2.1 Numerical methods

Solutions to the model described in the previous section are computed numerically using the methods of Aalto (2009) implemented in MATLAB R2013a (8.1.0.604). The equations of motion of the vocal folds (3.1) are solved using fourth order Runge-Kutta method, and when the glottis closes within a time step, polynomial interpolation is used to approximate the time of closure and the solution at that time. The glottal flow equation (3.5) is then solved using implicit Euler method.

Webster's equation is discretised spatially using finite element method (FEM) and temporally using Crank-Nicolson method to obtain update equations

$$\begin{cases} (\frac{\Delta t}{2}\mathbf{K} + \frac{2\rho}{\Delta t}\mathbf{M} + \rho\mathbf{R})\xi^n &= (-\frac{\Delta t}{2}\mathbf{K} + \frac{2\rho}{\Delta t}\mathbf{M} + \rho\mathbf{R})\xi^{n-1} + 2\mathbf{M}\mu^{n-1} + \Delta t\mathbf{b}(t_n), \\ \rho\xi^n - \frac{\Delta t}{2}\mu^n &= \rho\xi^{n-1} + \frac{\Delta t}{2}\mu^{n-1}, \end{cases} \tag{3.13}$$

where in the case of VT

$$\begin{aligned} \mathbf{M}_{ij} &= \tfrac{1}{2\rho c^2}\textstyle\int_0^{L_{VT}} v_i(s)v_j(s)\frac{A(s)}{\Sigma(s)^2}ds, \\ \mathbf{K}_{ij} &= \tfrac{1}{2}\textstyle\int_0^{L_{VT}} v_i'(s)v_j'(s)A(s)ds, \\ \mathbf{R}_{ij} &= \begin{cases} \frac{\pi\alpha}{\rho c^2}\int_0^{L_{VT}} v_i(s)v_j(s)\frac{W(s)}{\Sigma(s)^2}ds + \frac{A(L_{VT})}{2\theta\rho c}, & \text{when } i=j=N; \\ \frac{\pi\alpha}{\rho c^2}\int_0^{L_{VT}} v_i(s)v_j(s)\frac{W(s)}{\Sigma(s)^2}ds, & \text{otherwise}, \end{cases} \\ \mathbf{b}_j &= \begin{cases} \frac{H_1 h}{2}v_o(t), & \text{when } j=1; \\ 0, & \text{otherwise}, \end{cases} \end{aligned} \tag{3.14}$$

$\xi^n \approx \xi(t_n)$ and $\mu^n \approx \mu(t_n)$, and ξ and μ contain the coefficients of piece-wise linear basis functions $v_j(s)$, $j = 1, ..., N+1$ on each element of the VT in the approximate solution

$$\begin{bmatrix} \psi \\ \rho\psi_t \end{bmatrix} \approx \sum_{j=1}^{N+1}\left(\xi_j(t)\begin{bmatrix} v_j(s) \\ 0 \end{bmatrix} + \mu_j(t)\begin{bmatrix} 0 \\ v_j(s) \end{bmatrix}\right).$$

Further details on the derivation can be found in Appendix A.

In the case of the SGT, the mass, stiffness, and damping matrices are also calculated using Eq. (3.14) but replacing $A(s)$ with $A_s(s)$, $W(s)$ with $W_s(s)$, θ with θ_s, L_{VT} with L_{SGT}, and setting $\Sigma(s) = 1$. Eqs. (3.13) and (3.14) are valid for any VT (or SGT) geometry for which we can determine $A(s)$, $W(s)$ and $\Sigma(s)$ for $s \in [0, L_{VT}]$, where L_{VT} depends on the VT configuration as well. In fact, the integrals in Eq. (3.14) are evaluated numerically so it suffices to have $A(s)$, $W(s)$ and $\Sigma(s)$ defined at the discretisation points only. This might help to smooth out problem regions in some VT geometries.

3.2.2 Parameters

Input parameters for each simulation can be divided into three categories: (i) constants, i.e. parameters which take the same numerical value in all simulations, (ii) tuning parameters, i.e. parameters which take a constant predetermined value for each simulation, but the value may change between simulations, and (iii) derived parameters, i.e. parameters which depend on (i) and/or (ii) and hence must be determined for each simulation separately.

Justification for the choice of tuning parameters and the methods for choosing their values is left for the next chapter. Here, the four tuning parameters, the first eigenfrequency of the vocal folds, f_1, vocal fold length, h, inertial parameter in the glottal flow equation, C_{iner}, and subglottal pressure, p_{sub}, are taken as known constants. Also, for the purpose of this chapter, VT configuration is a special type of tuning parameter.

It is also convenient to parametrise the simulation output, namely glottal flow pulses, for the ease of both target definition and comparison. The output parameters used in this work are defined below.

Constants

Table 3.1 lists the numerical values of physiological and physical constants used. Based on these constants, VT loss coefficient, α, is approximated as

$$\alpha = \frac{\rho}{\rho_h c_h} \approx 7.6 \cdot 10^{-7} s/m, \tag{3.15}$$

and the viscous loss coefficient in the glottis, C_g, as

$$C_g = 12\mu H_1 L_g = 1.08 \cdot 10^{-20} Ns,$$

which approximates the pressure loss in the glottis using a rectangular tube of width h, length L_g, and height ΔW_1, and $\Delta W_1 << h$ (Aalto 2009).

For this work, the SGT lung termination coefficient, θ_s, is taken to be a constant which corresponds to an absorbing boundary condition.

Table 3.1: Constant physical and physiological parameters

Parameter	Symbol	Value
speed of sound in air	c	343 m/s
speed of sound in soft tissue	c_h	1540 m/s
kinematic viscosity of air	μ	18.27 μN s/m^2
density of air	ρ	1.2 kg/m^3
vocal fold tissue density	ρ_h	1020 kg/m^3
spring constant in contact	k_H	730 N/m [1]
location of superior vocal fold spring	l_1	0.15 [2]
location of inferior vocal fold spring	l_2	0.85 [2]
glottal gap at rest	g	0.3 mm
vocal fold shape parameter	a_1	1.858 [1]
vocal fold shape parameter	a_2	-319.722 m^{-1} [1]
vocal fold element length	L	6.8 mm [1]
control area height above glottis	H_1	1 mm
equivalent gap length for viscous loss	L_g	1.5 mm
SGT length	L_{SGT}	220 mm
normalised acoustic resistance at lungs	θ_s	1
glottal entrance/exit coefficient	k_g	0.175

Table 3.2: Simulation parameters

Parameter	Value
default time step	$2 \cdot 10^{-6}$ s
lower limit for glottal damping coeff.	10^{-4}
upper limit for glottal damping coeff.	1
feedback scaling (Q_{pc})	0.05
min. pulse amplitude in stable flow (C_b)	$5 \cdot 10^{-5}$ m^3/s
max. amplitude fluctuation in stable flow (ϵ_b)	0.01

Parameters for numerical methods can be found in Table 3.2. These have been chosen based on trial and error to give a stable output for long enough duration to determine output parameters reliably without making computational expense

[1] Horáček et al. (2005)
[2] Aalto (2009)

of each simulation too high. This compromise is important particularly during tuning when both unstable and needlessly long simulations may increase the time needed to find a suitable tuning parameter combination by hours.

In addition to the listed parameters, the number of area slices in the VT discretisation, N, is also needed. N affects the resonance structure of Webster's equation, so its value was chosen based on the resonances of the Webster's equation from the generalised matrix eigenvalue problem

$$\mathbf{K}\mu_\lambda = \lambda^2 \rho \mathbf{M} \mu_\lambda.$$

The first three of these resonances, λ_k, $k = 1, 2, 3$, for each of the four vowel configurations were used to compute the errors

$$\sqrt{\sum_{k=1}^{3} \frac{\left(\lambda_k - \tilde{\lambda}_k\right)^2}{\tilde{\lambda}_k^2}}, \tag{3.16}$$

where the reference resonances, $\tilde{\lambda}_k$, are solutions to the corresponding Helmholtz problem whose values can be found in Aalto et al. (2012).

The optimal value of N depends on the VT configuration. However, for simplicity and speed, $N = 28$ is used for all VT configurations. This is the value which minimises the sum of the relative errors (3.16) for the four VT configurations (Figure 3.2).

Derived parameters

Once VT configuration and other tuning parameters have been chosen, estimates for the following parameters are calculated:

1. Nominal value for the inertia parameter, C_{iner}, before tuning

$$C_{iner}^0 = \rho \int_0^{L_{VT}} \frac{ds}{A(s)} \tag{3.17}$$

   based on the estimation made by Aalto (2009) assuming incompressibility.

2. Normalised acoustic resistance at mouth

$$\theta = \frac{2\pi f^2 R(L_{VT})^2}{c^2},$$

   where $R(L_{VT})$ is the VT radius at mouth and f is a tuning frequency taken to be 2000 Hz. This estimate is based on a circular piston in an infinite baffle and often used to estimate radiation at the mouth (e.g. Aalto 2009; Flanagan 1972; Morse and Ingard 1968).

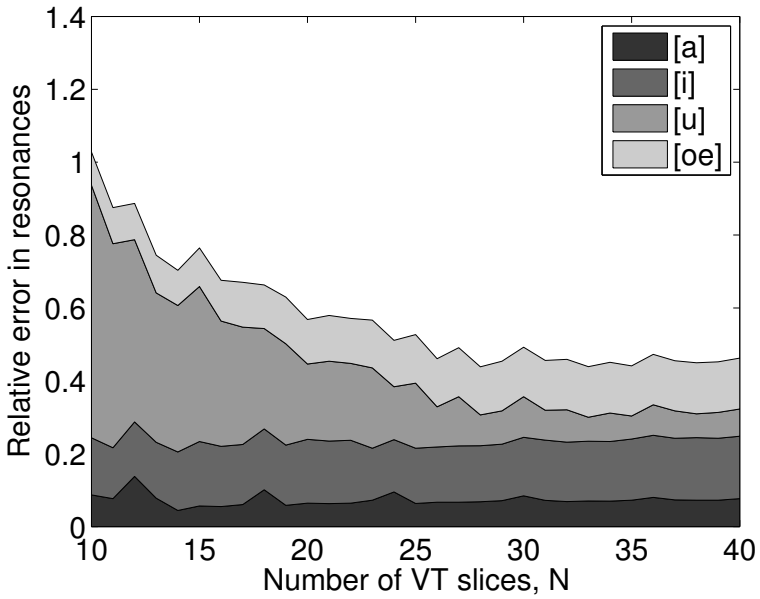


Figure 3.2: Relative error in resonances (Eq. 3.16) stacked for four vowel geometries as a function of number of VT area slices

3. Vocal fold masses are solved by matching the equivalent system (Figure 3.1) to a more realistic vocal fold geometry in terms of total mass (M), centre of gravity (T), and moment of inertia (I). In other words, we solve

$$\begin{cases} m_1 + m_2 + m_3 & = M = \quad h\rho \int_0^L a(x)dx, \\ \frac{L}{2}m_3 + Lm_1 & = T = \quad h\rho \int_0^L xa(x)dx, \\ \left(\frac{L}{2}\right)^2 m_3 + L^2 m_1 & = I = \quad h\rho \int_0^L x^2 a(x)dx, \end{cases}$$

where $a(x) = a_1 x + (a_2/2)x^2$ is the approximate shape function of the more realistic vocal folds (Horáček et al. 2005), and the rest of the symbols may be found in Section 3.1.1 and in Figure 3.1. Notice that the masses are derived parameters because they depend on h, which is a tuning parameter.

4. Vocal fold spring constants are solved by setting the two eigenfrequecies of the vocal folds to f_1 and $f_2 = 1.05f_1$, and solving the reverse eigenvalue problem

$$\begin{cases} r(2\pi i f_1) = 0, \\ r(2\pi i f_2) = 0, \end{cases}$$

where $r(s) = det(s^2 M + K)$, with M and K the mass and stiffness matrices defined in Eq. (3.2).

5. The damping coefficients in the glottis model, b_1 and b_2, are determined automatically using a search procedure. For simplicity, it is assumed that all the dampers in the vocal fold model are identical, i.e. $b_1 = b_2 = b$. The critical value of this damping coefficient is determined for each simulation separately using Golden Section search and looking for b such that if the total number of glottal cycles in a simulation is I_g,

$$\Delta U^i(b) > C_b \quad \text{and} \quad \frac{|\Delta U^i(b) - \Delta U^{i-1}(b)|}{\Delta U^{i-1}(b)} < \epsilon_b, \quad \text{for } i = I_g - k, ..., I_g, \tag{3.18}$$

where ΔU^i is the amplitude of the flow pulse in the i^{th} cycle (i.e. maximum flow minus minimum flow). k is chosen to give a sufficient number of glottal pulses over which to average phonation parameters, and C_b and ϵ_b are experimentally determined (positive) constants, which help to reliably determine critical damping without increasing the time of finding b excessively. For this same end, a simulation is terminated as overdamped if f_0 appears to be below 50 Hz, if the flow amplitude falls below C_b, or if the flow amplitude decreases by more than $\epsilon_b \cdot 100\%$ for three consecutive cycles. A simulation is terminated as underdamped if the flow amplitude increases by more than $\epsilon_b \cdot 100\%$ for three consecutive cycles.

Output parameters

The output of interest from each vowel synthesis simulation is the glottal flow pulse. Two parameters are used to characterise this pulse. First, as mentioned, is the fundamental frequency, f_0, which is calculated from a train of flow pulses as the inverse of the time difference between subsequent peaks, i.e. the fundamental period, T_0.

The second parameter is the closing quotient, ClQ, which is the proportion of the closing phase in the glottal cycle. ClQ is a time-domain ratio to parameterise glottal flow, and it has been used, for example, in glottal inverse filtering analyses of phonation types (Alku et al. 2002; Lehto et al. 2007). As phonation changes from pressed to normal to breathy, ClQ increases reflecting increasing symmetry in the pulses. Throughout this work, the phonation types reported for simulated pulse shapes are based on their ClQ-values. Whether ClQ alone is adequate for phonation type identification with this model is discussed in the next chapter.

As a back-up measure, a third parameter is also calculated. The open quotient, OQ, is the proportion of time the glottis is open in one glottal cycle. It is also a shape parameter for the glottal pulse, and hence characterises phonation type: the higher the OQ, the more breathy the phonation.

For the i^{th} glottal cycle these three parameters are defined as

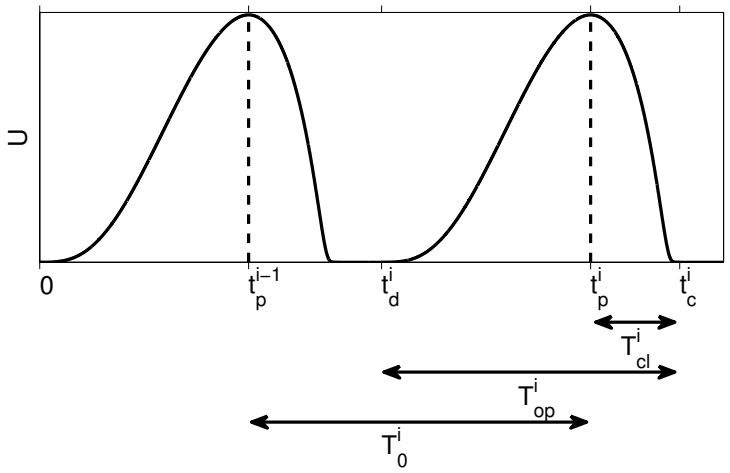


Figure 3.3: Parametrisation of output pulses

$$\begin{aligned} f_0^i &= \frac{1}{t_p^i - t_p^{i-1}} = \frac{1}{T_0^i}, \\ ClQ^i &= \frac{t_c^i - t_p^i}{T_0^i} = \frac{T_{cl}^i}{T_0^i}, \qquad \text{and} \\ OQ^i &= \frac{t_c^i - t_d^i}{T_0^i} = \frac{T_{op}^i}{T_0^i}, \end{aligned}$$

where t_d is the time of glottal opening, t_c the time of glottal closure (or minimum flow if the glottis does not close), and t_p the time of the flow peak as shown in Figure 3.3.

Unlike inverse filtering results, simulation results do not suffer from noise, so extraction of t_d, t_c, and t_p is straightforward. However, conditions (3.18) guarantee only approximately critical damping for any $\epsilon_b > 0$, and hence the time parameters vary between glottal cycles within each simulation. To minimise the error caused by this, f_0, ClQ, and OQ are averaged over glottal pulses which have been determined to be critically damped according to conditions (3.18).

3.3 Simulation results

The changes and additions described above were added to the model one at a time to investigate how they affect the model behaviour, in particular the output parameters f_0 and ClQ. First, the purely technical addition of automatic damping coefficient determination is tested. This addition is then used in all following simulations.

Attention is then turned to answering the question of whether tissue losses in the VT, turbulence losses in the glottis, or the SGT are necessary parts of a

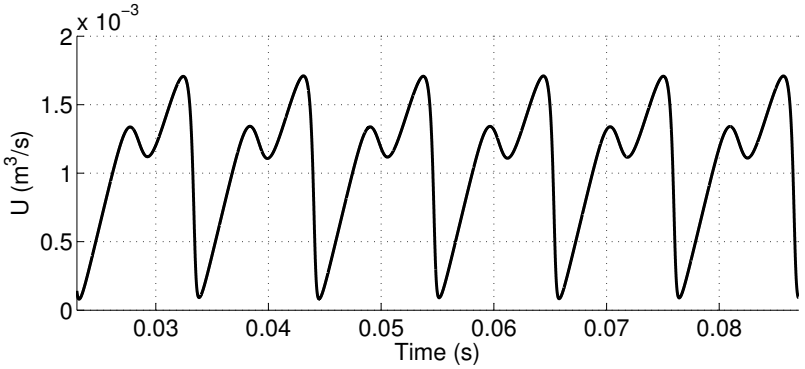


Figure 3.4: Finding critical b fails if the pulse form has a ripple, i.e. a local maximum typically on the rising edge. The shown flow is for [ɑ] with $f_1 \approx 218$ Hz, $h \approx 29$ mm, $Q_c \approx 7.93$ and $p_{sub} \approx 1540$ Pa.

minimal model. In other words, the aim is to determine whether the improved faithfulness (i f to reality is worth the computational costs (time, stability, etc.).

These latter investigations make use of the tuning method development and testing results described in the next chapter, particularly Section 4.3.2. Essentially, these results consist of 20 sets of tuning parameter combinations with each set corresponding to a vowel-f_0-combination, hereafter denoted ([v],f_0). For the sets in question, [v] is one of the four vowel geometries [ɑ, i, u, œ] and f_0 is one of five levels: 100 Hz, 130 Hz, 160 Hz, 190 Hz, or 220 Hz. The number of parameter combinations in the sets vary from 61 to 155.

3.3.1 Damping coefficient, b

Automatic damping coefficient determination detailed in Section 3.2.2 was added to the model to make it possible to handle a large number of simulations. The method was tested and suitable tolerance constants C_b and ϵ_b in Eq (3.18) were determined (see Table 3.2). Five different tuning parameter combinations were used for each of the four vowel geometries.

One search for the critical b took on average 33 s (using a server with Ubuntu 12.04, Intel Xeon X5650 CPU @ 2.67 GHz with 12 cores, and 53 GB RAM). If the search for b was successful, the average time was 28 s, whereas unsuccessful searches took on average 53 s. Out of the 20 searches, four did not find an approximate critical b.

The failures were caused by four different reasons which illustrate the limitations of this method for determining b:

1. Tuning parameter combination is outside the range which produces sustained oscillations. This is not, in fact, a failure of the method as no critical b exists.

2. Sustained oscillations happen with amplitude below C_b and hence are mislabelled overdamped. Decreasing the value of C_b reduces mislabelling but increases the time it takes to detect overdamping.

3. The range of b values that satisfy conditions (3.18) is so narrow it is difficult to find due to limited mesh tolerance in the Golden Section search (in these simulations $b_{up} - b_{low} \geq 0.001$). Decreasing the tolerance limit helps in this case but it also increases the number of futile simulations in the previous two cases.

4. Subharmonic frequencies are present in the flow signal which typically causes multiple local maxima in the pulses in sustained oscillations. This covers a continuum of cases from pulses with a ripple on the rising edge (see e.g. Figure 3.4) to signals where every other pulse has smaller amplitude.

   It may be possible to find a b-value corresponding to the sustained oscillations in these cases but automatic extraction of f_0 and ClQ fails because consecutive stable peaks cannot be identified. These cases typically occur when high f_1, which drives f_0 up, is used together with high values for some or all other tuning parameters, which drive f_0 down. If f_0 is measured using global maxima, it can often be observed that $f_0 \approx \frac{1}{2} f_1$.

Despite these failure modes, b can be found fairly reliably and quickly when the model is producing basic pulse shapes. Since this is what the model is mainly required to do, the method can be considered to work well. For the rest of this work, any failure to find the critical b is taken to be indicative of being outside the parameter range for sustained oscillations, regardless of the reason for the failure.

3.3.2 Vocal tract loss coefficient, α

Dissipation along the VT was added to original model through the use of a boundary loss coefficient, α, whose value increases as boundary dissipation increases. This model of losses includes only tissue losses, ignoring losses caused by viscosity, heat conduction, and abrupt changes in the VT radius. Because the real system is therefore expected to be more lossy than the model, the sensitivity of the glottal flow pulse to the value of α was investigated.

A nominal value, α_0, given by Eq. (3.15), was introduced and $\alpha = Q_\alpha \alpha_0$ was used in simulations. Five tuning parameter combinations were selected randomly out of each of the 20 ([v],f_0) sets, giving a total of 100 simulations settings. Q_α was varied in the range $[0, 10^{15}]$, which covers effects from no dissipation to virtually no phonation-like output at the mouth, and hence is expected to contain the value of α that best represents all losses in the VT. Note that $Q_\alpha = 1$ corresponds to

the physically motivated value given by Eq. (3.15). Figures 3.5a and 3.5b show some typical results as relative changes compared to the lossless case, i.e. $Q_\alpha = 0$.

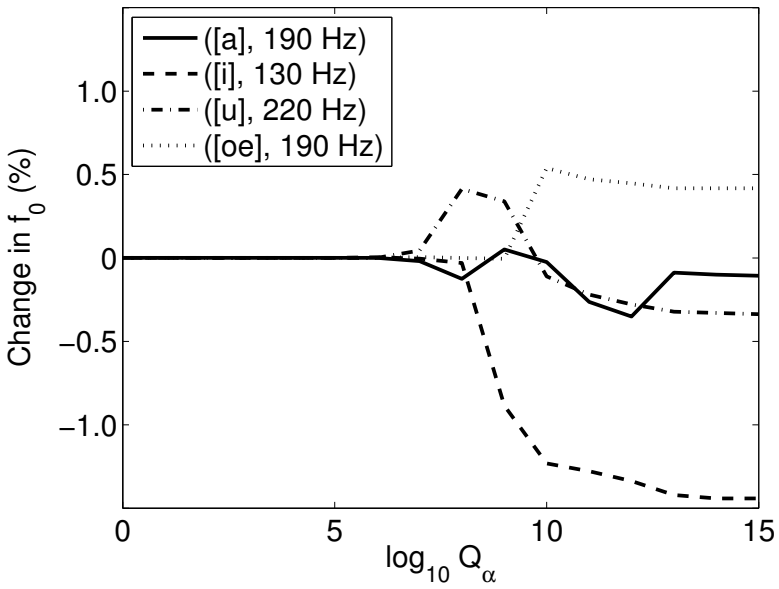
Three observations can be made about the impact of varying Q_α. First, neither f_0 nor ClQ change noticeably until tissue losses increase above a threshold level. This threshold is typically $Q_\alpha \approx 10^7$, and in all simulations it was in the range $[10^6, 10^9]$.

Second, increasing the loss coefficient beyond the threshold level causes generally a small change in both f_0 and ClQ. The direction of this change is most often negative, i.e. f_0 falls and phonation becomes more pressed. More rarely, the output parameters may increase, increase (or decrease) initially before decreasing (increasing), or they may fluctuate around the original level. Regardless of the direction of the change, it is at most 3% for f_0 and 10% for ClQ. These effects are comparable in magnitude to the inaccuracies caused by and handled in tuning the model (see Chapter 4).

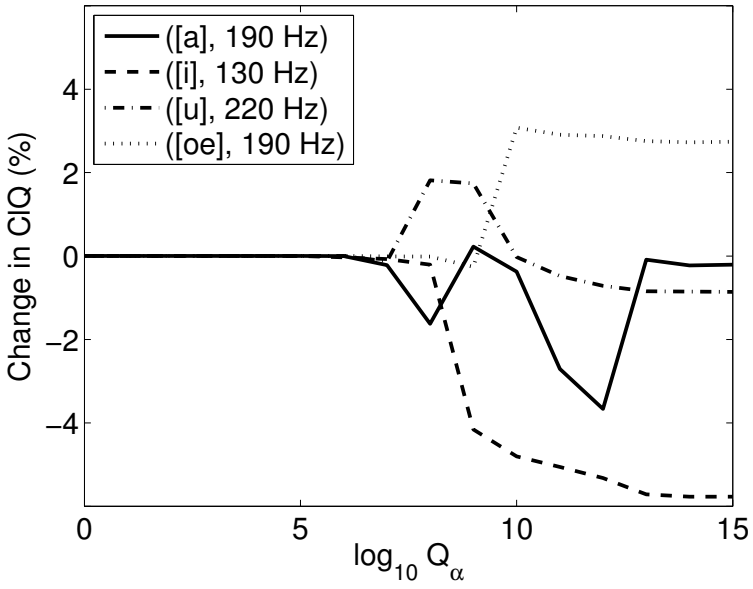
Thirdly, the second observation above is true in most but not in all cases. Sometimes significantly larger changes are observed when Q_α is increased beyond the threshold level. This effect can be up to 400% in magnitude and include absence of quasi-stable phonation at some values of Q_α. These cases were observed at four out of five f_0-values with [i] and [u] and at two f_0-values with [ɑ]. At all of these vowel-f_0-combinations, other parameter combinations produced only small changes.

The first two observations indicate that the glottal pulses produced by the model are generally not sensitive to the value of the tissue loss coefficient α. It seems reasonable to assume that the estimate given by Eq. (3.15) is accurate to at least six orders of magnitude, and hence the impact of adding tissue losses to the model is negligible. Of course, this is true only in regard to glottal pulses. Other signals, particularly pressures and velocities along the VT, are expected to be sensitive to the value of α.

The third observation is made when losses in the vocal tract are higher than would be expected under normal conditions. Nevertheless, it gives evidence of complex interaction between the vocal tract and the vocal fold models. Increasing α decreases vocal tract resonances, and simulations with a model similar to the one used in this work have indicated that interaction between f_0 and the lowest VT resonances can cause jumps in f_0 (Aalto et al. 2012; Titze 2008). However, the fact that the model can produce both typical and atypical behaviour at some vowel-f_0-combinations suggests that one or more of the tuning parameter values play an important role in producing large changes in f_0 and phonation type.



(a)



(b)

Figure 3.5: Some examples of the change in (a) f_0 and (b) ClQ compared to the values in the lossless case ($Q_\alpha = 0$). ([i],130 Hz) illustrates the most common case. In the other examples the direction of the change is less common, but its magnitude is still small.

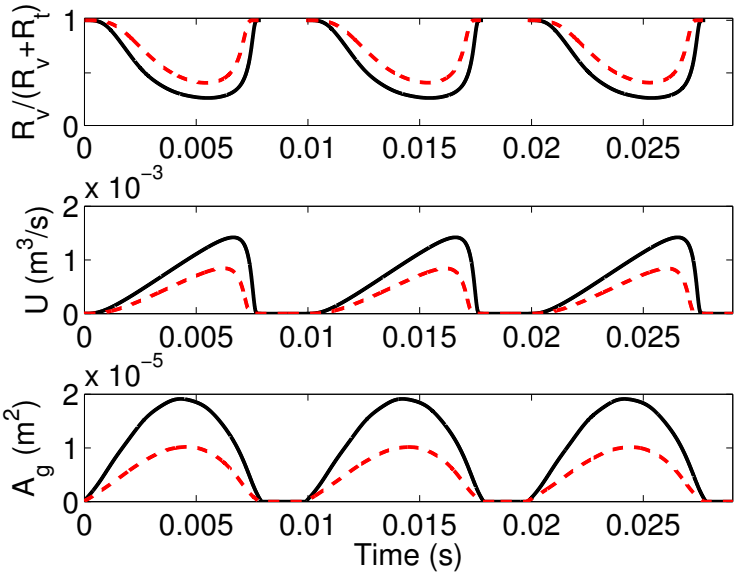


Figure 3.6: Ratio of viscous resistance (R_v) to viscous and turbulence resistance (R_v+R_t) when the model includes only viscous losses using two different parameter settings from ([ɑ], 100 Hz). The middle and lower figures show volume flow (U) and glottal area (A_g), respectively, for the same simulations.

3.3.3 Losses in the glottis

In its basic form model of glottal flow Eq. (3.5) includes only viscous losses, i.e. $R_g = R_v$ (see Eq. (3.6)). However, at large glottal opening values losses caused by turbulence become significant (van den Berg et al. 1957). Therefore it is relevant to ask if adding a turbulence loss term, i.e. using $R_g = R_v + R_t$ (see Eq. (3.7)), would improve the model performance and possibly reduce reliance on the experimentally determined damping coefficient, b.

First, the value for k_g, which scales the turbulence loss term in Eq. (3.7) was determined by trial and error. van den Berg et al. (1957) used $k_g = 0.875$ but this causes glottal pulses to be overdamped or very breathy in our model when $R_g = R_v + R_t$, indicating that the value is too large. Hence, k_g was decreased until the model remained stable and able to produce a variety of phonation types. The value used in the following simulations is $k_g = 0.175$.

In Figure 3.6, the viscous resistance, R_v, is compared with the sum of viscous and turbulence resistances when $R_g = R_v$. Both of the two simulations shown use tuning parameters from ([ɑ], 100 Hz) and they have ClQ-values corresponding to pressed phonation. The impact of different tuning parameter values can be seen in the maximum amplitudes of the glottal area and volume flow pulses.

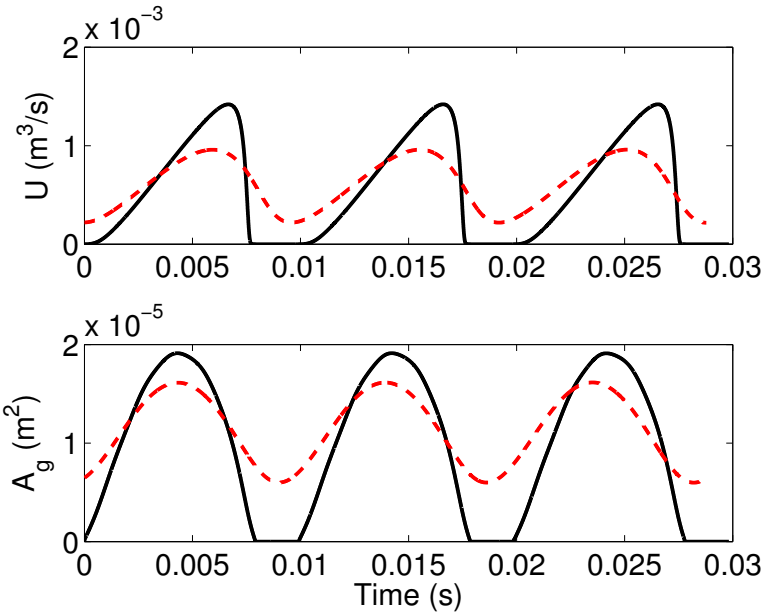


Figure 3.7: Volume flow (U) and glottal area (A_g) when model includes only viscous effects, $R_g = R_v$ (solid black line) and when model includes viscosity and turbulence, $R_g = R_v + R_t$ (red dashed line). Parameters are from ([ɑ], 100 Hz)

Figure 3.6 shows that the contribution of the loss types varies over the glottal cycle as expected based on Eq. (3.7): viscous losses dominate when the glottis is just opening or almost closed whereas losses caused by turbulence are particularly significant near maximum glottal opening and flow. Although the impact of turbulence decreases in simulations with smaller amplitudes of glottal flow and area, losses due to turbulence do not become insignificant, particularly at peak flow and opening. This same observation was made with other vowels, f_0-values and phonation types.

Next, the impact of adding turbulence losses into the model was explored by changing the model resistance to $R_g = R_v + R_t$. Five tuning parameter combinations where selected randomly out of each of the ([v],f_0) sets.

Out of these 100 simulations, 21 do not produce quasi-stable phonation after turbulence losses are added. [i] is particularly unstable with five failures with parameters from ([i], 220 Hz), four from ([i], 190 Hz) and one from ([i], 130 Hz). The rest of the failures are distributed more evenly among vowel-f_0-combinations. No clear relationship between original ClQ-value and failure is visible. All of the failed simulations exhibit the same mode of failure: initial oscillations, if any, are damped down until a steady state is reached with the glottis open to a fixed degree or oscillations are smaller than the threshold amplitude, C_b, in Eq. (3.18).

The change in f_0 in stable simulations varies in the range $[-20\%, 20\%]$, with no clear trend towards either increasing or decreasing f_0. In some cases, such as the example in Figure 3.7, there is virtually no change in f_0. ClQ increases in all stable simulations and the change varies from less than 1% to 400%.

The changes in ClQ are smallest when the phonation is breathy to start with. Phonation that is pressed or normal without the turbulence term, tends to change the predominant phonation type: normal to breathy and pressed to either normal or, as in Figure 3.7, to breathy. Full closure of the glottis during the minimum flow phase also becomes rare when turbulence is added. This can be explained by the removal of kinetic energy at the glottis to turbulence. Thus, there may not be enough energy left to be stored in the vocal folds to ensure closure or, in some cases, sustained oscillations.

Including turbulence in the model also causes a decrease in the critical damping constant, b, in stable simulations, except in three cases. In seven cases b decreases almost 100%, eliminating the need for damping in the vocal folds nearly completely. Without turbulence, b is virtually nonexistent only in one out of the 100 cases. This case remains stable with the same b when turbulence is added.

3.3.4 Subglottal tract

The final change to the model to be investigated is the addition of a SGT resonator. Five parameter combinations selected randomly from each of the ([v],f_0) sets, giving a total of 100 simulation settings, were again used.

The SGT resonator causes changes in both f_0 and ClQ. The changes in f_0 are typically less than 10% and towards increasing f_0, although decreasing f_0 is not rare and larger increases up to 20% were noted in a few cases. Likewise, ClQ shows a tendency to increase, i.e. move toward more breathy phonation, but decreasing ClQ was also observed. The changes were typically less than 20% but again, larger responses up to 60% were seen.

Figure 3.8 shows an example of model output time-series for a typical case with and without the subglottal resonator connected to the model. Parameters from ([ɑ], 100 Hz) were used for this figure and the original phonation type was pressed. In the shown case f_0 increases by 4% and phonation type remains pressed (ClQ increases from 0.15 to 0.20).

A few other observations can be made about the behaviour of the model when the SGT resonator is connected. In 15 cases, the model does not produce quasi-stable phonation. Failures caused by both over- and underdamping of the vocal fold oscillations are present.

In some cases, the glottal area pulses become distorted. With SGT the area pulses can have multiple local maxima or they can be angular near the peaks. These distortions decrease when simulations are run longer, indicating that the

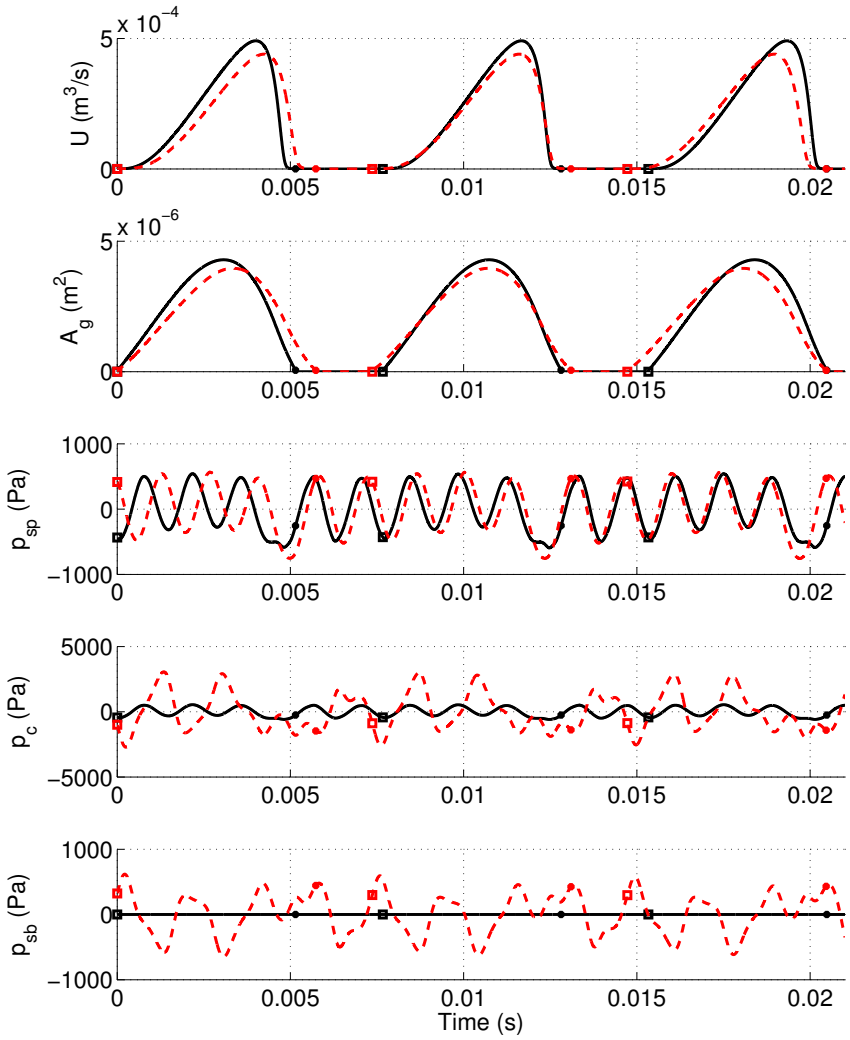


Figure 3.8: Volume flow (U), glottal area (A_g), supraglottal pressure (p_{sp}) just superior to the vocal folds, counter pressure (p_c), and subglottal pressure just inferior to the vocal folds (p_{sb}) without SGT (solid black line) and with SGT (dashed red line) for parameters from ([ɑ], 130 Hz). Glottal closure is indicated by a dot and glottal opening by square.

problem is caused by a transient response in the system. The effect this response has on the glottal volume flow is not as easily detected. In fact, from the point of view of determining the value of b, these simulations appear stable.

Slowly decaying transient responses are also seen in the pressure signals, especially in the subglottal pressure just inferior to the vocal folds and the equivalent counter pressure that depends on it. The severity of the transient response depends on the vowel configuration, tuning parameter values, the termination condition, θ_2, and the level of feedback, Q_{pc}. It can take up to 20 times longer to reach steady state when SGT is connected.

Previous studies have noted that glottal closure occurs at or is immediately followed by a peak in the subglottal pressure - the so called "water hammer" effect. If the glottis remains closed long enough, the first damped echo of this pressure peak can also be seen during the closed phase (Ho et al. 2011). These effects can be seen in Figure 3.8, although the echo pulse appears very damped. The shown case is one of the few cases where these effects are visible, however, because in most cases adding the SGT resonator to the model prevents the glottis from closing fully, possibly due to too strong coupling from SGT. When this happens, minima in the glottal area pulses occur close to a peak in the subglottal pressure, but the temporal order of these events is not constant.

3.4 Discussion on the model

In this chapter, a vowel synthesis model that can be used as a glottal pulse generator has been described. The model by Aalto (2009) has been taken as a starting point and the effect of making additions to the model have been investigated with the aim of finding a minimal model that is able to produce the essential features of glottal pulses. An automatic procedure to find the glottal damping coefficient has been developed, tissue and turbulence losses have been added to the VT and glottal flow models, respectively, and a whole new subsystem representing the SGT has been introduced.

3.4.1 Glottal damping

The procedure to find the glottal damping coefficient, b, was found to work sufficiently well for the application considered. The method is unfortunately somewhat slow, leading to long computation times when a large number of simulations is needed, such as in optimisation routines discussed in the next chapter.

Golden Section search was selected as the search algorithm as it is an efficient line search algorithm that does not require the evaluation of derivatives. It also has the benefit that it is sufficient to know whether the system is over- or underdamped

at the point of evaluation; the degree of over- or underdamping is of no interest. On the other hand, if this degree information could be reliably estimated, then using, for example, interpolation search could improve the speed of convergence.

Another issue with Golden Section search is that it requires predetermined lower and upper limit for the parameter. Constant limits were used for this work, meaning that the algorithm always started at the same b-value. By the end of this work, a large data set exists that could be used to investigate which factors affect b. An intelligent guess as a starting point could greatly affect the performance of the algorithm.

The cases in which the search for the damping coefficient fails have been listed in Section 3.2. While the failures were judged to be acceptable for this application, it pays to keep in mind that this same vowel synthesis model could be used for purposes other than as a source signal generator for a wave equation solver. If the model is, for example, used to investigate phenomena in which ripples in the pulse form are of interest, then the method for finding the critical b-value needs to be redesigned.

It should also be noted that the damping coefficient used in the model of this work serves two functions. Firstly, it represents damping in the vocal fold tissue, which is a physiological phenomenon that needs to be included in the model. Secondly, the coefficient is used to produce approximately quasi-stable phonation in simulations. This indicates that either the model is missing some essential stabilising element or that we are artificially stabilising a system that is not inherently robustly stable. These stability issues are discussed further below. In any case, estimating the magnitude of vocal fold tissue damping and comparing it with the damping found numerically would give valuable insight into the model.

3.4.2 Vocal tract losses

VT tissue losses were added to the model and represented via a boundary loss coefficient, α. The model was found not to be very sensitive to this addition. In particular, at reasonable values of α, the impact of tissue losses on glottal pulses was found to be negligible. However, if the model is used generate glottal flow pulses for a 3D acoustic simulator, boundary dissipation should be comparable in the two models.

In contrast to glottal area and flow, changes to the pressure at mouth were significant. At high values of α the resonances of the VT were suppressed and only the harmonics of the fundamental frequency were visible in the spectrum. Other pressure and velocity signals along the VT are also affected. Therefore, care must be taken in applications where these signals are of interest.

It was observed that at very high values of α, some parameter combinations produce very large changes in either f_0 or ClQ or both. In a few cases at high values

of α, absence of quasi-stable phonation was also observed. While these observations do not directly affect use of the model as glottal source generator under normal conditions, they reveal interesting interaction between tuning parameter values and the vocal tract.

As was mentioned earlier, the coefficient α only accounts for one out of several types of losses present in the VT. Losses due to viscosity, heat conduction at the VT walls, or abrupt changes in the cross sectional area of the VT have been left out of the model. Out of these, particularly viscous losses in narrow VT sections could prove to be noticeable.

3.4.3 Losses in the glottis

The current glottal flow model was constructed to include viscous losses only. To investigate the sensitivity of the model output to this assumption, a rough estimate for the losses caused by turbulence in the glottis was obtained and compared to the model losses. This estimated loss was then added to the flow model, and the impact of this addition was investigated.

The observations made were similar to previous studies: turbulence is significant at large glottal openings and flows (van den Berg et al. 1957). The model hence underestimates losses for a part of the glottal cycle, leading to pulse shapes which are notably more pressed and have different f_0 than pulses obtained with turbulence included in the model.

It was also observed that in a significant number of cases adding turbulence resulted in a failure of the model to produce quasi-stable phonation. There are two factors that contribute to this. First, kinetic energy is lost to turbulence and hence there may not be enough left to cause sustained oscillations in the vocal folds without altering tuning parameter values. And second, in the turbulence-free model, only the relative values of the subglottal pressure and the flow parameters C_{iner} and C_g matter, and hence an accurate estimate of the viscous loss parameter, C_g, is not necessary. When turbulence is added, this is no longer true. The disappearance of quasi-stable phonation may therefore be caused overestimation of either turbulence or viscous losses, or both.

One question left open by these simulations is whether including turbulence in the model causes changes in the pulse shapes that are not reflected in f_0 and ClQ. Indeed, this is an essential question for future work. The findings indicate that future work is needed on refining the turbulence resistance estimate. However, such work brings little added value to the model if the pulses produced by the turbulence-augmented model are to a large extent similar to pulses of matching f_0 and ClQ produced by the original model.

3.4.4 Subglottal tract

The final addition to the model considered was a SGT resonator. It was implemented as a simple expanding horn in parallel with the VT. The length of the horn was selected to produce the first subglottal formant at 500 Hz. The simulation results produced were in line with existing literature as long as f_0 was low and the glottis closed fully. For example, subglottal pressure peaks were observed at or immediately following glottal closures and the first echo of this peak was seen during the closed phase. The SGT-augmented model produced glottal pulses with f_0 and pulse shapes differing from the original model. These differences varied from insignificant to very clear.

The question of interest is, of course, whether the SGT is a necessary or desirable addition to the vowel synthesis model. The answer depends on what the model is used for, as usual. Glottal flow pulses are fairly non-sensitive to the SGT, particularly in applications where the need for computational speed and stability require the feedback level to be low. The transglottal pressure and other pressure measures, on the other hand, reflect these changes more strongly.

In any case, if the SGT resonator is used, methods for stabilising it are needed. Reducing the feedback level is one option but it has the unfortunate effect of limiting the extent to which the SGT can affect the rest of the model and removing phenomena such as coincidence of subglottal pressure peaks and glottal closure. Increased losses along the SGT or at its termination would also have a stabilising influence and should be carefully considered. A more complex model with subdivision and sufficiently absorbing boundaries might also perform better. If the final subdivisions get very narrow, however, viscous effects become significant and need to be added to the model as well.

Initial conditions play a key role in reducing transient responses. The difficulty in selecting appropriate initial pressure and velocity distributions arises from the fact that these depend not only on the VT configuration but also on tuning parameter values. Since slow transient responses are a problem in the SGT, it might be worth investigating non-zero initial pressures and velocities in the SGT only.

3.4.5 The vocal tract model

Having investigated the impact of adding the SGT to the model, it is natural to ask whether side branches to the VT should also be considered. In particular, the nasal tract is essential if we wish to extend the model to VT geometries corresponding to nasals, e.g. [m,n]. The velum may also be open during vowel production coupling the nasal tract to the VT. If the model is to be used, say, with patient data where this cannot be ruled out, then including the nasal tract may be necessary to produce all essential features of the glottal pulses.

Another point to note is that approximating the VT as a tube works better in some regions and in some vowel geometries than in others. For example, the oral cavity for [i] is fairly tubular to begin with whereas the position of the tongue for [u] produces very non-convex area slices making the fit much worse. Another example is the region of the epiglottis, where discretisation plays a key role in determining the complexity of the area slice geometries, and hence quality of the tube fitting. This is an inherent problem with using a simple model, such as Webster's equation, and explains for its part discrepancies between model output and reality.

Despite the simplicity of the model, literature and computations show that Webster's equation in the vocal tract gives practically the same first three resonances as the Helmholtz model. However, at the fourth resonance, sinuses start to have dynamics, and as a result the models differ significantly.

3.4.6 Stability of phonation in the model

When each of the additions to the model was investigated, it was noted that in some cases the additions caused the model to stop producing quasi-stable phonation. Particularly in the case of turbulence losses and the SGT, the number of cases where phonation disappeared was not insignificant. There are two issues to consider regarding this apparent increase in instability.

On one hand, it cannot be concluded without further investigation whether the more extensive models are less able to produce quasi-stable phonation or if the effect is merely caused by the way the simulations were done. The results were computed by comparing model output with and without the additional elements and holding everything else constant. The simulations without the additions were always done first so that the tuning parameter values used in both simulations were selected from the original parameter space. Furthermore, the values of some of the constants were also selected based on the behaviour of the original model. A fixed value of the glottal gap at rest, for example, might be one of the reasons why the glottis does no close fully in many of the turbulence simulations.

Changing the model changes the parameter space causing the boundaries of the phonation-producing set to shift and leaving some points which were previously in the set outside it. The clearest examples of this are seen when addition of turbulence losses in the glottis damps out all glottal oscillations because the the driving force (i.e. subglottal pressure) cannot be increased to match the increased losses in the system. Of course, if the shifting boundaries cause the phonation-producing set to shrink, the ability of the model to produce quasi-stable phonation can be considered to decrease.

On the other hand, one can ask whether the aim of producing quasi-stable phonation is realistic. Given the general trade-off between stability and agility, one

could argue that the articulatory system, whose main function requires high maneuverabilty, may, without active neural control, be just marginally, non-robustly stable or not stable at all. If the underlying phenomenon is not robustly stable then the more realistic the model becomes, the less likely it is to produce quasi-stable phonation.

Chapter 4

Tuning the model

The aim of this work is to produce a vowel synthesis model which can be used to produce glottal flow pulses using magnetic resonance imaging (MRI) data of the vocal tract (VT). A model suitable for this purpose was described in the previous chapter. Out of the four additions to the model that were discussed, automatic damping coefficient and dissipation in the VT are included in the model for the rest of this work. The subglottal tract resonator and turbulence in the glottis were excluded because, unless these submodels are developed further, they tend to make quasi-stable phonation more difficult to simulate.

For any such glottal pulse generator to be of practical use, it must be possible to run it with alternative settings so that it can produce a variety of flow pulses for each VT configuration. Indeed, even validation of the model relies on producing glottal signals which are similar to a reference signal, such as recorded sound or glottal area signal, in terms of some essential features.

Given the simplicity of the model, perfect replication of the reference signal is neither possible nor often desirable due to, for example, noise. Instead, the aim is to capture the most important source-related characteristics of the signal. For this model, two characteristics of glottal flow are used: (i) the fundamental frequency of phonation, f_0, and (ii) the phonation type, which is parametrised with the closing quotient, ClQ, as discussed in Section 3.2.2. This approach also has the benefit that these target characteristics are simple enough that they can be set arbitrarily without extracting them from a reference signal.

Tuning the model to match any given target output parametrisation (f_0^t,ClQ^t) is an inverse problem. How ill-posed this problem is depends on the number of input (i.e. tuning or control) parameters considered and how continuously the output parameters depend on the input parameters. It should be kept in mind that the exact parameter combination is of no interest, *per se*, in this application. Instead, we are interested in the glottal flow when the parametrisation of the pulses in that flow matches (f_0^t,ClQ^t). This means that any input parameter

combination which produces the desired output parameters is a valid solution to the tuning problem.

This chapter starts off by describing the tuning parameters chosen for this model. Next, the tuning problem is set up more formally in Section 4.2 and a strategy for solving the problem is shown in Section 4.3. In the last two sections of this chapter, the results of implementing the chosen tuning strategy are shown and discussed.

4.1 Tuning parameters

The four tuning parameters, the first eigenfrequency of the vocal folds, f_1, vocal fold length, h, inertial parameter, C_{iner}, and subglottal (lung) pressure, p_{sub}, were already mentioned in Section 3.2.2 when the other parameters for the model were discussed. These four parameters were chosen because there is reason to expect that they can impact at least on one of the output parameters.

Given that there are only two output parameters, redundancy is to be expected in the input parameters. Although such redundancy may not be desirable when the model is used in final applications, mainly due to computational speed considerations, it was kept for this work for three main reasons.

Firstly, sufficient prior information was not available for identifying which, if any, of the parameters could be left out.

Secondly, because most of the input parameters may affect more than one output parameter, it is likely that using only two input parameters limits the output range of the model making it impossible to, for example, produce certain phonation types at some f_0-values. The more input parameters are included in the model, the less such limitations there are in the output space. On the other hand, the computational effort needed to tune the model also increases with each added tuning parameter. As a compromise, 1-2 input parameters above the dimension of the output space was considered optimal.

And thirdly, most commonly used glottal pulse models, such as those based on the LF-model (Fant et al. 1985), use three independent parameters. Typically, these parameters are f_0 and two shape parameters, for example ClQ and the open quotient, OQ. The reason three output parameter were not used in the first place is that it makes selecting the target values harder, particularly as the shape parameters are not always truly independent in our model (see discussion later on). In any case, it is possible that in some cases a third parameter is needed to describe the model output. In these cases, four input parameters are needed to ensure sufficient redundancy.

4.1.1 Vocal fold parameters

The vocal fold model includes a large number of parameters which must either be estimated or set as tuning parameters. Vocal fold length, h, is chosen as a tuning parameter to reflect both gender and individual differences and to account for the speaker actively adjusting the length and mass of the vocal folds participating in phonation (Scimarella and d'Alessandro 2004). Notice that in this model, changing h also changes the vocal fold masses proportionally.

Manipulating both masses and spring constants is a commonly used strategy for adjusting f_0 and is often carried out through a special tension parameter (e.g. Aalto et al. 2009; Ishizaka and Flanagan 1972; Lous et al. 1998). Since masses are already tuned through h, however, a choice was made to tune the spring constants separately. To do this, the approach taken for example by Aalto (2009); Horáček et al. (2005) is used: eigenfrequencies, f_1 and f_2, of the vocal folds are chosen and the spring constants are calculated based on them (see Section 3.2.2). To reduce the number of tuning parameters, the ratio f_2/f_1 is kept constant and tuning is done through f_1 alone.

4.1.2 Glottal flow loss parameters and subglottal pressure

Aalto et al. (2009) used tuning of the glottal flow loss parameters C_{iner} and C_g (see Eq. (3.5)) and the subglottal pressure, p_{sub}, to match their flow pulses to LF-pulses for three different phonation types (breathy, normal, and pressed). Only the relative magnitudes of the three parameters matter, leaving essentially a choice of two out of these parameters for tuning.

The subglottal pressure is chosen as one of the tuning parameter because of the well studied relationship between it and the fundamental frequency, f_0 (see e.g. Scimarella and d'Alessandro 2004; Titze 1989).

The results of Aalto et al. (2009) suggest that the inertia parameter is of particular interest as it is related to phonation type. Hence C_{iner} is taken as the last tuning parameter. Notice, however, that C_g is not a constant but rather a derived parameter whose value depends on h (see Section 3.2.2).

4.1.3 Normalisation

The four tuning parameters are normalised in order to avoid scaling problems. Tuning algorithms use the Q-set

$$Q_f = \frac{f_1}{100\,Hz}, \quad Q_h = \frac{h}{18\,mm}, \quad Q_c = \frac{C_{iner}}{C^0_{iner}}, \quad \text{and} \quad Q_p = \frac{p_{sub}}{1000\,Pa},$$

where the nominal C^0_{iner} is calculated using Eq. (3.17). However, for the sake of clarity, the tuning parameters are generally referred to by the corresponding non-normalised symbols (f_1, h, C_{iner}, p_{sub}) in the following discussion.

4.2 Problem setup

Let $\mathbf{x} \in \mathbb{R}^4$ denote the vector of input parameters, i.e. $\mathbf{x} = (Q_f, Q_h, Q_c, Q_p)$, and let $f(\mathbf{x}) : \mathbb{R}^4 \to \mathbb{R}$ be the simulation process producing f_0 of the output of a simulation with input parameters $\mathbf{x}$, and $g(\mathbf{x}) : \mathbb{R}^4 \to \mathbb{R}$ be the process for producing ClQ for the same simulation. If simulation with a particular parameter combination does not produce quasi-stable phonation, $f = g = 0$ so that $S = supp(f) = supp(g)$ is a convenient definition of the set of parameters which produce phonation.

Both $f(\mathbf{x})$ and $g(\mathbf{x})$ are non-linear and furthermore cannot be assumed to be even continuous. These properties are a result of not only the mechano-acoustic interactions in the model but also caused by the numerical method of determined damping coefficient $b(\mathbf{x})$, which in effect limits b to discrete values.

Finding an input parameter combination which produces the output of interest (in this case f_0^t and ClQ^t) has been previously approached as an optimisation problem (e.g. Döllinger et al. 2002; Pinheiro et al. 2012; Schwarz et al. 2006; Yang et al. 2011). Translating this approach to the problem at hand, requires solving an unconstrained multi-objective optimisation problem for every target pair (f_0^t, ClQ^t):

$$\underset{\mathbf{x}\in\mathbb{R}^4}{\text{minimise}} \quad \mathbf{F}_m(\mathbf{x}), \tag{4.1}$$

where the objective function is

$$\mathbf{F}_m(\mathbf{x}) = \begin{cases} \left| \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix} - \begin{bmatrix} f_0^t \\ ClQ^t \end{bmatrix} \right|, & \text{when } \mathbf{x} \in S, \\ \\ \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}, & \text{otherwise,} \end{cases}$$

where C_1 and C_2 are large positive constants that place a high penalty to venturing outside S. Thus, they reflect a preference of any kind of phonation over none at the optimum.

Solving Eq. (4.1) directly is possible but suffers from some drawbacks. The first of these drawbacks is the large number of useless evaluations of $f(\mathbf{x})$ and $g(\mathbf{x})$ caused by not knowing before a simulation is completed whether $\mathbf{x} \in S$ or not. Clearly, this problem would be removed by performing the minimisation over

$\mathbf{x} \in S$, but finding the exact boundaries of S is very expensive computationally. Instead, S can be approximated with boundary constraints so that Eq. (4.1) becomes

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimise}} & \mathbf{F}_m(\mathbf{x}) \\ \text{subject to} & l_i \leq x_i \leq u_i \quad i = 1, ..., 4, \end{array} \tag{4.2}$$

where l_i and u_i are real positive constants which can be determined with a reasonable number of simulations.

The second major issue with solving problems (4.1) and (4.2) is their multi-objective nature. Most multi-objective optimisation methods require articulation of preferences regarding the different objectives either *a priori* or *a posteriori* (Marler and Arora 2004). Fortunately, the nature of the target parameters suggests a special way to deal with preferences.

It is straightforward to set a value for f_0^t but ClQ^t is more complicated. It is known, for example, that pressed phonation corresponds to low values of ClQ but what values constitute as low is not so clear. Inverse filtering gives a ballpark but due to differences in noise and possible pulse shapes, the exact values are not necessarily the same. Furthermore, simulations can produce unrealistic pulse shapes even if the ClQ-value is reasonable. Since posterior checks are required to account for these problems anyway, the solution set can be extended to cover a range of ClQ values and the choice of pulse shape can be integrated into the final checking and selection process. Although this approach is based on practical considerations, finding a much better solution does not seem feasible.

4.3 Solution strategy

The above idea leads to a three phase solution strategy.

1. **Optimisation.** The first phase consists of solving a single objective optimisation problem with boundary constraints

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimise}} & F(\mathbf{x}) \\ \text{subject to} & l_i \leq x_i \leq u_i \quad i = 1, ..., 4, \end{array} \tag{4.3}$$

   where

$$F(\mathbf{x}) = \begin{cases} |f(\mathbf{x}) - f_0^t|, & \text{when } \mathbf{x} \in S \\ C_1, & \text{otherwise.} \end{cases} \tag{4.4}$$

   Let us denote the solution to this phase as $\mathbf{x}^0$.

2. **Exploring the iso-f_0-set.** In the second phase a sequence of points $\{\mathbf{x}^j\}_{j=1}^J$ is computed such that $F(\mathbf{x}^j) \leq \epsilon_f$, where ϵ_f is a fixed tolerance level. Furthermore, the exploration algorithm must be able to access each point $\mathbf{x}^j$

from the previous point $\mathbf{x}^{j-1}$. For the purpose of this work, a point is considered *accessible* if the algorithm can find a direction $\mathbf{p}_j \in \mathbb{R}^4$ and step size $\mu_j \in \mathbb{R}\backslash\{0\}$ such that $\mathbf{x}^j = \mathbf{x}^{j-1} + \mu_j \mathbf{p}_j$. Possible directions and step sizes depend on the algorithm.

3. **Selection.** The final phase consists of identifying the points in the iso-f_0-set which approximately match ClQ^t, inspecting the corresponding pulse shapes, and selecting the final optimal point.

The methods used to solve the first two phases must be able to deal with the properties of $F(\mathbf{x})$: it is nonlinear, non-smooth, and possibly discontinuous. Further, its values cannot be estimated with an expression given in closed form, and evaluating its values computationally is slow. On average a single simulation takes 33 s.

4.3.1 Optimisation

From the point of view of the optimisation step, the above properties of the objective function give rise to two essential observations:

1. $F(\mathbf{x})$ cannot be assumed to be continuously differentiable, and
2. $F(\mathbf{x})$ cannot be assumed to be convex, i.e. multiple non-strict local and/or global minima are possible.

Given the expected redundancy in the input parameter space, the second observation is extended to assuming that

2b. $F(\mathbf{x})$ has multiple non-strict global minima where $f(\mathbf{x}) \approx f_0^t$ (else the exploration step fails), and these minima do not necessarily form a convex set.

Algorithms

Given the first observation about the nature of the objective function, gradient based methods are not a viable choice for solving Eq. (4.3). Direct search methods are a better choice and MATLAB offers a library of three such methods: pattern search, simulated annealing, and genetic algorithm. All three algorithms will be tested to determine which of them is best suited to the problem.

Pattern search is based on polling (function evaluations) around the current point and selecting a polling point which reduces the objective function value as the next point. Polling points are generated from the current point based on the chosen polling method and the current step size. For this problem, 8 polling

directions are used, one in the positive and one in the negative direction of each input parameter. The current step size depends on past polls: a successful poll increases the step size by a factor of 2, an unsuccessful poll decreases it by the same factor.

Pattern search is robust at the boundary of S even if C_1 is set to ∞ or `NaN`. With a minimum of eight function evaluations at each point, convergence rate is moderate. A potential problem with applying pattern search to a non-convex problem is that it may get stuck on a local minimum when the current step size is very small. This might account for the fact that this algorithm is not among the most popular used in existing vocal fold parameter tuning studies by other groups.

Simulated annealing is based on importance sampling the parameter space (Ingber 1996). Using the current point as a starting point, a test point is drawn randomly from a generating function. If the test point improves the objective function value, it is accepted as the next point. If this is not the case, the test point is accepted as the next point with a probability determined by an acceptance function. Both the generating and the acceptance functions depend not only on the current point and its objective function value but also on annealing temperature, a quantity which decreases with increasing number of iterations according to the chosen annealing function.

Simulated annealing can have a high convergence rate and does not easily get stuck in local minima. However, this requires that provided acceptance, generating, and annealing functions have been chosen and tuned well. This algorithm has been used successfully to determine optimal vocal fold parameters both in a two-mass model (Wurzbacher et al. 2006) and in a much more complex model (Yang et al. 2011).

Genetic algorithms are a wide class of algorithms which make use of the idea of biological evolution. Optimisation is started with an initial population of points in the parameter space each with a level of fitness, in this case the value of the objective function. Individuals (points) in the population are selected for reproduction, crossover, and mutation based on their fitness, producing a new generation with improved fitness. After some generations the best individual in the population will reach optimal fitness value, i.e. it is a global minimum.

The genetic optimisation algorithm is robust against local minima and finds reliably at least one global minimum. It is one of the more popular global optimisation methods used for vocal fold parameter tuning (Pinheiro et al. 2012; Pinheiro and Kerschen 2013; Schwarz et al. 2006). Sometimes a single run of the algorithm produces more than one distinct global minimum. However, for a four-dimensional parameter space 30-40 individuals are needed in each population and hence the computational cost of each optimisation tends to be high unless an optimum can be found in very few generations.

These three algorithms are compared in Section 4.4 in terms of their running times and convergence reliability in this particular application. The most promising of these three algorithms is then selected as the primary optimisation algorithm for the problem. In order to improve reliability, hybrid schemes are also considered.

4.3.2 Exploring an iso-f_0-set

From the point of view of the exploration step, the properties of the objective function listed in Section 4.3 lead to the following observations regarding any "iso-f_0-set", $\mathcal{F} := \{\mathbf{x} : F(\mathbf{x}) \leq \epsilon_f,\ l_i \leq x_i \leq u_i,\ i = 1, ..., 4\}$:

1. $\mathcal{F}$ cannot be assumed to be convex, and
2. $\mathcal{F}$ cannot be assumed to be connected.

Notice that $\mathcal{F}$ depends both on the value of f_0^t and on the VT configuration. For the third phase in the solution strategy, it is hoped that every $\mathcal{F}$ contains a large range of ClQ-values.

Algorithm

The aim of the exploration phase is to produce a sequence of points $\mathbf{x}^j \in \mathcal{F}$, $j = 1, ..., J$, such that each point is accessible from the previous one. The meaning of "accessible" in this context can be found in Section 4.3. While full isospace extraction would be ideal, this path exploration approach was chosen due to the high cost of computing the mesh of data points required by common isosurface algorithms such as marching cubes (Lorensen and Cline 1987) and the possible complications in applying the algorithms in a four-dimensional space.

A sequences starting from an arbitrary point in $\mathcal{F}$ may terminate before J points have been found, and the larger J is, the more likely this is to occur particularly if $\mathcal{F}$ is not connected. A simple solution is to use K starting points $\mathbf{x}^{0,k} \in \mathcal{F}$, $k = 1, ..., K$, for computing a set of sequences $\{\mathbf{x}^{j,k}\}_{j=1}^{J/K}$. The higher K is used, however, the more optimisation runs are required, and hence the higher the total computational cost becomes.

Now, supposing a starting point $\mathbf{x}^{0,k}$ is available, a modified pattern search method is used for exploring $\mathcal{F}$. At each current point, polling is performed using the same polling method and step size adaptation as in the pattern search optimisation scheme. One of the poll points within $\mathcal{F}$ is chosen randomly as the next point. If a poll is unsuccessful, step size is decreased and polling is repeated.

This method of choosing the next point translates to selecting a direction in which the largest step produces a tolerable change in $F(\mathbf{x})$. This a desirable property, as it seems reasonable to assume that if ClQ varies within $\mathcal{F}$, more of these variations can be discovered by exploring a larger range of points in $\mathcal{F}$.

Cycle prevention

If polling is performed in all 8 direction, the exploration algorithm can easily get stuck in cycles. To prevent this, two different cycle prevention schemes are tested. In the first scheme, henceforth referred to as *close point removal*, polls are not performed at points which are within a 4-orthotope of side lengths $2\delta(u_i - l_i)$, $i = 1, ..., 4$, centred at any of the points in the already computed sequence. In other words, point $\mathbf{x}^{p,l}$ is not polled if for any $\mathbf{x}^{j,k}$, $j = 0, ..., p-1$, $k = 1, ..., l-1$,

$$|x_i^{p,l} - x_i^{j,k}| \leq \delta(u_i - l_i), \quad \text{for all } i = 1, ..., 4. \tag{4.5}$$

The second cycle prevention scheme, henceforth *poll direction removal* scheme, also removes poll points which fulfil criterion (4.5). In addition, poll directions which appear already to have been explored are also removed. This check is done at each point before the first poll by computing poll points with half the starting step size. If any of these poll points meets (4.5), the corresponding direction is removed from all polls at the current point. Hence, at each point, the direction from which the algorithm has just arrived is not polled. Depending on the exploration history, other directions may get removed as well.

These two schemes are tested and compared in two 2D cases reflecting the observations made on iso-f_0-set, i.e. non-convexity and disconnectedness.

Starting points

As mentioned above, the exploration algorithm requires K starting points in $\mathcal{F}$. These starting points are solution points found by the optimisation algorithm. With the occasional exception of the genetic algorithm, the considered optimisation algorithms converge to a single solution point. Hence some method is needed to produce the K different points.

Randomisation based optimisation algorithms, i.e. simulated annealing and genetic algorithm, will likely produce different solutions depending on the state of the random number generator. Genetic algorithm also produces its initial population as a random draw from a uniform distribution over the feasible region in (4.3). Hence running the genetic algorithm K times will produce the desired set of solutions.

For pattern search, as well as for simulated annealing (to introduce additional variation), K different starting points for the algorithms are used. These starting points are generated using Latin Hypercube design (McKay et al. 1979) over the feasible region. In order to save computation time, an optimal point is not used as a starting point for exploration if its Euclidean distance from an already used starting point is less than a predefined tolerance level.

Table 4.1: Performance comparison of optimisation algorithms

		Pattern search	Simulated annealing	Genetic algorithm
function evaluations	average	50	46	180
	worst case	171	90	570
failure rate		0.02	0.08	0.01

4.4 Tuning results

4.4.1 Comparison of optimisation algorithms

The three optimisation algorithms described in Section 4.3.1 were compared in terms of number of function evaluations and their failure rate (fraction of runs for which $F(\mathbf{x}) > 0.005 f_0^t$ at termination). Five different starting points or populations were used for each algorithm for each of the four vowel geometries, [ɑ, i, u, œ]. Five different target frequencies were used for each of these runs: 100 Hz, 130 Hz, 160 Hz, 190 Hz, and 220 Hz. These vowel-f_0-combinations will be denoted with ([v], f_0), where [v] is the vowel geometry and f_0 one of the five levels. Thus, there are 20 ([v], f_0) combinations, and altogether 100 searches were performed with each algorithm.

A summary of the results is shown in Table 4.1. The table shows that pattern search and genetic algorithm proved to be more reliable at finding an optimum than simulated annealing, although this may reflect a failure to find suitable settings for simulated annealing rather than unsuitability of the algorithm for this application. No evidence was seen of pattern search getting stuck on local minima. Either the objective function has very few of them or local peaks are narrow compared to step sizes.

The table also confirms the high computational cost of genetic algorithm. In particular, the number of function evaluations in the worst case is very high. If parallel computation is not available, finding a set of exploration starting points for a new vowel geometry could take over a day for each f_0-value using the genetic algorithm as opposed to about eight hours with pattern search and four hours with simulated annealing (using a server with Ubuntu 12.04, Intel Xeon X5650 CPU @ 2.67 GHz with 12 cores, and 53 GB RAM.).

If parallel computation is possible, however, the relative expenses of the algorithms change due to their differing abilities to make use of it. This is illustrated in Figure 4.1. Simulated annealing is not easily parallelisable and its MATLAB implementation does not benefit from using parallel workers. Pattern search and genetic algorithm perform multiple function evaluations before moving to a new point and hence can better utilise parallel computation. In this application pattern

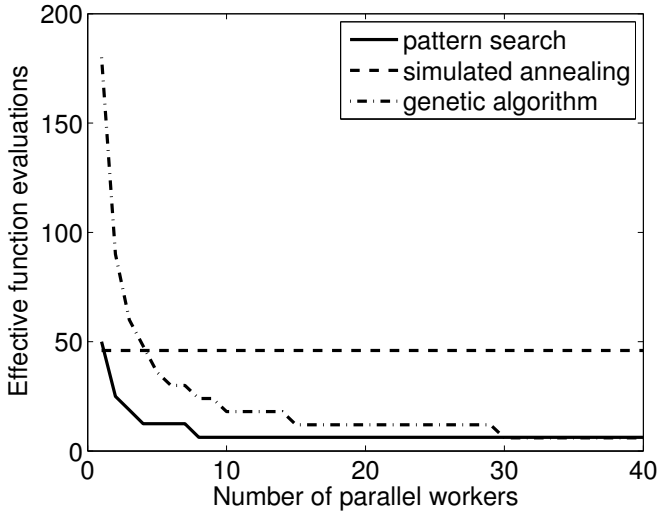


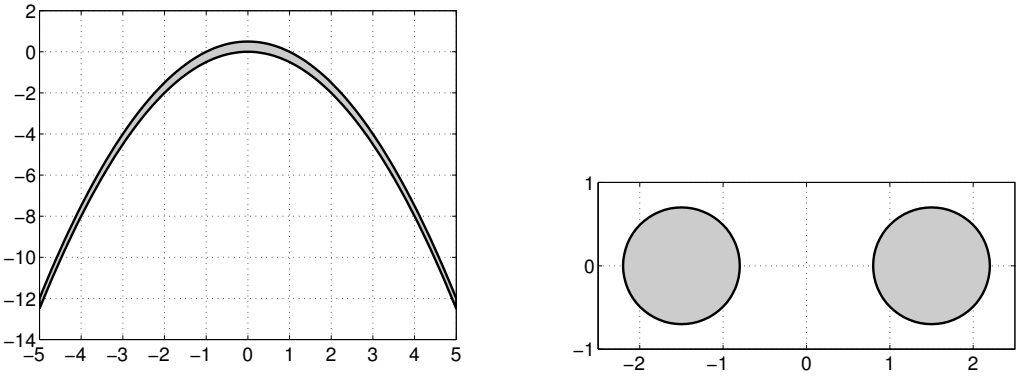
Figure 4.1: Effective cost of the optimisation algorithms when parallel computation is utilised.

search performs eight polls before changing current point and/or step size. Genetic algorithm computes fitness value of 30 individuals in each generation before producing the next generation. Hence for a modest number of parallel workers, pattern search is the fastest algorithm available.

Differences in optimisation performance between vowels and f_0 levels was not analysed in detail. However, it was noted that pattern search appears to run into difficulties with ([i], 220 Hz). The algorithm failed in two out of the five runs at this vowel-f_0-combination, which accounted for two of its highest function evaluation counts. Simulated annealing, starting from the same initial points, showed no particular signs of problems. Out of the three algorithms, simulated annealing performed most consistently in the sense that no vowel, f_0 or their combination required significantly more effort than the others. For genetic algorithm, high frequencies appeared to be more difficult than low with 190 Hz and 220 Hz requiring on average 3-6 generations more to converge than 100 Hz and 130 Hz.

4.4.2 Testing cycle prevention schemes

Figures 4.2a and 4.2b show two example cases in two-dimensions which are used to test and compare the cycle prevention schemes. In both figures, $\mathcal{F}$ is shown as the shaded region. In Figure 4.2a, $\mathcal{F}$ is the region between two parabolas separated by a constant vertical distance of 0.5. This was chosen as a test case due to its non-convexity, tendency of the exploration algorithm to cycle and because then narrow ends are nearly at 45° to the poll directions and hence exploring them



(a) (b)

Figure 4.2: Example cases: (a) non-convex region between two parabolas and (b) two disconnected circles.

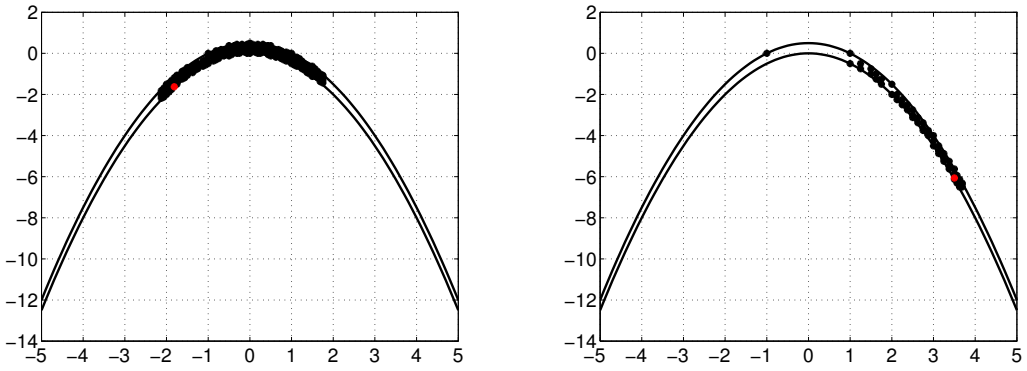
require slow zig-zagging.

The second case (Figure 4.2b) tests the performance of the algorithm for a space that is not only non-convex and cycle-inducing but also disconnected. The dimension of the circles is fairly large ($r = 0.7$) compared to the minimum step size ($d_{min} = 0.05$) so that exploring within the circles is relatively easy. On the other hand, jumping between the regions is made non-trivial by choosing the minimum distance between the regions to be larger than $2r$ (1.6 is used here). It is worth noting that jumping becomes impossible in two situation. First, the minimum distance between the circles exceeds the maximum step size that can be attained in this geometry ($4r$). And second, the centres of the circles are separated both horizontally and vertically by more than $2r$. In the second case, no matter where the current point is in one circle, no point in the other circle lies in any of the poll directions.

The points of interest in the performance of the cycle prevention schemes are how many points they generate before termination, how many polls they perform to obtain those points, and how the points are distributed over the set. The first two measures can be easily summarised quantitatively. The distribution of the points is assessed visually and in the second case also via the number of jumps between circles.

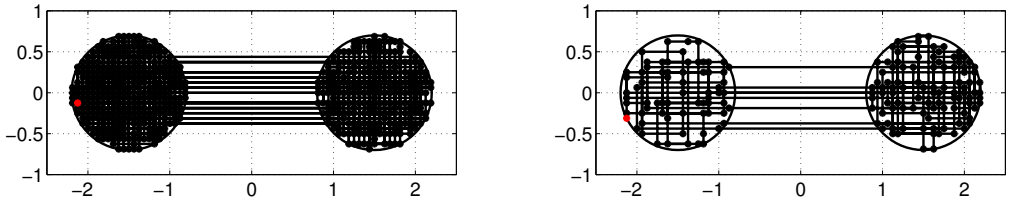
In both cases the algorithm starts at the same state: starting point $(-1, 0)$ has been explored, and the current point is $(1, 0)$ and starting step size $d = 4$. Proceeding from this state, the first poll will fail and step size will be reduced to $d = 2$. If no steps are taken to prevent cycles, this produces one successful poll at $(-1, 0)$ and the algorithm thus gets stuck cycling between the two points.

The performance numbers for the two cycle prevention schemes are summarised



(a) (b)

Figure 4.3: Non-convex parabola exploration case: (a) close point removal scheme and (b) poll direction removal scheme.



(a) (b)

Figure 4.4: Disconnected circles exploration case: (a) close point removal scheme, and (b) poll direction removal scheme

in Table 4.2. The table shows that close point removal produces more points before termination than poll direction removal, but the number of polls needed to find a valid exploration point is lower for poll direction removal.

Typical results for each algorithm in the parabola case are shown in Figures 4.3a and 4.3b and in the circle case in Figures 4.4a and 4.4b. In the case of the non-

Table 4.2: Cycle prevention scheme performance on average in 50 runs.

Cycle prevention scheme	Non-convex		Disconnected	
	Points	Polls	Points	Polls
Close point removal	250	1740	593	3963
Poll direction removal	65	380	102	540

convex parabola, the schemes show clear tendency to explore different regions of the parabola. Close point removal tends to explore the curving region and perform U-turns rather than explore the narrow ends. Poll direction removal explores the right narrow leg and typically terminates soon after doing a U-turn.

In the case of the disconnected circles, close point removal produces a more uniform grid of points than poll direction removal. The latter scheme also shows a tendency to terminate without exploring the two circles equally. In any case, for an equal number of points explored, the coverage of the exploration space is similar for the two algorithms, unless the distance between disconnected regions becomes very large.

These two examples have illustrated key features in the performance of the two cycle prevention schemes. Because close point removal allows backtracking, it is less likely to terminate before the entire set has been explored. On the other hand, the same backtracking property tends to lead to "stalling", i.e. exploring some part of the set at a finer grid than necessary rather than moving on. Hence simply taking the J/K first points produced by the algorithm may not give much coverage of the region.

As the poll direction removal scheme allows backtracking only through U-turns, it shows a higher tendency to move away from already explored regions. This same property makes the algorithm more likely to "corner" itself, i.e. terminate due to badly chosen next point. As an extreme case, if the algorithm arrives at a locally convex boundary of the exploration space from a direction normal to the boundary, the result is always termination. In the actual four-dimensional vowel application, occurrence of such an extreme case is possible but unlikely.

Since a number of different starting points will be used anyway, stalling is considered more problematic than cornering. Hence based on the above observations, poll direction removal will be used in the exploration algorithm.

4.4.3 Exploration

Simulations were run using the exploration strategy detailed in Section 4.3.2 with poll direction removal as cycle prevention method. A two-dimensional search space example is shown in Figure 4.5 for [ɑ]. A single exploration sequence with target frequency $f_0^t = 130$ Hz starting from the point (1.15, 2.15) is shown as a black line with dots indicating explored points. Behaviour similar to the non-convex parabola example can be seen: U-turns and slow zig-zagging, indicating a narrow leg of $\mathcal{F}$ that is not parallel to the coordinate axes. The sequence covers a range of ClQ values from 0.32 to 0.43 with 30 explored points. A second exploration sequence with a different starting point can be expected to improve this range significantly, and the strategy is considered successful.

On an unrelated note, Figure 4.5 illustrates a search space where at medium

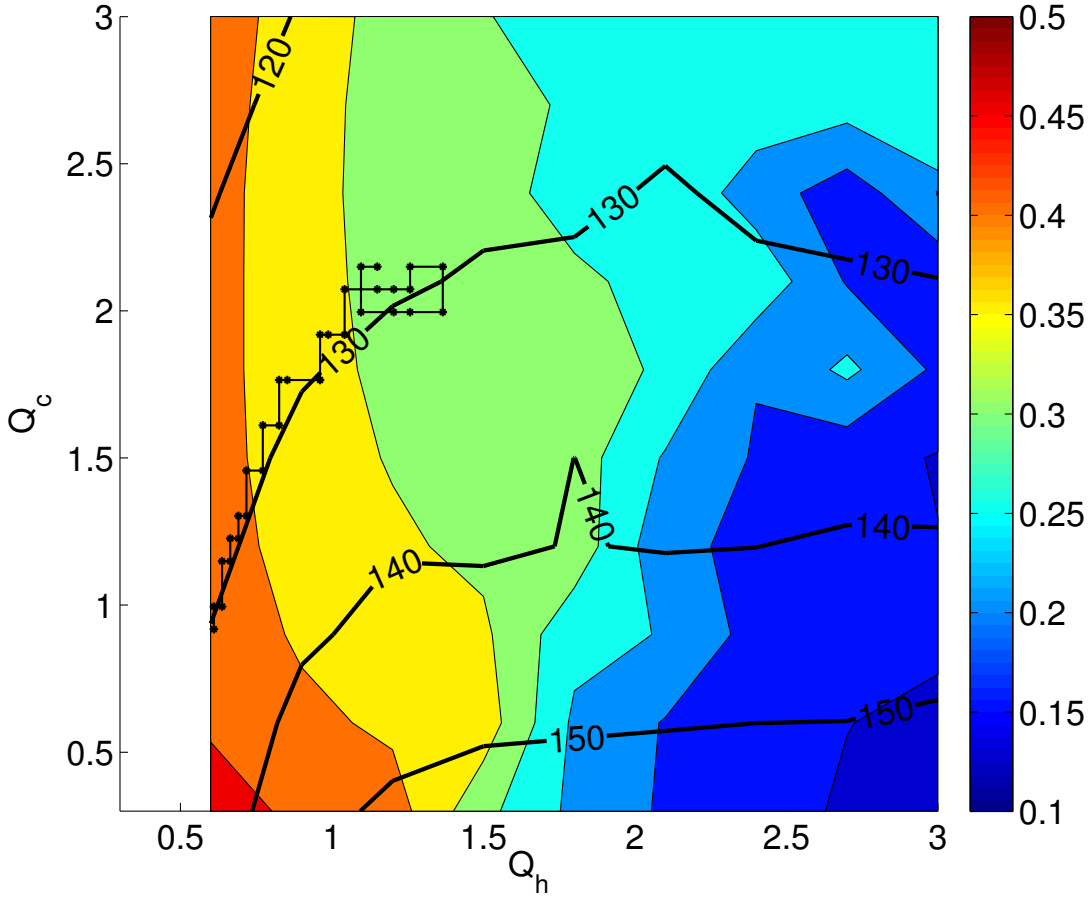


Figure 4.5: A single exploration sequence for ([a], 130 Hz) in the Q_c-Q_h -space ($Q_f = 0.9$ and $Q_p = 0.8$) is shown as line with dots. Coloured contour plot and colour bar indicate ClQ-values in the search space and thick black lines f_0-values, both computed on a grid of 10×10 points with $Q_c, Q_h \in [0.3, 3.0]$.

f_0 two input parameters would suffice to tune the model to both f_0^t and ClQ^t. In fact, the output appears so well-behaved that it might be possible to "read off" the parameter values that produce the desired output. However, at low and high f_0 only limited phonation types are available and even at medium f_0 very high ClQ-values are not possible as the boundary of the phonation producing search space, S, runs between $Q_h = 0.3$ and $Q_h = 0.6$ for all the Q_c shown. This limitation in available f_0-ClQ-combinations is one of the reasons why higher number of input parameters was considered desirable. It should also be noted that finding subsets of S where the model output is as well-behaved as in Figure 4.5 is not a trivial task.

Next, performance of the exploration strategy is tested in the full four-dimensional input parameter space for a range of vowel-f_0-combinations. The same set of combinations are used as with testing optimisation algorithms: ([v], f_0) where [v] is one of the vowel geometries [ɑ, i, u, œ] and f_0 is 100 Hz, 130 Hz, 160 Hz, 190 Hz or 220 Hz. $K = 5$ distinct starting points were used, as produced by the optimisation phase, except with ([u], 100 Hz) where two of the optimisation runs converged to a single point and hence only four starting points were available. From each starting point, exploration is carried out for $J/K = 30$ steps unless the exploration algorithm terminates before this. The maximum number of points produced is hence $J_{max} = 155$ (124 for ([u], 100 Hz)).

Figure 4.6 shows the exploration results in terms of the number of points successfully explored, J_{true}, and the distribution of the ClQ values for these points. In most cases (85%), at least one of the exploration sequences was terminated before 30 points were explored and hence J_{max} was not reached. In 75% of the cases $J_{true} \geq 0.7J_{max}$ but for ([ɑ], 190 Hz) and ([u], 100 Hz) $J_{true} < 0.5J_{max}$. Visual inspection suggest, however, that premature termination does not necessarily mean that the range of ClQ-values covered is worse than when full number of exploration points are reached. For example, compare ([u], 100 Hz) and ([œ], 220 Hz).

There is some evidence that the exploration algorithm could be stalling: very high peaks (e.g. ([i], 220 Hz), ([œ], 220 Hz)) and gaps in the ClQ-range covered (e.g. ([ɑ], 190 Hz), ([i], 160 Hz)). However, there are fewer peaks and gaps than there were starting points. Hence either not all exploration sequences stall, or they tend to stall at certain ClQ-values regardless of starting point. In fact, an alternative explanation for the peaks and the gaps is that certain ClQ-values are more prevalent while other values might be missing in the (possibly disconnected) regions of the iso-f_0-set that are being explored.

ClQ range approximately from 0.1 to 0.4 is covered fairly well. Comparison with the ranges found using inverse filtering suggests that this gives coverage of all three phonation types (Table 4.3). However, ClQ-values below those shown in Table 4.3 occur rather often in simulations, and ClQ-values in the upper end of

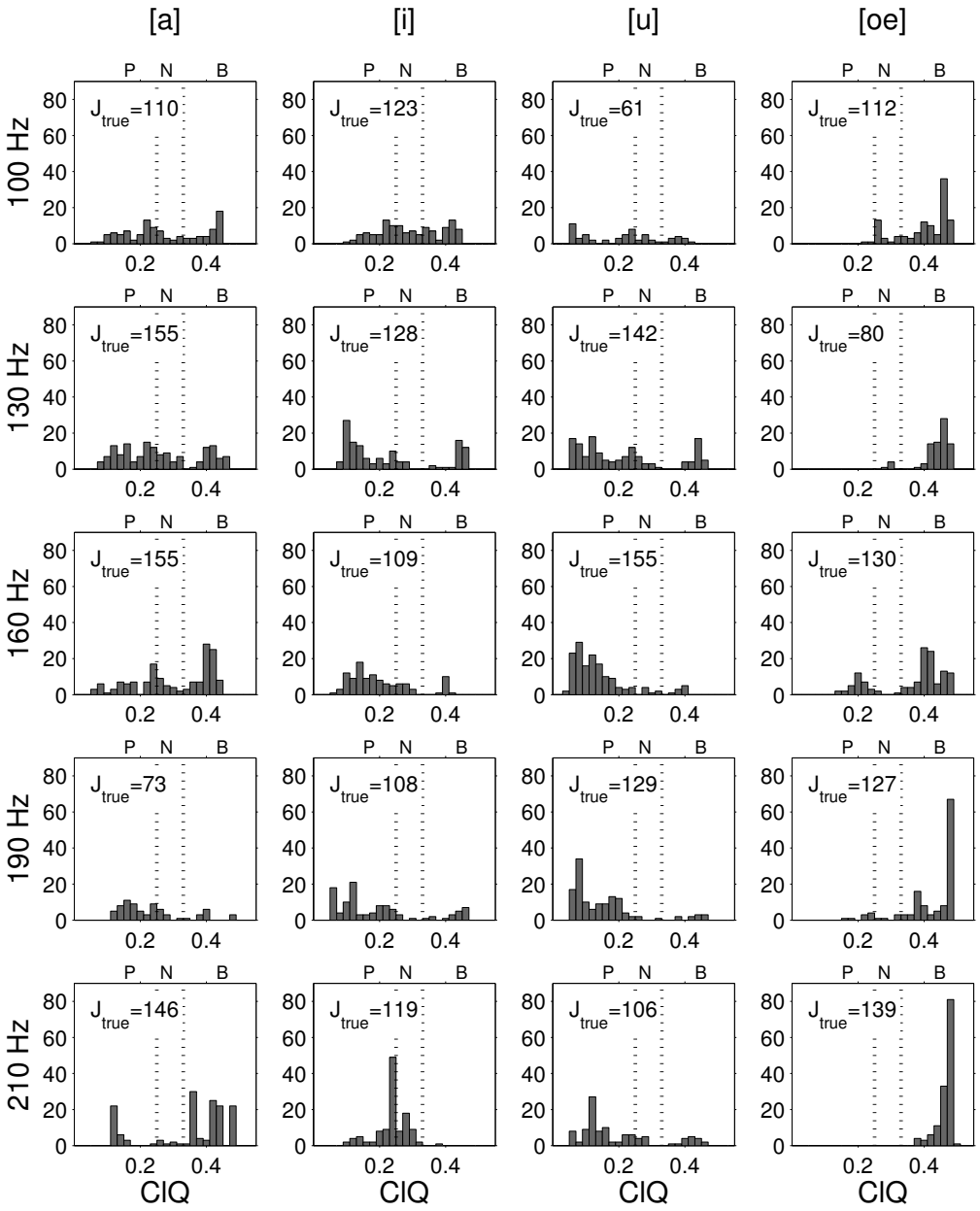


Figure 4.6: Histograms of ClQ values for each vowel-f_0-combination. J_{true} is the number of points successfully explored (the aim was 155 points). The vertical lines divide the ClQ values roughly into regions of pressed (P), normal (N), and breathy (B) phonation.

Table 4.3: ClQ-values corresponding to different modes of phonation in inverse filtering as reported by Alku et al. (2002: Tables I and II)

	Male		Female	
	mean	range	mean	range
Pressed	0.22	0.18-0.25	0.26	0.22-0.32
Normal	0.27	0.24-0.30	0.29	0.26-0.36
Breathy	0.45	0.38-0.51	0.40	0.29-0.48

breathy range rare. The latter observation is explained by the fact that the model produces symmetric pulses when $C_{iner} \to 0$, so that the model ceases to produce phonation before the theoretical maximum for the model, $ClQ = 0.5$, is reached.

Generally speaking, the pressed end of the ClQ scale is covered better than the breathy end. The clear exception to this is [œ], with virtually no pressed phonation at three f_0 levels. It can also been seen that in many cases, for example ([ɑ], 130 Hz), ([i], 160 Hz), and ([u], 220 Hz), the transitional region between normal and breathy is covered poorly if at all. Since the algorithm does not explore the entire $\mathcal{F}$, it cannot be said for certain whether these observations are caused by general model behaviour or by the choice of the starting points. It is also possible that simulated pulse shapes differ from inverse filtered pulses in such a way that ClQ does not accurately identify the different phonation types from the simulated pulse shapes. Changes in predominant phonation types may occur at lower ClQ values in simulations, leading to apparent over representation of pressed phonation.

Some insight into the model behaviour can be gained by comparing the ClQ-values with the open quotient, OQ (Figure 4.7). Normally, as phonation moves from pressed to breathy, the glottis remains open for a larger fraction of the glottal cycle. At pressed and normal phonation the glottis is expected to close fully in most cases, while glottal leakage, i.e. incomplete closure of the glottis, becomes more common as phonation moves towards breathy.

Figure 4.7 shows that increasing ClQ is accompanied by increasing OQ until the open quotient saturates out at $OQ = 1$. However, at most vowel-f_0-combinations the glottis closes fully only when phonation is pressed. In the rest of the cases, except for ([ɑ], 190 Hz), some tuning parameter combinations produce glottal closure at normal phonation while other combinations do not. ClQ in the breathy range is nearly always accompanied by glottal leakage. These observations support the inference that the ClQ ranges obtained from inverse filtering cannot be directly applied on the simulation results or the model due to differences in pulse shapes.

At a few vowel-f_0-combinations (e.g ([ɑ], 100 Hz)) a least-squares regression line can be fitted fairly successfully to the data after removing the points where OQ

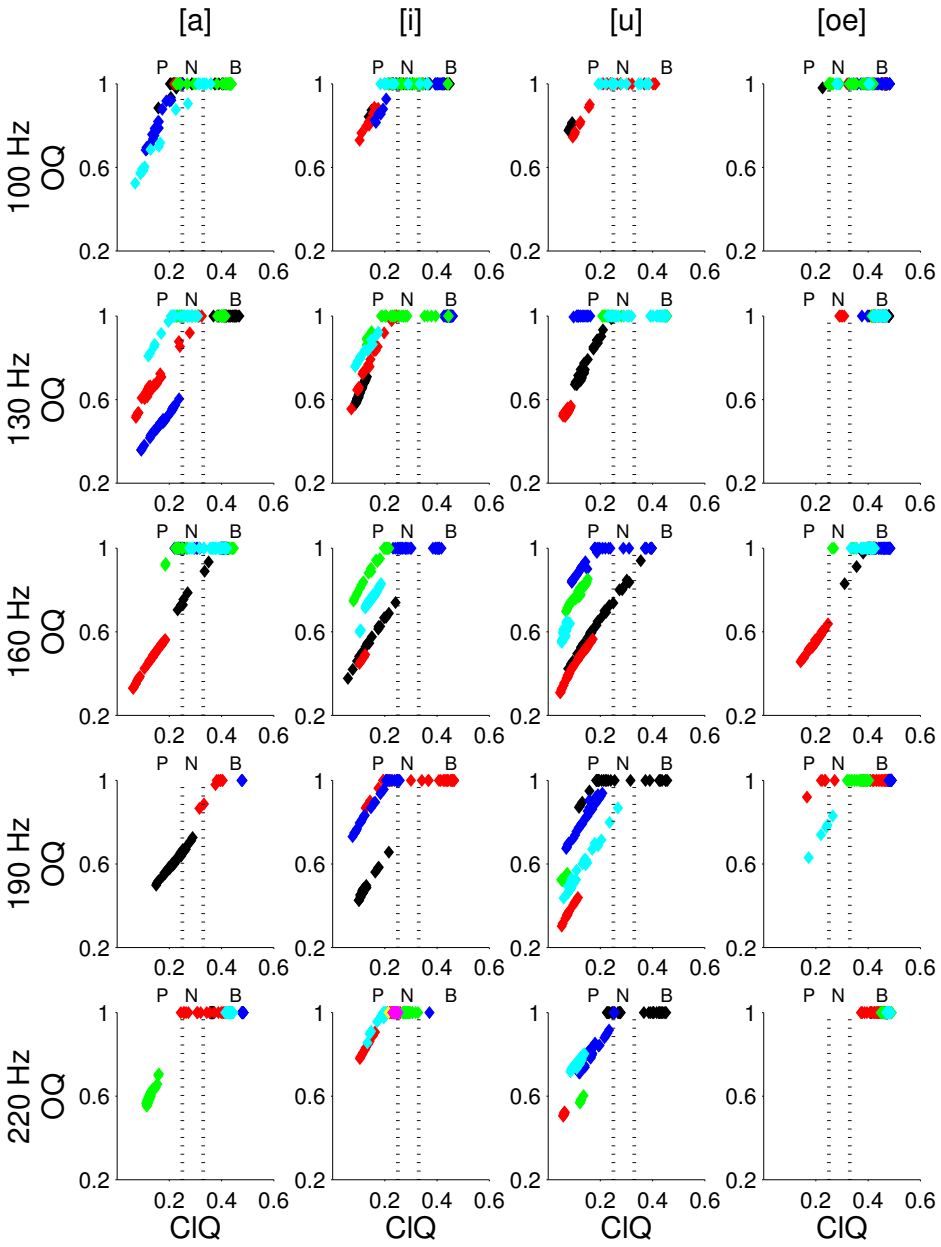


Figure 4.7: *OQ* versus *ClQ* for each vowel-f_0-combination. Different colours indicate different exploration sequences, i.e. from different starting points. The vertical lines divide the *ClQ* values roughly into regions of pressed (P), normal (N), and breathy (B) phonation.

has saturated due the glottis remaining open throughout the cycle. On the other hand, there are several cases (e.g ([ɑ], 130 Hz)) where a single linear regression line does not fit the data very well. Instead, it seems that a linear relationship applies to the exploration sequences separately, although in some cases (e.g. ([u], 160 Hz)) some sequences can be combined. Full analysis on relationship between the two output parameters has not been carried out as part of this work, but in these later cases it seems that ClQ and OQ contain different information about the pulses.

4.4.4 Choice of pulse shapes

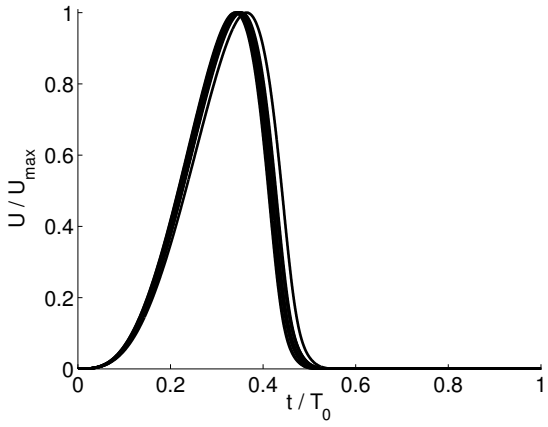
The last phase in solving the tuning problem, Eq. (4.2), is choosing a pulse shape from $\mathcal{F}$ that best matches user preferences. Results presented in the previous section indicate that manual pulse shape selection is necessary as ClQ does not fully represent the shape of the pulses produced by simulations, as was supposed to begin with.

Generally, it can be said that groups of points in $\mathcal{F}$ with similar ClQ fall under one of three categories. First, there are cases in which specifying an approximate ClQ^t suffices to determine a pulse shape. Figure 4.8a shows an example where a set of points with distinct input parameters nevertheless produce very similar pulse shapes with $f(\mathbf{x}) \approx f_0^t$ and $g(\mathbf{x}) \approx ClQ^t$. This is typical at low ClQ-values, that is, at pressed phonation.

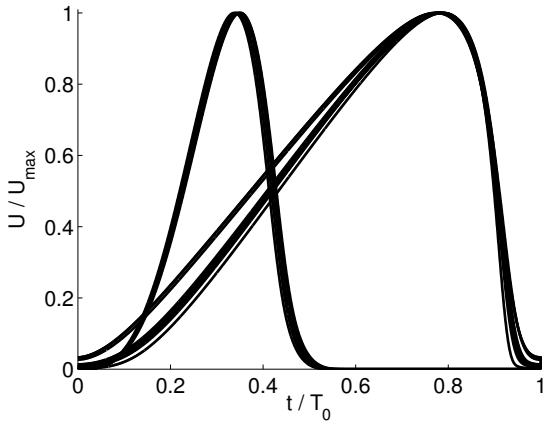
The second category is the most common one at normal and breathy phonation. In this category the pulse shapes form a few subgroups with very similar pulses within each subgroup. Differences between the subgroups vary from very clear to so small they form a continuum. Figure 4.8b shows an example of this category with clear subgroups. If the glottis closes in some or all of the subgroups, the clearest difference between them is typically the open quotient. In these cases, using both ClQ and OQ as target parameters would be beneficial. If the glottis remains open in all subgroups, open quotient is of no use, but looking at the minimum value of the normalised flow can help to differentiate the subgroups.

The final category is why posterior checks were deemed necessary in the first place. In this category some individual or subgroups of pulses show atypical features. Figure 4.8c shows a group of pulses typical for the ClQ-value (pressed) and another group where the pulses have a "ripple" on the rising edge. These atypical pulses are mild cases of the multiple local maxima phenomenon that was observed in Section 3.3. Neither ClQ nor OQ is able to reflect these difference in the rising edge.

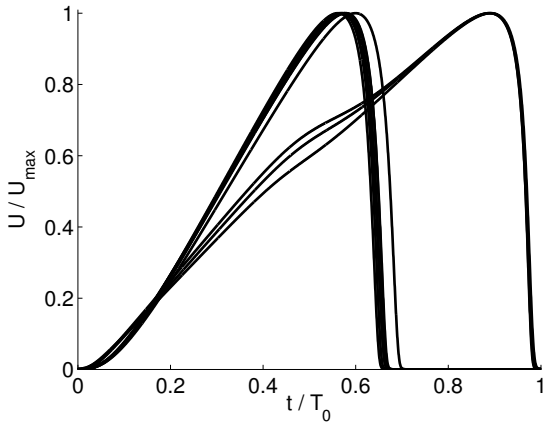
Nevertheless, these three example figures show that ClQ can be used as an starting indicator for the phonation type, as long as it is not relied too heavily on.



(a)



(b)



(c)

Figure 4.8: Pulse shapes corresponding to (a) [œ], $f_0 = 160$ Hz and $ClQ \approx 0.19$, (b) [ɑ], $f_0 = 130$ Hz and $ClQ \approx 0.22$, and (c) [u], $f_0 = 130$ Hz and $ClQ \approx 0.11$.

4.5 Discussion on tuning

Methods for tuning the model have been developed and tested. First, four model parameters were selected to function as tuning parameters. Second, the tuning problem was set up formally, and the strategy for solving it was described. And finally, simulations were carried out to select the best algorithms and to test their performance. Overall, a working solution was found.

4.5.1 Tuning parameters

Four parameters from the vocal fold and glottal flow models were used as tuning parameters: vocal fold length, h, the first eigenfrequency of the vocal folds, f_1, subglottal pressure, p_{sub}, and inertia parameter in the flow equation, C_{iner}.

This tuning parameter set is not the only possible one. For example, masses and spring constants could have been tuned together using a tension parameter (such as in Aalto et al. 2009; Ishizaka and Flanagan 1972), instead of separately through h and f_1. Or, the masses could have been tuned independently of h (as was done by Scimarella and d'Alessandro 2004) so that the latter would have affected force and flow equations only.

There are also a number of other vocal fold parameters which have been treated as constants in this work (see Section 3.2.2) but which could have been chosen as tuning parameters. Based on studies on other two-mass-models (Ishizaka and Flanagan 1972; Horáček and Švec 2002), the production of quasi-stable phonation in the model is expected to be sensitive to the glottal gap at rest. Hence, including this parameter in the tuning parameter set could expand the output space, that is, the model might be able to produce a wide range of f_0 and ClQ combinations. On the other hand, the impact of altering the glottal gap at rest can be countered to some extent by adjusting p_{sub} in particular if turbulence losses are not included.

The parameter C_{iner} represents the inertia of the air column in the VT. Naturally, this is not something that a speaker can control without altering the VT configuration. Instead, it was selected as a tuning parameter to compensate for the simplifications in the model. It is prudent to ask whether it is necessary to use C_{iner} as tuning parameter or if it could be replaced by some other parameter without reducing the output space.

As long as we are not interested in the values of the parameters, however, any reasonable choice of tuning parameters suffices. Indeed, when the model is used as a glottal pulse generator, only the produced pulse matters, not the details of the model. The most compelling reason to reconsider the choice of tuning parameters in the future is the need to reduce the dimension of the search space. In that case the tension parameter approach is a good place to start but would require further investigation to ensure that it does not limit the f_0 and ClQ combinations that

the model can produce.

For this work, the area function and centreline of the VT have been taken as given. However, extracting these from the MRI data suffers from some ambiguities. Particularly, the length of the centreline could be adjusted without changing the area function, and this would lead to changes in the resonance structure of the VT representation. This tuning method has been left out of this work completely, but it could be used to match Webster's resonances to some particular targets, as may be necessary in future development of the model.

4.5.2 Problem setup

The tuning problem was set up as an unconstrained multi-objective optimisation problem that was reduced to a constrained single-objective optimisation problem. The main problem with this optimisation approach is the high computing time when it is used for a large number of targets and VT configurations.

Computationally lighter alternatives are few, however. The most promising alternative is to fit a response function to a training set of simulations, i.e. construct approximate functions $\hat{f}(\mathbf{x}) \approx f(\mathbf{x})$ and $\hat{g}(\mathbf{x}) \approx g(\mathbf{x})$, so that a solution, $\mathbf{x}^*$, satisfies

$$\begin{cases} \hat{f}(\mathbf{x}^*) &= f_0^t, \\ \hat{g}(\mathbf{x}^*) &= ClQ^t. \end{cases} \tag{4.6}$$

The initial computational cost of such an approach is high, but can be reduced by using adaptive sampling.

For example, Gaussian process trees have been used successfully to fit a response surface to computational fluid dynamics simulations of a space vehicle without excessive computing cost (Gramacy et al. 2004). The problem with a response surface approach is that Eq. (4.6) is notably faster to solve than the original problem only if $\hat{f}$ and $\hat{g}$ are invertible or if the corresponding optimisation problem can be solved easier. Unfortunately, if the response functions are good approximations for the non-smooth and discontinuous f and g, neither of these is true. Nevertheless, response functions could prove to be useful for narrowing down the search space, particularly if similarities between responses for different vowels can be found.

4.5.3 Sensitivity of the model to tuning parameters

How sensitive the model is to changes in its parameter values is a question closely related both to the selection of the tuning parameters and to the response surface approach for solving the tuning problem (see above). Sensitivity of model can be determined using extensive computation and analysis of the approximate functions

$\hat{f}$ and $\hat{g}$. Besides helping to identify efficient tuning parameters and to indicate good starting points for optimisation, this would also answer a number of questions about the model.

One such question is whether the need for a high glottal damping coefficient, b, is caused by unrealistic parameter combinations. The tuning parameters are allowed to vary independent of each other and the rest of the system which might lead to, say, female vocal folds (length and eigenfrequencies) coupled with male vocal tract and lung pressure. In addition, these are combined with parameters with fixed values, such as vocal fold thickness and glottal gap at rest. This mismatch could produce situations where the system must be artificially damped with b because the "natural" methods, such as adjusting the subglottal pressure, p_{sub}, have been removed.

A sensitivity study gives insight both to the model and to the underlying system but it was left for the future as a satisfactory treatment would be more extensive than what could fit into this thesis.

4.5.4 Optimisation and exploration algorithms

Since optimisation appears unavoidable, the only solution left is to reduce its cost. Constricting the search space and using a single-objective function served this function. Of course, while using a single-objective function is a benefit in the optimisation step, it requires the exploration step to complete the search.

Out of the three optimisation algorithms considered, the final choice, pattern search, is not one of algorithms commonly used in vocal fold parameter optimisation studies.

One possible reason for this is that pattern search may suffer from getting stuck on local minima in non-convex problems. However, no signs of this were seen during simulations. Another reason for the surprising outcome is that none of the algorithms were fully optimised for solving the tuning problem. It may have been coincidental that the settings used for pattern search produced the best match for the problem. Nevertheless, this good match was consistent across all optimisation runs, and hence it can be said with a fair degree of certainty that pattern search works well in this application.

The exploration algorithm was also developed with one primary criterion in mind: it needs to work. In this respect, the pattern search based algorithm that removes explored poll directions is a success. It does have some weaknesses, however. Because the algorithm is based on an optimisation tool, it has an inherent tendency to stall, i.e. circle in the vicinity of an single point in the exploration space. Cycle removal can reduce this tendency but it cannot be removed completely without compromising the ability of the algorithm to cope with encountering the boundary of $\mathcal{F}$. The only way to effectively eliminate the problem is to

approach the exploration problem differently. Some alternative approaches that could be considered are Marching Cubes (Lorensen and Cline 1987) or seeded region growing (Adams and Bischof 1994), although both approaches would require adaptation to account for the high-dimensional search space and to avoid excessive simulation time.

Even setting the issue of stalling aside, exploration may be slowed down by use of fixed orthogonal search directions, constantly changing step size, and random selection of the next point among successful polls. Some degree of adaptation in the algorithm regarding search directions and step sizes and selecting the next point based on largest ClQ change would improve the rate of ClQ-range coverage at least where the target functions are smooth. The performance of such adaptive algorithms near points of discontinuity would need to be studied carefully, however.

4.5.5 Pulse shapes

The conclusions drawn about how well the three phase tuning strategy works were based on the output parametrisation selected in the previous chapter. Simulation results indicated, however, that in some cases the strategy relies quite heavily on the final manual checking and selecting phase. Closing quotient, ClQ, which was selected to describe phonation type, does not always fully describe this aspect of the pulses.

Problems were noted when the simulated ClQ-values were compared with phonation type ranges obtained using inverse filtering. Looking at values of the open quotient, OQ, also indicated that the pulse shapes obtained from simulations differ from inverse filtered pulses. The model has a tendency to produce lower ClQ and higher OQ, and glottal leakage ($OQ = 1$) is common. One possible reason for this is the choice of (constant) parameter values. It is not inconceivable that, for example, a too high value of the glottal gap at rest, g, could cause glottal leakage. Unfortunately, the model is very sensitive to g, and finding a value that improves the pulse shapes without causing significant changes to stability is not trivial.

Curiously enough, it was noted that these problems with the pulse shapes are reduced significantly by taking the glottal output velocity, $v_o(t)$, to be the velocity through the glottal area $A_g(t) = h\Delta W_1(t)$ instead of the constant control area, hH_1, both when scaling the VT input velocity (Eq. (3.10) and when extracting the output parameters. This causes sufficient changes in the parameter space to make the exploration step easier as well. Whether this behaviour is indicative of the model missing a crucial element, e.g. flow separation at the vocal folds, or caused by a particular oversimplification in the model or merely coincidental, remains an open question.

The reliance of the system on the final manual phase can be reduced by improving the output parametrisation. A combination of ClQ with another parameter

such as the open quotient, speed quotient, or normalised amplitude quotient, would reduce the observed ambiguity in the above mentioned problem region. In contrast to LF-pulse based models (Fant et al. 1985), however, these parameters cannot be assumed to be independent in general. There appears to be a linear relationship between ClQ and OQ in our model although this relationship seems to depend in some cases on the tuning parameter values. Similar dependence is also true for speed quotient and normalised amplitude quotient.

One of the problems with quotient parameters such as ClQ and OQ is that they only utilise durations of different parts of the pulses, and hence cannot detect abnormalities that occur between the time points. In order to rule out such occurrences, the target pulse needs to be described more fully, for example as an LF-pulse which can then be matched, say, in the least squares sense.

The more fully determined the target pulse is, the more expertise setting it up requires. Furthermore, using a detailed target pulse increases the risk of over-fitting: the model could match the target pulse to a higher level of detail than to which the target is accurate, and other details, such as the impact of the VT, could be optimised out entirely making use of the coupled model redundant.

Another alternative is to use a measured pulse such as a glottal area pulse obtained by high-speed video (see e.g. methods use by Pinheiro et al. 2012; Yang et al. 2011) or an inverse filtered flow pulse (as used e.g. by Gómez-Vilda et al. 2007). Of course, this does not completely remove the possibility of over-fitting, although the major concerns would shift from model accuracy to noise and mismatch between pulse and VT data. The latter issue arises because it would be difficult to perform the measurements required for pulse approximation simultaneously with MR imaging of the VT.

Chapter 5

Conclusion

In this work, an existing biomechano-acoustic model for vowel synthesis (Aalto 2009) has been revised. The aim has been to modify the model so that it can be used as a tunable glottal pulse generator for a 3D acoustic simulator based on the wave equation and environmental acoustic models. Tunability of the model has been achieved by selecting suitable control parameters and developing methods for finding values for these parameters such that the produced glottal pulses have certain target characteristics. In this final chapter, the work done is summarised (Section 5.1) and avenues for further work are discussed (Section 5.2).

5.1 Summary

New elements have been introduced to the model and their impact on simulation output, in particular the glottal flow pulses, has been investigated. It was found that the impact of tissue losses along the vocal tract is negligible.

A rough estimate of the losses caused by turbulence in the glottis was also obtained and noted to be significant at large glottal openings and flows. Adding this estimate to the viscous losses in the glottal flow model resulted in pulses with lower fundamental frequency and increased breathiness.

The impact of adding a subglottal tract resonator in parallel with the vocal tract resonator is also noticeable when a high level of feedback is used. The subglottal tract has a tendency to increase the fundamental frequency of phonation and to make phonation more breathy. Some distortion in the glottal flow and area pulses can also be observed.

Other changes made to the model are more practical in nature. The revised model can take any vocal tract geometry in the form of a centreline and an area function as an input. For this work, four geometries corresponding to Finnish vowels [ɑ, i , u] and [œ] extracted from magnetic resonance imaging data were

used. The second major change is the automated search for the critical glottal damping coefficient.

Besides changes made to the model, the second topic addressed in this thesis is how the model can be tuned to obtain glottal pulses that meet the requirements. This is essentially a question of how to find values for selected tuning parameters, so that the glottal pulses produced by the model exhibit chosen characteristics. In this work, fundamental frequency and mode of phonation, as parametrised by the closing quotient, have been used as target characteristics.

A three phase strategy is used to solve this problem. First, optimisation is carried out in the tuning parameter space to find points where the fundamental frequency of the simulated pulses matches the target. This optimisation is carried out using pattern search. Second, exploration is started from each optimal point found in the previous phase. This exploration makes use of a pattern search-type algorithm that only accepts points where the fundamental frequency matches the target. Cycles are prevented by removing already explored directions from polling. In the final phase, pulse shape is chosen manually using the closing quotient as a guideline.

This strategy was tested and found to work acceptably well. Its main weaknesses were found to lie in long computation times unless the process is parallelised and in the use of the closing quotient alone to parametrise mode of phonation.

5.2 Further work

Further work regarding details of the vowel synthesis model or its tuning methods has been discussed together with results that prompted it at the end of Chapters 3 and 4, respectively. In addition, there are more general avenues requiring further work.

For example, before the model can be used as a glottal source signal generator for a large number of vocal tract geometries, data management issues, such as which results are saved and in what format, need to be addressed. The solutions, in turn, depend on how the model is used. If only a few discrete fundamental frequencies are of interest for a vocal tract geometry, but the desired phonation type may vary at each frequency, it makes sense to save the exploration results. The same saving method wastes a lot of space and is of little value, if one is interested in a single mode of phonation but a large range of fundamental frequencies.

The prospect of using a large number of vocal tract geometries extracted from different individuals also leads to asking to what extent results are translatable, say, between different individuals producing the same vowel. Individual differences in the vocal tract geometries are expected to affect the glottal pulse shape, else the use of a coupled model as a pulse generator would be redundant. This does

not preclude a set of tuning parameters from producing similar output in different individuals, particularly if the output is parametrised with few parameters.

The same question can be asked about translatability of results between different vocal tract geometries from the same subject. Both the methods and the geometries exist already, so that all that is needed is selecting a suitable data set, running the simulations and analysing the results.

It would also be interesting to compare the output of the model to glottal signals obtained using other methods, such as high-speed imaging, electroglottography, or inverse filtering. While such studies have been carried out before, the model presented in this study is particularly well suited for studying the interaction between vocal folds and vocal tract using a large data set of realistic vocal tract geometries.

In fact, the main outcome of this work is not so much increased understanding of vowel production, *per se*, but rather a tool that can be used for studying it. There are still numerous open questions in phonetics and other related fields for which this model, with or without the addition of the 3D acoustic simulator, may be able to shed some light.

Appendix A

Discretisation of Webster's equation

The discretisation of Webster's equation used in this work follows the steps taken by Aalto (2009). The difference here is the addition of the dissipation term, which is followed closely in the following.

Spatial discretisation is done using Finite Element Method. First, the weak formulation of the problem is found using an auxiliary function $\varphi = \rho\psi_t$ and defining $\mathbf{X} = \frac{1}{A(s)}\frac{\partial}{\partial s}\left(A(s)\frac{\partial}{\partial s}\right)$. Using these, Webster's equation may be written a system of first order differential equations

$$\frac{\partial}{\partial t}\begin{bmatrix}\psi\\ \varphi\end{bmatrix} = \begin{bmatrix}0 & \rho^{-1}\\ \rho c^2\Sigma^2\mathbf{X} & -\frac{2\pi\alpha W}{A}\end{bmatrix}\begin{bmatrix}\psi\\ \varphi\end{bmatrix}. \tag{A.1}$$

So adding dissipation changes one element in $\mathbf{L} = \begin{bmatrix}0 & \rho^{-1}\\ \rho c^2\Sigma^2\mathbf{X} & -\frac{2\pi\alpha W}{A}\end{bmatrix} : Z \to X$, where

$$Z := \left(H^1(0, L_{vt}) \cap H^2(0, L_{vt})\right) \times H^1(0, L_{vt}), \tag{A.2}$$

$$X := H^1(0, L_{vt}) \times L^2(0, L_{vt}); \tag{A.3}$$

and the Hilbert space X is equipped with the inner product corresponding to the physical energy norm

$$\left\langle \begin{bmatrix}y_1\\ y_2\end{bmatrix}, \begin{bmatrix}x_1\\ x_2\end{bmatrix} \right\rangle_X = \frac{1}{2}\left(\rho\int_0^{L_{vt}} y_1'(s)x_1'(s)A(s)ds + \frac{1}{\rho c^2}\int_0^{L_{vt}} y_2(s)x_2(s)\frac{A(s)}{\Sigma(s)^2}ds\right). \tag{A.4}$$

The vocal tract model may be written as a linear boundary control system without change

$$\begin{cases} \dot{z}(t) &= \mathbf{L}z(t) \\ \mathbf{G}z(t) &= \begin{bmatrix} c_1 v_o(t) \\ 0 \end{bmatrix}, \\ p_c(t) &= \mathbf{H}z(t) \\ z(0) &= z_0 \end{cases} \tag{A.5}$$

where the end point and observation operations are defined as

$$\mathbf{G}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} := \begin{bmatrix} -z_1'(0) \\ z_2(L_{vt}) + \theta c z_1'(L_{vt}) \end{bmatrix} \qquad \text{and} \qquad \mathbf{H}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} := z_2(0), \tag{A.6}$$

where $(z_1, z_2)^T \in Z$.

The weak formulation of Webster's equation is found by computing the inner product A.4 of the first line of A.5 with a test functions $(v(s), 0)^T \in X$ and $(0, v(s))^T \in X$, i.e.

$$\left\langle \begin{bmatrix} \psi_t \\ \varphi_t \end{bmatrix}, \begin{bmatrix} v(s) \\ 0 \end{bmatrix} \right\rangle_X = \left\langle \mathbf{L}\begin{bmatrix} \psi \\ \varphi \end{bmatrix}, \begin{bmatrix} v(s) \\ 0 \end{bmatrix} \right\rangle_X, \tag{A.7}$$

leading to

$$\frac{\rho}{2}\int_0^{L_{vt}} \psi_{ts}(t,s) v_s(s) A(s) ds \;=\; \frac{1}{2}\int_0^{L_{vt}} \varphi_s(t,s) v_s(s) A(s) ds, \tag{A.8}$$

and similarly

$$\left\langle \begin{bmatrix} \psi_t \\ \varphi_t \end{bmatrix}, \begin{bmatrix} 0 \\ v(s) \end{bmatrix} \right\rangle_X = \left\langle \mathbf{L}\begin{bmatrix} \psi \\ \varphi \end{bmatrix}, \begin{bmatrix} 0 \\ v(s) \end{bmatrix} \right\rangle_X, \tag{A.9}$$

leading to

$$\begin{array}{l} \frac{1}{2}\frac{1}{\rho c^2}\int_0^{L_{vt}} \varphi_t(s,t) v(s) \frac{A(s)}{\Sigma(s)^2} ds = \frac{1}{2}\Big|_0^{L_{vt}} A(s)\psi_s(s,t)v(s) \\ -\frac{1}{2}\int_0^{L_{vt}} A(s)\psi_s(s,t)v_s(s)ds - \frac{\pi\alpha}{\rho c^2}\int_0^{L_{vt}} \varphi(s,t)v(s)\frac{W(s)}{\Sigma(s)^2} ds. \end{array} \tag{A.10}$$

Notice that the dissipation term appears in the last term of A.10.

Now, instead of general test functions $v(s)$ in A.7 and A.9, piecewise linear basis functions (hat function) $v_j(s)$, $j = 1, ..., N$ on each element of the vocal tract are used. Hence the approximate solution we are looking for is of the form

$$\begin{bmatrix} \psi \\ \varphi \end{bmatrix} = \sum_{j=1}^{N+1} \left(\xi_j(t) \begin{bmatrix} v_j(s) \\ 0 \end{bmatrix} + \mu_j(t) \begin{bmatrix} 0 \\ v_j(s) \end{bmatrix} \right). \tag{A.11}$$

This will give us, instead of A.8 and A.10, a set of equations which can be written in matrix form as

$$\begin{cases} \rho\dot{\xi}(t) = \mu(t) \\ \mathbf{M}\dot{\mu}(t) = -\mathbf{K}\xi(t) - \mathbf{R}\mu(t) + \mathbf{b}(t) \end{cases}, \tag{A.12}$$

where

$$\begin{aligned} \mathbf{M}_{ij} &= \tfrac{1}{2\rho c^2}\textstyle\int_0^{L_{vt}} v_i(s)v_j(s)\tfrac{A(s)}{\Sigma(s)^2}ds, \\ \mathbf{K}_{ij} &= \tfrac{1}{2}\textstyle\int_0^{L_{vt}} v_i'(s)v_j'(s)A(s)ds, \\ \mathbf{R}_{ij} &= \begin{cases} \frac{\pi\alpha}{\rho c^2}\int_0^{L_{vt}} v_i(s)v_j(s)\frac{W(s)}{\Sigma(s)^2}ds + \frac{A_m}{2\theta\rho c}, & \text{when } i=j=N; \\ \frac{\pi\alpha}{\rho c^2}\int_0^{L_{vt}} v_i(s)v_j(s)\frac{W(s)}{\Sigma(s)^2}ds, & \text{otherwise,} \end{cases} \\ \mathbf{b}_j &= \begin{cases} \frac{H_1 h}{2}v_o(t), & \text{when } j=1; \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{A.13}$$

Here $A_m = A(L_{vt})$ is the mouth opening area and $H_1 h$ is the area of the control surface next to the glottis. Adding tissue losses along the vocal tract has caused changes in the elements of the dissipation matrix $\mathbf{R}$ only.

Finally, time discretisation is done using the Crank-Nicolson method. Denoting $\xi^n \approx \xi(t_n)$ and $\mu^n \approx \mu(t_n)$ to obtain the update equations

$$\begin{cases} (\frac{\Delta t}{2}\mathbf{K} + \frac{2\rho}{\Delta t}\mathbf{M} + \rho\mathbf{R})\xi^n = (-\frac{\Delta t}{2}\mathbf{K} + \frac{2\rho}{\Delta t}\mathbf{M} + \rho\mathbf{R})\xi^{n-1} + 2\mathbf{M}\mu^{n-1} + \Delta t\mathbf{b}(t_n) \\ \rho\xi^n - \frac{\Delta t}{2}\mu^n = \rho\xi^{n-1} + \frac{\Delta t}{2}\mu^{n-1} \end{cases}. \tag{A.14}$$

Bibliography

A. Aalto. A low-order glottis model with nonturbulent flow and mechanically coupled acoustic load. Master's thesis, Helsinki University of Technology, Institute of Mathematics, 2009.

A. Aalto and J. Malinen. Composition of passive boundary control systems. *Mathematical Control and Related Fields*, 3(1):1–19, 2013. doi: 10.3934/mcrf.2013.3.1.

A. Aalto, P. Alku, and J. Malinen. A LF-pulse from a simple glottal flow model. In *Proceedings of the 6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA2009)*, pages 199–202, Firenze, Italy, December 2009.

A. Aalto, D. Aalto, J. Malinen, and M. Vainio. Modal locking between vocal fold and vocal tract oscillations. *ArXiv e-prints*, November 2012.

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

D. Aalto, A. Huhtala, A. Kivelä, J. Malinen, P. Palo, J. Saunavaara, and M. Vainio. How far are vowel formants from computed vocal tract resonances? *ArXiv e-prints*, August 2012.

D. Aalto, J. Helle, A. Huhtala, A. Kivelä, J. Malinen, J. Saunavaara, and T. Ronkka. Algorithmic surface extraction from MRI data: modelling the human vocal tract. In *Proceedings of BIODEVICES 2013*, 2013.

D. Aalto, O. Aaltonen, R.-P. Happonen, P. Jääsaari, A. Kivelä, J. Kuortti, J.-M. Luukinen, J. Malinen, T. Murtola, R. Parkkola, J. Saunavaara, T. Soukka, and M. Vainio. Large scale data acquisition of simultaneous {MRI} and speech. *Applied Acoustics*, 83:64 – 75, 2014. doi: 10.1016/j.apacoust.2014.03.003.

R. Adams and L. Bischof. Seeded region growing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(6):641–647, 1994. doi: 10.1109/34.295913.

F. Alipour, D. A. Berry, and I. R. Titze. A finite-element model of vocal-fold vibration. *The Journal of the Acoustical Society of America*, 108(6):3003–3012, 2000. doi: 10.1121/1.1324678.

P. Alku, T. Bäckström, and E. Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *The Journal of the Acoustical Society of America*, 112(2):701–710, 2002. doi: 10.1121/1.1490365.

P. Birkholz, D. Jackel, and B. Kroger. Simulation of losses due to turbulence in the time-varying vocal system. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(4):1218–1226, 2007. doi: 10.1109/TASL.2006.889731.

Comspeech@Aalto. Mathematical and numerical modelling of speech, 2013. URL `http://speech.math.aalto.fi/about.html`.

D. J. Daily and S. L. Thomson. Acoustically-coupled flow-induced vibration of a computational vocal fold model. *Computers & Structures*, 116:50 – 58, 2013. doi: 10.1016/j.compstruc.2012.10.022.

M. de Oliveira Rosa, J. C. Pereira, M. Grellet, and A. Alwan. A contribution to simulating a three-dimensional larynx model using the finite element method. *The Journal of the Acoustical Society of America*, 114(5):2893–2905, 2003. doi: 10.1121/1.1619981.

M. P. de Vries, H. K. Schutte, and G. J. Verkerke. Determination of parameters for lumped parameter models of the vocal folds using a finite-element method approach. *The Journal of the Acoustical Society of America*, 106(6):3620–3628, 1999. doi: 10.1121/1.428214.

M. P. de Vries, H. K. Schutte, A. E. P. Veldman, and G. J. Verkerke. Glottal flow through a two-mass model: Comparison of navier–stokes solutions with simplified models. *The Journal of the Acoustical Society of America*, 111(4): 1847–1853, 2002. doi: 10.1121/1.1323716.

M. Döllinger, U. Hoppe, F. Hettlich, J. Lohscheller, S. Schuberth, and U. Eysholdt. Vibration parameter extraction from endoscopic image series of the vocal folds. *Biomedical Engineering, IEEE Transactions on*, 49(8):773–781, 2002. doi: 10.1109/TBME.2002.800755.

G. Fant, J. Liljencrants, and Q.-g. Lin. A four-parameter model of glottal flow. *STL-QPSR*, pages 1–13, 1985.

J. Flanagan and L. Landgraf. Self-oscillating source for vocal-tract synthesizers. *Audio and Electroacoustics, IEEE Transactions on*, 16(1):57–64, 1968. doi: 10.1109/TAU.1968.1161949.

J. L. Flanagan. *Speech Analysis Synthesis and Perception*. Springer-Verlag, 1972.

L. P. Fulcher, R. C. Scherer, and T. Powell. Pressure distributions in a static physical model of the uniform glottis: Entrance and exit coefficients. *The Journal of the Acoustical Society of America*, 129(3):1548–1553, 2011. doi: 10.1121/1.3514424.

P. Gómez-Vilda, R. Fernández-Baillo, A. Nieto, F. Díaz, F. Fernández-Camacho, V. Rodellar, A. Álvarez, and R. Martínez. Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters. *Journal of Voice*, 21 (4):450–476, 2007. doi: 10.1016/j.jvoice.2006.01.008.

R. B. Gramacy, H. K. H. Lee, and W. G. Macready. Parameter space exploration with gaussian process trees. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 45–53, New York, NY, USA, 2004. ACM. doi: 10.1145/1015330.1015367.

A. Hannukainen, T. Lukkari, J. Malinen, and P. Palo. Vowel formants from the wave equation. *The Journal of the Acoustical Society of America*, 122(1):EL1–EL7, 2007. doi: 10.1121/1.2741599.

J. C. Ho, M. Zañartu, and G. R. Wodicka. An anatomically based, time-domain acoustic model of the subglottal system for speech production. *The Journal of the Acoustical Society of America*, 129(3):1531–1547, 2011. doi: 10.1121/1.3543971.

J. Horáček and J. G. Švec. Aeroelastic model of vocal-fold-shaped vibrating element for studying the phonation threshold. *Journal of Fluids and Structures*, 16(7):931 – 955, 2002. doi: 10.1006/jfls.2002.0454.

J. Horáček, P. Šidlof, and J. G. Švec. Numerical simulation of self-oscillations of human vocal folds with hertz model of impact forces. *Journal of Fluids and Structures*, 20(6):853 – 869, 2005. doi: 10.1016/j.jfluidstructs.2005.05.003.

L. Ingber. Adaptive simulated annealing (ASA): Lessons learned. *Control and Cybernetics*, 25(1):33–54, 1996.

K. Ishizaka and J. L. Flanagan. Synthesis of voiced sounds from a two mass model of the vocal cords. *Bell System Technical Journal*, 51:1233–1268, 1972.

K. L. Kelly and C. C. Lochbaum. Speech synthesis. In *Proceedings of the Fourth International Congress on Acoustics, Paper G42*, pages 1–4, 1962.

L. Lehto, M. Airas, E. Björkner, J. Sundberg, and P. Alku. Comparison of two inverse filtering methods in parameterization of the glottal closing phase characteristics in different phonation types. *Journal of Voice*, 21(2):138–150, 2007.

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

N. Lous, G. Hofmans, R. Veldhuis, and A. Hirschberg. A symmetrical two-mass vocal-fold model coupled to vocal tract and trachea, with application to prosthesis design. *Acta Acustica united with Acustica*, 84(6):1135–1150, 1998.

C. Lu, T. Nakai, and H. Suzuki. Finite element simulation of sound transmission in vocal tract. *Journal of the Acoustical Society of Japan (E)*, 14(2):63–72, 1993.

T. Lukkari and J. Malinen. Webster's equation with curvature and dissipation. *Submitted*, apr 2013.

J. Malinen and O. J. Staffans. Impedance passive and conservative boundary control systems. *Complex Analysis and Operator Theory*, 1:279–300, 2007.

R. T. Marler and J. S. Arora. Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization*, 26(6):369–395, 2004.

M. D. McKay, R. J. Beckman, and W. J. Conover. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979. doi: 10.1080/00401706.1979.10489755.

P. Mokhtari, H. Takemoto, and T. Kitamura. Single-matrix formulation of a time domain acoustic model of the vocal tract with side branches. *Speech Communication*, 50(3):179–190, 2008.

P. M. Morse and K. U. Ingard. *Theoretical acoustics*. McGraw-Hill, 1968.

A. P. Pinheiro and G. Kerschen. Vibrational dynamics of vocal folds using nonlinear normal modes. *Medical Engineering & Physics*, 35(8):1079 – 1088, 2013. doi: 10.1016/j.medengphy.2012.11.002.

A. P. Pinheiro, D. E. Stewart, C. D. Maciel, J. C. Pereira, and S. Oliveira. Analysis of nonlinear dynamics of vocal folds using high-speed video observation and biomechanical modeling. *Digital Signal Processing*, 22(2):304 – 313, 2012. doi: 10.1016/j.dsp.2010.11.002.

R. Schwarz, U. Hoppe, M. Schuster, T. Wurzbacher, U. Eysholdt, and J. Lohscheller. Classification of unilateral vocal fold paralysis by endoscopic digital high-speed recordings and inversion of a biomechanical model. *Biomedical Engineering, IEEE Transactions on*, 53(6):1099–1108, 2006. doi: 10.1109/TBME.2006.873396.

D. Scimarella and C. d'Alessandro. On the acoustic sensitivity of a symmetric two-mass model of the vocal folds to the variation of control parameters. *Acta Acoustica united with Acoustica*, 90:746–761, 2004.

H. Suzuki, T. Nakai, N. Takahashi, and A. Ishida. Simulation of vocal tract with three-dimensional finite element method. Technical report, IEICE EA93-8, 1993.

P. Švancara, J. Horáček, and L. Pešek. Numerical modelling of production of Czech vowel /a/ based on FE model of the vocal tract. In *Proceedings of International Conference on Voice Physiology and Biomechanics*, 2004.

I. R. Titze. On the relation between subglottal pressure and fundamental frequency in phonation. *The Journal of the Acoustical Society of America*, 85(2):901–906, 1989. doi: 10.1121/1.397562.

I. R. Titze. Nonlinear source-filter coupling in phonation: Theory. *The Journal of the Acoustical Society of America*, 123(5):2733–2749, 2008. doi: /10.1121/1.2832337.

T. Vampola, J. Horáček, and J. G. Švec. FE modeling of human vocal tract acoustics. Part I: Production of Czech vowels. *Acta Acustica united with Acustica*, 94 (3):433–447, 2008. doi: doi:10.3813/AAA.918051.

J. van den Berg, J. T. Zantema, and J. P. Doornenbal. On the air resistance and the bernoulli rffect of the human larynx. *The Journal of the Acoustical Society of America*, 29(5):626–631, 1957. doi: 10.1121/1.1908987.

K. van den Doel and U. M. Ascher. Real-time numerical solution of Webster's equation on a nonuniform grid. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(6):1163–1172, 2008.

T. Wurzbacher, R. Schwarz, M. Döllinger, U. Hoppe, U. Eysholdt, and J. Lohscheller. Model-based classification of nonstationary vocal fold vibrations.

The Journal of the Acoustical Society of America, 120(2):1012–1027, 2006. doi: 10.1121/1.2211550.

A. Yang, M. Stingl, D. A. Berry, J. Lohscheller, D. Voigt, U. Eysholdt, and M. Döllinger. Computation of physiological human vocal fold parameters by mathematical optimization of a biomechanical model. *The Journal of the Acoustical Society of America*, 130(2):948–964, 2011. doi: 10.1121/1.3605551.

A. Yang, D. A. Berry, M. Kaltenbacher, and M. Döllinger. Three-dimensional biomechanical properties of human vocal folds: Parameter optimization of a numerical model to match in vitro dynamics. *The Journal of the Acoustical Society of America*, 131(2):1378–1390, 2012. doi: 10.1121/1.3676622.