Voice Synthesis Using the Three-Dimensional Digital Waveguide Mesh

Matthew David Adam Speed

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Electronics
University of York

March, 2012

Abstract

The acoustic response of the vocal tract is fundamental to our interpretation of voice production. As an acoustic filter, it shapes the spectral envelope of vocal fold vibration towards resonant modes, or formants, whose behaviours form the most basic building blocks of phonetics.

Physical models of the voice exploit this effect by modelling the nature of wave propagation in abstracted cylindrical constructs. Whilst effective, the accuracy of such approaches is limited due to their limited geometrical analogue. Developments in numerical acoustics modelling meanwhile have seen the formalisation of higher dimensionality configurations of the same technologies, allowing a much closer geometrical representation of an acoustic field. The major focus of this thesis is the application of such a technique to the vocal tract, and comparison of its performance with lower dimensionality approaches.

To afford the development of such models, a body of data is collected from Magnetic Resonance Imaging for a range of subjects, and procedures are developed for the decomposition of this imaging into suitable, efficient data structures for simulation. The simulation technique is exhaustively validated using a combination of bespoke measurement/inversion techniques and analytical determination of lower frequency behaviours.

Finally, voice synthesis based on each numerical model is compared with acoustic recordings of the subjects involved and with equivalent simulations from lower dimensionality methods. It is found that application of a higher dimensionality method typically yields a more accurate frequency-domain representation of the voice, although in some cases lower dimensionality equivalents are seen to perform better at low frequencies.

Contents

List of Tables

List of Figures

Acknowledgements

I have been very fortunate to have two extraordinary supervisors in the course of this thesis. Dr Damian Murphy and Prof. David Howard have both provided inspiration and enthusiasm without fail. This project would not have been possible without their experience and knowledge, and the willingness with which they have shared it. I am immensely grateful for all they have done to support me during my studies.

I'd like to thank all the staff and students, past and present, of the Audio Lab in York for their company, good humour and of course cake, biscuits and cups of tea. I've enjoyed several happy years here, and I'm sure the lab will continue to go from strength to strength.

I'm grateful to Sten Ternström at KTH, Stockholm, for inspiring my interest in voice research, and also to Olov Engwall, who kindly made his own MRI data available to help me get to grips with this project.

I am indebted to those who have given up their own time to participate in this experiment, and also grateful to Ross Devlin at the York Neuroimaging Centre for his expertise and patience at the helm of the MRI scanner.

I would like to express my sincere thanks to my parents, Jenny and David, for their support throughout my studies. They have made everything possible. Finally, I'd like to say a huge thankyou to Maria, for her love, support and fantastic cooking.

Declaration

I hereby declare that this thesis is entirely my own work and all contributions from outside sources, through direct contact or publications, have been explicitly stated and referenced. I also declare that some parts of this program of research have been presented previously, at conferences and in journals. These publications are listed as follows:

- **Three-Dimensional Digital Waveguide Mesh Simulation of Cylindrical Vocal Tract Analogs**, Matt Speed, Damian Murphy, David Howard, submitted journal paper to *IEEE Transactions on Audio, Speech and Language Processing*, in review.
- **Acoustic Coupling in Multi-Dimensional Finite Difference Schemes for Physically Modeled Voice Synthesis**, Matt Speed, Damian Murphy, David M Howard, paper and oral presentation at the *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, October 2009.
- **Characteristics of Two-Dimensional Finite-Difference Techniques for Vocal Tract Analysis and Voice Synthesis**, Matt Speed, Damian Murphy, David M Howard, paper and poster presentation at *ISCA Interspeech 2009*, Brighton, UK, September 2009.
- **Natural Voice Synthesis: The potential relevance of high-frequency components**, David M Howard, Sten Ternström, Matt Speed, paper at *3rd Advanced Voice Function Assessment International Workshop (AVFA 2009)*, Madrid, Spain, May 2009.

Chapter 1

Introduction

1.1 Interest

> It would be a considerable invention indeed, that of a
> machine able to mimic speech, with its sounds and articulations.
> I think it is not impossible.
>
> *Leonhard Euler 1761*

Given man's dependence on the vocal anatomy for communication, it is perhaps unsurprising that he has long sought to better understand and reproduce its form and function. Despite this, the complexity of the mechanism and its acoustic productions are such that much remains unknown.

Acoustic phonetics is a field of research quite predictably focussed on the interrelationship between phonetics and consequent acoustic behaviours. Much of our current understanding is informed by the decomposition of the voice into smaller lumped elements of more easily calculable acoustic performance. This defines an articulatory-acoustic transform, whereby broad changes in the anatomical configuration are seen to exhibit predictable, characteristic changes in the acoustic response. Our most basic understanding of this transform is perhaps by the definition of formants. These are low frequency resonant modes of the vocal tract, seen to change for different articulatory configurations. The centre frequency of the first two formants

governs the listener's interpretation of the sound as a particular vowel. The third and fourth formants are consequently considered to affect the vowel timbre. These relationships can be expanded by the development of a more complete acoustical analogue, considering the impact of branched resonators to describe the coupling of the nasal tract for example.

Analytical models as basic as these will always constitute an anthropomorphically poor representation of the complete vocal anatomy. While such approaches can be effective in terms of vowel identification, they are rarely convincingly natural. Beyond a certain complexity the analytical determination of resonant mode frequencies becomes particularly difficult, which complicates the objective assessment of performance.

Physical modelling encompasses an approach to representation of the vocal tract whereby the acoustic behaviours responsible for the formation of resonant modes are directly reproduced. The first of such models were one-dimensional, constituting the representation of the vocal tract as a chain of cylinders of changing cross-sectional area. Axial acoustic behaviours are then reproduced by modelling wave propagation under the assumption of planar wave propagation. While these models are able to demonstrate accurate reproduce formant behaviours, the approximation represented by a simple one-dimensional representation limits the frequency-domain accuracy of the simulation to the first two formants alone.

Consequent approaches have extended the modelling technique to higher dimensionalities, allowing a more complete reproduction of the acoustic field. These techniques are based on the development of vocal tract analogues which are significantly simplified in terms of their anatomical accuracy to achieve a comprehensible output within the bounds of reasonable computational loading. This thesis considers an extension to these techniques, whereby a full three-dimensional simulation of the acoustic field in the vocal tract is performed. This will provide as accurate a geometrical match as possible to each configuration of the vocal tract.

The long-term goal of this research is the development of a more accurate representation of the acoustic correlates of vocal tract articulation (through a dynamic physical model). The subject is hence the suitability of

three-dimensional simulation to the vocal tract in its function as a resonator. Other physical models have aimed to represent the salient traits of the vocal tract such as changing vocal tract cross-sectional areas, whilst omitting detailed geometrical features, and/or assuming symmetry about the midsaggital axis to reduce model complexity. The study will address whether full three-dimensional simulation results in adequate improvement to the accuracy of the resulting synthesised voice to justify its additional computational load.

1.2 Hypothesis

Statement of Hypothesis

> Time-domain acoustical simulation of the vocal tract using a three-dimensional digital waveguide mesh can result in more accurate voice reproduction than lower dimensionality equivalents.

Decomposition of Hypothesis

Time-Domain Acoustical Simulation

Time-domain variable scattering techniques are well suited to physical modelling of the vocal tract since they have been shown to be capable of real-time synthesis. Additionally, there is potential for changes to the shape of the tract to be made in real-time, with consequent acoustic changes.

The Vocal Tract

Developing an accurate model of the vocal tract demands an accurate geometrical representation. Recent voice research has made extensive use of Magnetic Resonance Imaging (MRI) to collect three-dimensional data regarding the vocal anatomy. Ever improving MRI technologies have lead to reduced capture times and increased resolution in imaging. In this study,

a body of data will be gathered using the latest MRI technologies, providing accurate three-dimensional imaging of the vocal anatomy under varying articulatory configurations. This corpus will be used to develop accurate numerical acoustical models of the vocal tract.

Three-Dimensional Digital Waveguide Mesh

While not the only means of performing time-domain scattering-based numerical acoustics simulation, the digital waveguide mesh is particularly attractive to voice synthesis due to its potential for dynamic manipulation via impedance mapping. Previous studies have used the digital waveguide and two-dimensional digital waveguide mesh to model the acoustics of the vocal tract. The current hypothesis will be tested using a static, uniform three-dimensional digital waveguide mesh using standard wave-variable formulations alone.

Voice Reproduction

The particular advantage of time-domain modelling lies in the opportunities it presents for real-time voice synthesis, using a dynamic representation of the vocal tract. The first stage in exploring this capability is in determining the ability of such a technique to reliably reproduce the acoustic response of a stationary vocal tract. Simulation here focusses on reproduction of the vocal tract transfer function, incorporating the effect of lip radiation but using basic reflective methodologies. A nasal tract model is not incorporated and the simulation operates under the assumption of linear separability of the vocal source and tract.

The consequent simulation is assessed in terms of its frequency-domain characterisation of human phonation. Where simulation demonstrates traits closely matched to that of a human speaker, the technique is determined to constitute a good representation of the natural voice production system.

1.3 Motivation

The natural trend of physical modelling is towards a more complete representation of a given system. In the vocal tract this includes progression towards a higher dimensionality representation and reducing the approximation inherent to a geometrical analogue. A fundamental compromise which demands consideration is that between the complexity of the model and the resulting computational load. This study seeks to address whether there is value in further increasing the dimensionality of a vocal tract analogue. If this is the case, attention can be turned to techniques for reduction of the computational load and bringing the model towards feasibility for real-time speech synthesis. If three-dimensional representation has little improvement to offer above comparable one and two-dimensional models, attention can instead be turned towards improving lower-complexity analogues.

1.4 Thesis Outline

The thesis is organised as follows:

Chapter 2 - Acoustics

Chapter 2 begins with formal definitions of applicable concepts in acoustics. The nature of wave behaviour is described, followed by a development of the acoustics of simple cylinders and the introduction of the wave equation.

Chapter 3 - Time-Domain Acoustics Simulation

Chapter 3 explores techniques for the time-domain numerical simulation of acoustic wave propagation. The digital waveguide mesh is introduced, in addition to further techniques for the manipulation of sampling grids. Methods employed for boundary modelling and injection are explored. The chapter concludes with a treatment of the numerical dispersion error encountered

in numerical simulation and an assessment of computational cost in these simulation schemes.

Chapter 4 - Human Voice

In Chapter 4, the human voice is introduced and explored in terms of its anatomy. In addition, a simple treatment of the processes inherent to voice production is presented. The chapter continues to consider the contribution of the source and nasal cavity to the vocal process. An initial exploration of means for practical voice measurement is then performed.

Chapter 5 - Voice Modelling

Once the voice has been introduced in Chapter 4, this chapter continues to explore existing approaches to its numerical modelling and more specifically reproduction of the vocal tract transfer function. Particular attention is paid to the development of digital-waveguide based methodologies, including the Kelly-Lochbaum model and the digital waveguide mesh. The recent application of impedance mapping to the two-dimensional digital waveguide mesh is considered before a brief exploration of alternative numerical models of voice production.

Chapter 6 - 3D DWM Simulation of the Vocal Tract

This chapter describes the techniques used for the development of a three-dimensional digital waveguide mesh model of the vocal tract. It begins by describing the processes used for magnetic resonance imaging capture then continues to describe how the collected data is processed and manipulated to generate a sampling grid. The computational implementation of the numerical technique is then explored, before the techniques used to generate lower-dimensionality derivative simulations are introduced.

Chapter 7 - Benchmarking and Validation

Chapter 7 encompasses the testing and validation of the techniques introduced and developed in Chapter 6. Simulations are performed on structures of increasing geometrical complexity and compared with mathematical approximation and benchmark acoustic measurements. The chapter begins with a treatment of a simple cuboid before presenting a technique developed for broadband acoustic measurement of mechanical vocal tract analogues. Validation and measurement is then performed for cylinders, concatenated cylinders and complex vocal tract models in turn. The chapter concludes with a presentation of the approach taken to benchmark the frequency-domain consistency of phonations produced during MR scanning by supine acoustic measurement.

Chapter 8 - Results of Simulation

In Chapter 8 the results of simulation are presented, beginning with those for the three-dimensional case on a subject-by-subject basis. The results of derivative simulation are then explored, including one- and two-dimensional and impedance-mapped two-dimensional simulation. The chapter concludes with an appraisal of the influence of the source waveform on perceived naturalness, followed by a description of the collected body of data.

Chapter 9 - Summary and Conclusions

In Chapter 9 the results of this study are summarised, followed by reconsideration of the initial hypothesis. Potential areas for further research are then explored.

1.5 Contributions

The novel contributions of this thesis are as follows:

- Development of a three-dimensional digital waveguide mesh-based numerical model of the vocal tract based on MRI data.
- Validation of the application of the three-dimensional digital waveguide mesh against mechanical vocal tract analogues.
- A novel technique for broadband measurement of the acoustic response of mechanical vocal tract analogues.
- A corpus of MRI images specifically targeting phoneme reproduction in a range of trained subjects, including benchmark audio recordings.
- A comparison of three-dimensional, two-dimensional, impedance mapped and one-dimensional digital waveguide-based models of the vocal tract.

Chapter 2

Acoustics

2.1 Introduction

Voice communication is amongst our most fundamental faculties as humans, forming a link between the vocal anatomy and our highly specialised hearing apparatus. This connection is made through the domain of acoustics, the study of which encompasses the physics of audible vibration and wave propagation phenomena. The voice represents an acoustic instrument of remarkable sophistication. It offers an intuitive means of layered information encapsulation and broadcast based on manipulation of an air flow, anatomical articulation and consequent modification of acoustic behaviours. An appropriate understanding of acoustics is hence a pre-requisite for any study of the vocal tract.

In this chapter the fundamentals of acoustics are introduced, particularly reflecting on applications of relevance to structures closely matched to the vocal tract. Section 2.1.1 begins by considering the physical principles underpinning acoustics, including gaseous wave propagation. Section 2.2 then introduces the concept of acoustic impedance, explaining its relevance and application. In section 2.4 acoustical concepts with respect to simple cylindrical structures are explored. These are important due to the approximate physical analogue which they draw with the human voice production system. The meaning and application of the wave equation is introduced in section

2.5, before the transition between flow and acoustics is considered in section 2.6.

2.1.1 Gas Behaviours

Particle motion in a gas can be simplified to a model whereby each particle exhibits an effectively random behaviour. It will travel in an arbitrary direction, at a speed relative to its kinetic energy until collision with another particle causes a mutual change in velocity and potentially direction (according to the conservation of momentum).

The gas will have a density, ρ. This describes the number of particles within a given volume, V. If the volume were to consequently change, the density would hence change in sympathy. The random motion exhibited in a simplified gas model suggests that the gas will always move towards its lowest possible energy state, as a localised increase in particle velocity will induce distribution of this energy through expansive collision. This concept is termed entropy [14, 5], and provides the second law of thermodynamics in that all systems will act to minimise their potential energy.

The ambient pressure of a given gaseous volume, P, describes the average momentum of its particles in an effective entropic equalisation. In a system with a human listener, this describes the inaudible 'DC' condition.

Small fluctuations in air pressure, p carry with them a localised change in density, as particles are either driven together (compression) or separated (rarefaction). If each is considered on an individual basis this density is translated to a change in the effective volume of the particles. For example, if there are 5 particles of an entropically distributed gas in a geometrically defined volume, the effective volume of each particle will be 1/5 of the total volume. If the number of particles in the same geometric volume is doubled, each will represent only 1/10 of the total volume. The relationship between pressure and volume is described by the Bulk Modulus of a gas, K [15], defined as in (2.1) where V is the gaseous volume.

$$K = -V\frac{\partial P}{\partial V} \tag{2.1}$$

Increasing the temperature of a gas is equivalent to increasing the energy and hence particle velocity. This can change both its volume and pressure. The three are related through the Ideal Gas Law [5, 16], as in (2.2) where T is the temperature of the gas (in degrees Kelvin), N is the number of gas particles and k_B is the Boltzmann constant [14].

$$PV = Nk_BT \tag{2.2}$$

The same relationship is provided in (2.3) for temperature given in degrees Celsius.

$$PV = Nk_B(T_c + 273.15) \tag{2.3}$$

Gaseous media are fundamentally different to solids in that gas particles present no direct resistance to separation. In a solid this would be termed elasticity, defining a restoring force that acts to hold a medium together. Since air does not present such a resistance it will not support transverse wave motion, in which the forces driving the wave act in a direction perpendicular to the motion of the wave itself.

In air it is only localised changes in pressure which can support wave motion. These constitute compressions and subsequent rarefactions in air pressure, travelling outwards as a response to particle collisions and particle sparsity respectively. Localised distribution changes lead to gradients in pressure and a corresponding movement of particles. This is a longitudinal, or compressional wave, and is analogous to a spring. Whether compressed or expanded (and assuming its elastic limit is not exceeded), a free spring will always return to its lowest possible energy state and the same is true of air. Where a spring's motion is governed by its elasticity through Hooke's law, the motion of air is directly determined by the properties of the gas.

It is these properties which also govern the speed with which a wave can propagate in air. The Newton-Laplace equation [5] describes the relationship between wave propagation speed in a gas, its Bulk Modulus, and density ρ, as per 2.4 where c is the wave propagation speed.

$$c^2 = \frac{K}{\rho} \tag{2.4}$$

Since air cannot support transverse wave movement, shear force results in a net movement of particles and a consequent flow condition. This can be considered the acoustical DC condition.

2.2 Acoustic Impedance

Acoustic impedance is a conceptual unit describing the relationship between the potential and kinetic elements of the total energy in an acoustic wave. As with all simple harmonic motion moved from rest, the system oscillates between conditions of maximum potential or kinetic energies, according to restoring forces. The phase relationship between the two elements determines the manner in which the wave reacts to further interfering forces.

In acoustics there are two commonly used definitions of acoustic impedance, specific and characteristic. Both are equally valid, but the two cannot be simultaneously resolved. The definitions of the potential and kinetic components differ in each, as do several key assumptions about the nature of the acoustic wave. In whatever domain it is defined, acoustic impedance is the ratio of sound pressure (p), to the velocity of movement induced through the particular area over which that pressure acts (U):

$$Z = \frac{p}{U} \tag{2.5}$$

Specific

Specific acoustic impedance, z, is a very low level characteristic of the medium in which the acoustic wave travels. Completely independent of the medium's shape, volume or termination behaviours, it describes the ratio of the acoustic pressure (per unit area) to the resulting flow through that same area. Discounting changes in density and humidity (through wave speed) the specific acoustic impedance would remain constant in a given medium. It can

hence be described as the product of this medium volume density and wave speed [5], as in (2.6).

$$z = \rho c \tag{2.6}$$

Specific acoustic impedance is vital to modelling wave propagation in compressible systems, since under compression a change in volume density would result. This in turn would affect the phase relationship of the system components (pressure, velocity) and hence alter the nature of the wave behaviour itself. In this study the medium is assumed to be homogeneous and incompressible, hence specific acoustic impedance is uniform. Under these assumptions the specific acoustic impedance actually represents the *characteristic* acoustic impedance of the medium.

Characteristic

While specific acoustic impedance is normalised to a unit volume, characteristic acoustic impedance (Z) is defined as a function of cross-sectional area S [5], as in (2.7).

$$Z_c = \frac{\rho c}{S} \tag{2.7}$$

It is used to characterise a given component in a one-dimensional sense, assuming the cross-sectional area to account for the further two dimensions in a three-dimensional space. Since it is a one-dimensional approximation it is only ever valid under the assumption of planar wavefront propagation.

The use of characteristic acoustic impedance can greatly simplify mathematical approximation of the acoustic behaviour of a given structure. This in turn allows the response of the system to some driving function to be estimated, within the accuracy constraints of the assumption of planar wave propagation.

While specific acoustic impedance can be considered uniform under the assumption of incompressibility, characteristic acoustic impedance can be seen to change with the geometry of the medium. Such changes lead to

transitions in the phase relationship of pressure and volume velocity and will contribute to the overall acoustic response of the system.

2.3 Wave Behaviour

Acoustic pressure and velocity exist in a potential/kinetic exchange relationship similar to the voltage/current exchange in electrical circuits. For a steady-state condition the phase relationship between the two becomes fundamental to acoustic power delivery and resonant behaviour. For a plane-wave travelling in an infinite medium the pressure will maintain a $\frac{\pi}{2}$ radians phase lead over the volume velocity. This condition is known as quadrature, representing a quarter-cycle shift. The nature of reflection and transmission will affect this phase relationship, hence characteristic acoustic impedance is often complex valued as defined in (2.8) where r is the resistive (pure real) component of impedance and x is the reactive (imaginary) component.

$$Z_c = r + jx \tag{2.8}$$

The reactive component can introduce a phase lead or lag, depending on the capacitive/inductive elements of its behaviour.

Most common approaches to acoustic modelling are based on reproducing key principles of wave behaviour. These could include geometrical modelling of wave behaviours, simple mathematical approximation of changes in characteristic acoustic impedance or complex models incorporating detailed descriptions of the system physics. Multi-dimensional time-domain models can implicitly reproduce diffraction effects (see section 2.3) and in certain models more complicated boundary behaviours such as a frequency dependence and diffusion (see sections 3.6 and 2.3 respectively). Yet more complex frequency-domain approaches such as finite-volume modelling aim to reproduce the acoustic circumstances as closely as possible, including taking aerodynamics into account (see section 5.9).

Reflection and Transmission

The response of a wave when it meets a change in the impedance of a medium is fundamental to its overall behaviour. As in much of acoustics, many of these characteristics can be simplified under certain conditions. Reflection and transmission are functions of changes in acoustic impedance. Whenever such a change is encountered, a characteristic reflection and transmission of components occurs which is in turn a function of the relationship between the acoustic pressure and volume velocity.

The most simple conceivable model is perhaps that of a planar wave normally incident on a completely reflective wall. Assume for simplicity that the acoustic impedance of this wall is entirely real, infinite and constant for all frequencies. When a particle arrives at the boundary its velocity is constrained to zero since it cannot move through the boundary, or move the boundary itself. Without another force acting on the system the conservation of momentum must be observed.

$$|mu_\bullet||^-_{x=W} = |mu_\bullet||^+_{x=W} \tag{2.9}$$

$$u_\bullet|_{x=W} = 0 \tag{2.10}$$

The condition for conservation of momentum at the boundary is described by (2.9), where x is axial position, and the wall is at $x = W$. The volume velocity for a particular particle unit incident at the wall is represented by $u_\bullet$, and the superscripts $^+$ and $^-$ denote conditions immediately before and after collision. To simplify (2.9) assume that the total velocity before and after collision must be constant. Since particle velocity at the point of collision must equal zero (as expressed by (2.10)), the velocity must be equal but reversed to satisfy the conservation of momentum. This change in polarity constitutes a complete reflection. Reversal of momentum also implies a reversal of force, through (2.11).

$$F = ma = m\frac{dv}{dt} \tag{2.11}$$

This leads to an instantaneous doubling of pressure at the point for which velocity passes through $u = 0$ (or when the nominal particle is actually incident at the wall). No phase shift occurs after relaxation of this pressure doubling, hence such a reflection is phase preserving.

Next, consider transmission from one medium to another, across an infinite interface as in Fig. 2.1. The first medium has a finite, real and frequency independent impedance Z_1, and the second a similarly finite, real and frequency independent impedance Z_2, where $Z_2 > Z_1$. When a normally incident wave arrives at the interface, it is intuitive that some proportion of the wave will be transferred into the second medium, while another will be reflected back into the first medium. Reflection in these conditions is governed by two principles [5].

- Pressure is equal on both sides of the interface (so as not to exert a force)
- Total velocity is equal on both sides of the interface

The first principle can be described by (2.12), where p is the incident pressure, subscripts $_{Z_1}$ and $_{Z_2}$ denote the occupied medium and superscripts $^+$ and $^-$ describe wave directionality, as per Fig. 2.1.

$$p^+_{Z_1} + p^-_{Z_1} = p^+_{Z_2} \tag{2.12}$$

The second point can be similarly stated, as in (2.13) where the notation is the same as for (2.12), but u represents velocity.

$$u^+_{Z_1} + u^-_{Z_1} = u^+_{Z_2} \tag{2.13}$$

Define two coefficients, one for the fraction of transmission and the other for reflection, as per (2.14) and (2.15).

$$R = \frac{p^-_{Z_1}}{p^+_{Z_1}} \tag{2.14}$$

$$T = \frac{p^+_{Z_2}}{p^+_{Z_1}} \tag{2.15}$$

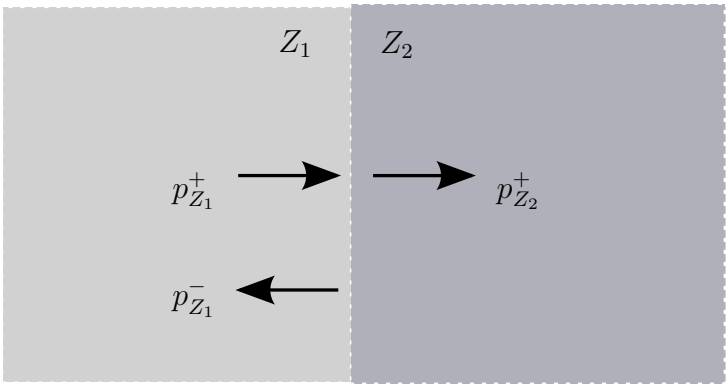


Figure 2.1: Sound Transmission and Reflection at an Acoustic Impedance Interface, after [5]

By dividing through by $p^+_{Z_1}$, (2.12) can be rewritten as (2.16).

$$1 + R = T \tag{2.16}$$

Considering the velocity-impedance relationship of (2.5), it is then possible to rewrite (2.13) as (2.17).

$$\frac{p^+_{Z_1}}{Z_1} - \frac{p^-_{Z_1}}{Z_1} = \frac{p^+_{Z_2}}{Z_2} \tag{2.17}$$

The rearrangement of (2.17) then results in (2.18).

$$\frac{p^+_{Z_1}}{p^+_{Z_2}} - \frac{p^-_{Z_1}}{p^+_{Z_2}} = \frac{Z_1}{Z_2} \tag{2.18}$$

Rearrangement and substitution of (2.14) and (2.15) into (2.18) results in (2.19).

$$1 - R = \frac{Z_1}{Z_2} T \tag{2.19}$$

Equation (2.19) can then be rearranged to provide an expression for the reflection or transmission coefficients, based on the values of Z_1 and Z_2. These are given in (2.20) and (2.21) respectively.

$$R = \frac{Z_2 - Z_1}{Z_2 + Z_1} \tag{2.20}$$

$$T = \frac{2Z_2}{Z_1 + Z_2} \tag{2.21}$$

These definitions of the reflection and transmission coefficients allow calculation of the fraction of an incident pressure wave that is reflected from, or transmitted across an interface in acoustic impedance (as per (2.22) and (2.23)). The previous example addressed the case of a boundary with infinite acoustic impedance, hence no sound power was transmitted. In a real world situation all boundary media are likely to feature a finite amount of transmission, dependent on the system architecture.

$$p_{Z_1}^{-} = Rp_{Z_1}^{+} \tag{2.22}$$

$$p_{Z_2}^{+} = Tp_{Z_1}^{+} \tag{2.23}$$

When $Z_2 < Z_1$ in Fig. 2.1, (2.20) determines the reflection coefficient will be negative while (2.21) suggests that the transmission coefficient will be positive. The reason for this is perhaps best investigated by considering the limiting case of an infinitely soft boundary, of zero frequency-independent acoustic impedance. This system behaves like an inverse of the infinitely hard boundary. At the interface, the pressure is forced to zero, as a force cannot result from a system with effectively zero area. A boundary displaying these characteristics is known as a pressure releasing boundary for this reason. Just as pressure doubles when velocity is forced to zero, in this case the velocity doubles, drawing particles towards the interface. Since the polarity of velocity is constant, the reflected component of pressure is inverted, explaining the negative reflective coefficient. In this limiting case the transmission coefficient (and hence transmitted pressure component) goes to zero, however in a real case (for non-zero Z_2) the coefficient will be finite and positive.

The values of R and T for an interface can go some way towards describing its acoustic behaviour. Fundamentally, a positive reflection coefficient describes wave incidence at an interface to a larger acoustic impedance. The reflected and transmitted pressure components will be a weighted form of

the incident wave. Conversely, a negative reflection coefficient describes incidence of a wave at an interface to a smaller acoustic impedance. In this case the reflected wave component will be an inverted, weighted form of the incident wave. The transmitted component meanwhile will not be inverted, rather a weighted form of the incident wave.

The assumption of normal wave incidence held for previous derivations is clearly a luxury that cannot often be fulfilled in real systems. Wave motion is very rarely planar, and can only be considered so under certain conditions relating wavelengths to containing geometries. Where specific acoustic impedance changes during sound transmission, there is the additional concern of refraction, a change of wave direction with speed. Under the assumption of incompressibility in acoustics, it is possible to assume homogeneity of the medium. Where characteristic acoustic impedance is used to describe discontinuities due to the geometry there are no implications for the wave speed, only the pressure to volume velocity relationship.

The implications of oblique incidence are particularly dependent on the assumptions made in boundary modelling. These are discussed in this context in section 3.6.

Diffraction

The nature of entropic relaxation in a gaseous medium causes localised changes in ambient pressure to propagate in a three-dimensional manner. Even in cases where the planar wavefront assumption is close to holding, a simple contribution from the surrounding geometry can provide a situation where the assumption no longer holds. The reason for this is rooted in Huygen's wavefront principle. This describes each point in a propagating wavefront as an individual, spherical point source. In an assumed planar wave, these spherical sources will overlap and add to produce what is apparent as a continual planar front contained between geometrical limits. When incident upon an aperture, the waveform will spread in a manner dependent upon the width of the opening and the wavelength. This spreading is termed diffraction, and is fundamental to acoustic radiation whereby a wave propa-

gates into a larger space in a spatial distribution dependent on the aperture geometry.

Diffraction is often characteristic of how sound waves will interact and interfere when exposed to changes in the containing geometry. Its mathematical derivation is non-trivial, demonstrations for various circumstances are given in [5, 14]. Both observe that within these diffraction patterns maxima and minima in pressure variation can be observed at angles dependent on the interference patterns generated by the aperture-wavelength relationship. The crucial factor for a single aperture is the path length difference from one side of the opening to the other, an element central to what is known as single-slit diffraction. It is observed that for different wavelengths, at different angles this path difference will vary. For some frequencies the path difference at a given angle might correspond to a phase shift of π, causing destructive interference, cancellation and hence a minima in the radiated wavefront. For other angles and/or frequencies the path difference might correspond to a 2π phase shift, constructive interference and a consequent maxima in the wavefront. For more complex geometries such as those featuring multiple apertures the calculation of maxima and minima positions becomes particularly complex.

Diffraction effects begin to become apparent where aperture widths are in the same order as incidental wavelengths, since this is when the path-differences and corresponding phase shifts begin to interact. When geometrical features are small compared to a given wavelength, their interaction is minimised. When geometrical features approach the same order as a given wavelength, the waves will begin to interact with the feature, generating a particular diffraction pattern.

Diffusion

In real-world acoustics surfaces are rarely completely flat. Any real surface will therefore present an effectively rough surface, even if this roughness occurs only on a very small scale. On a larger scale, it's much easier to see that wave components should be scattered at different angles. This scattering

is known as diffusion [17]. The very nature and shape of a reflecting surface, coupled with the incident wavelength determines the degree of diffusive behaviour.

2.4 The Acoustics of Cylinders

As shown in section 2.2, the characteristic acoustic impedance of a cylinder is given by (2.24), where S is its cross-sectional area.

$$Z_c = \frac{\rho c}{S} \tag{2.24}$$

This implies interfaces between successive cylindrical sections of changing cross-sectional area can be neatly characterised by a single macroscopic acoustic impedance, without having to consider the microscopic changes in specific acoustic impedance on a particle-by-particle basis.

Where the cylinder diameter d is acoustically small ($d << \lambda$) it is also fairly safe to make the assumption of planar wave propagation.

Where a propagating sound wave is incident at a discontinuity in characteristic acoustic impedance (whether a change in cylinder diameter or simply a closed/open end), corresponding wave reflection and transmission will occur. To augment the reflective/transmittive theory developed in section 2.3, consider the steady state acoustic response of a cylinder with one closed end and one open end.

Beginning with a periodic pressure signal of the form $pe^{j\omega t}$ describe the components of the scaled signal $e^{j\omega t}$ by Euler's relation, as in (2.25) and (2.26).

$$e^{j\omega t} = cos(\omega t) + jsin(\omega t) \tag{2.25}$$

$$e^{-j\omega t} = cos(\omega t) - jsin(\omega t) \tag{2.26}$$

$$cos(\omega t) = \frac{e^{j\omega t} + e^{-j\omega t}}{2} \tag{2.27}$$

$$sin(\omega t) = \frac{e^{j\omega t} - e^{-j\omega t}}{2j} \tag{2.28}$$

In this case (2.27) and (2.28) give the zero-phase and quadrature components respectively. In this case the zero-phase component is the propagating sound pressure and the quadrature component its velocity. By de-Moivre's theorem define the real and imaginary components of the pressure signal in (2.29) seperately, as (2.30) and (2.31). In this case a describes the real component and b the imaginary.

$$pe^{j\omega t} \tag{2.29}$$

$$a = pcos(\omega t) \tag{2.30}$$

$$b = psin(\omega t) \tag{2.31}$$

Consider now the reflection and transmission coefficients for this steady state system to be phasors, whereby each has a real and imaginary component and hence magnitude and phase responses. A typical reflection coefficient is given in polar form in (2.32) and as a typical transmission coefficient in (2.33). In this case R and T represent the magnitude of reflection/transmission and ϕ represents the phase response for both.

$$Re^{j\phi t} \tag{2.32}$$

$$Te^{j\phi t} \tag{2.33}$$

Incidence with the boundary will correspond to multiplication by the reflection / transmission phasor for the reflected and transmitted components accordingly. This will impart a magnitude multiplication and phase shift. The boundaries explored in section 2.3 are pure real. This means there will be no imaginary component and hence zero phase shift (or similarly a phase shift of 2π). While the pressure releasing boundary represents a pure real boundary (at least in the simplified limiting case) it can be argued that

this apparently negative pure-real reflection is caused by a positive reflection coefficient with a complete phase inversion (or phase shift of π). Phase shifting by π does not affect the zero-phase/quadrature relationship, hence no additional imaginary component is introduced to the phasor.

The acoustic power transfer of a cylinder, or a concatenated sequence of cylinders is governed by the phase relationship between sound pressure and velocity. This is largely established by the wave propagation behaviour with respect to reflective functions [5]. In a plane wave propagating in infinite space, the velocity will always remain in quadrature with the sound pressure, as previously determined.

To further investigate the nature of this power transfer consider the cylinder of Fig. 2.2 (terminated at one end with an infinite impedance and zero-impedance at the other) for which the assumption of planar wave propagation holds. In a one-dimensional steady-state treatment, visualise the two components of (2.34) as the left- and right-going components p^- and p^+ respectively, where x denotes axial position.

$$p(x,t) = p^+(x,t) + p^-(x,t) \tag{2.34}$$

For maximum power transfer at any termination, the reflected component should be in phase with the incident component to cause constructive interference.

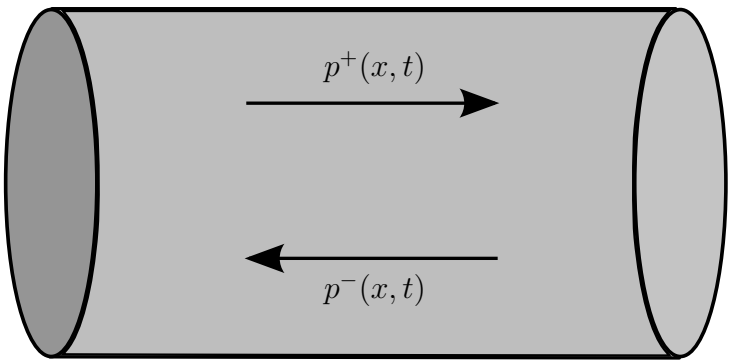


Figure 2.2: Uniform Cylinder Showing Left- and Right-Going Pressure Components

This requirement is expressed for the left-hand phase-preserving rigid termination in (2.35), (2.36) and (2.37).

$$e^{j(\omega+\phi)x} = e^{-j(\omega+\phi)x} \tag{2.35}$$

$$cos((\omega+\phi)x) + jsin((\omega+\phi)x) = cos((\omega+\phi)x) - jsin((\omega+\phi)x) \tag{2.36}$$

$$2jsin((\omega+\phi)x) = 0 \tag{2.37}$$

Note that the condition of (2.37) occurs at phase offsets described by (2.38).

$$\phi_L = n\pi : n \in \mathbb{N}^0 \tag{2.38}$$

For the right-hand (infinitely-soft) termination the same phase condition applies. The boundary itself is phase inverting, as demonstrated by the condition of (2.39).

$$e^{j(\omega+\phi)x} = -e^{-j(\omega+\phi)x} \tag{2.39}$$

$$cos((\omega+\phi)x) + jsin((\omega+\phi)x) = -cos((\omega+\phi)x) + jsin((\omega+\phi)x) \tag{2.40}$$

$$2cos((\omega+\phi)x) = 0 \tag{2.41}$$

Observe that the condition of (2.41) is met at phase offsets given in (2.42):

$$\phi_R = \left[\frac{2n+1}{2}\right]\pi : n \in \mathbb{N}^0 \tag{2.42}$$

The frequencies for which maximum power transfer occurs are those with a wavelength such as to meet both conditions (those established in (2.38) and (2.42)). This phase difference can be described by ϕ_R alone, since the left-hand face forms the reference point and through (2.38) will always be null for natural values of n.

For the example of Fig. 2.2 the phase difference is described as a fraction of a complete cycle by (2.43).

$$\frac{\frac{(2n+1)\pi}{2}}{2\pi} = \frac{2n+1}{4} : n \in \mathbb{N}^0 \tag{2.43}$$

To meet the conditions of maximum power transfer, the tube length L must then relate to the wavelength by this same fraction. Where $n = 0$, the relationship is $\frac{1}{4}$, meaning the tube length is a quarter of the first maximum transfer (or *resonant*) wavelength. For this reason such a cylindrical arrangement is commonly known as a quarter-wave resonator.

Based on the phase-difference method derived above, the frequencies of resonance for a quarter-wave resonator are given by (2.44).

$$f_R = \frac{(2m+1)c}{4L} : m \in \mathbb{N}^0 \tag{2.44}$$

In real systems, determining the resonant response is often complicated by the involvement of a steady-state (or otherwise) input. A typical example is that of a piston moving in a cylinder [14]. In this case the resonant response of the system is determined by simultaneous resolution of the mechanical input impedance and the load impedance presented by the cylinder. In this case, the source itself can be modelled as a phasor hence its acoustic impedance will have a resistive (real) and reactive (imaginary) component. To represent this system, a suitable transfer function will be developed and resonance will be seen to occur for frequencies at which the reactance vanishes.

While modelling a pressure releasing boundary with a negative, pure-real reflection coefficient appears an attractive approximation, it introduces a significant error. The region of air at such an aperture does not behave in the same manner as a solid boundary. Instead, the region slightly beyond the boundary is involved in the acoustic exchange [18]. Further, pressure-releasing boundaries can support a complicated three-dimensional behaviour whose compatibility with the planar wave assumption is limited. For a one-dimensional approximation of a pressure releasing boundary in a cylinder, this excitation of the region beyond the aperture implies the introduction of a reactive element to the characteristic acoustic impedance. This has the overall effect of increasing the apparent length of the cylinder, partially as a function of its cross-sectional area. It is hence widely known as end cor-

rection. The derivation of an appropriate characteristic acoustic impedance incorporating this effect is non-trivial, involving treatment of an aperture's radiation function. While complicated derivations are available for many aperture types [19, 20], a typical approach is to use functions which approximate the condition for common circumstances, such as an infinitely flanged, or unflanged pipe.

2.5 The Wave Equation

The wave equation is a fundamental statement in fluid dynamics, describing the propagation of travelling waves in multi-dimensional acoustic or mechanical systems by means of a second order partial differential equation in sound pressure and spatial distribution. Its derivation is included in Appendix A.

Different formulations of the wave equation are possible, dependent on the number of spatial dimensions over which the expression applies. The one-dimensional version is stated in (2.45), the two-dimensional wave equation in (2.46) and the three-dimensional equivalent in (2.47). In this case p represents sound pressure, t represents the temporal variable and x,y,z represent spatial coordinates for each axis.

$$\frac{\partial^2 p(x,t)}{\partial t^2} = c^2 \frac{\partial^2 p(x,t)}{\partial x^2} \tag{2.45}$$

$$\frac{\partial^2 p(x,y,t)}{\partial t^2} = c^2 \left(\frac{\partial^2 p(x,y,t)}{\partial x^2} + \frac{\partial^2 p(x,y,t)}{\partial y^2} \right) \tag{2.46}$$

$$\frac{\partial^2 p(x,y,z,t)}{\partial t^2} = c^2 \left(\frac{\partial^2 p(x,y,z,t)}{\partial x^2} + \frac{\partial^2 p(x,y,z,t)}{\partial y^2} + \frac{\partial^2 p(x,y,z,t)}{\partial z^2} \right) \tag{2.47}$$

The three can be neatly encapsulated by using the Laplacian operator ∇ to denote the coordinate system, as demonstrated in Appendix A and shown in (2.48).

$$\frac{\partial^2 p}{\partial t^2} = c^2 \nabla^2 p \tag{2.48}$$

The wave equation is derived from application of the principles of conservation of momentum and conservation of mass to wave motion (whether transverse or longitudinal) in a medium. It can for example be used to describe wave motion in strings, membranes or as is most applicable in this case, a gaseous medium. The wave equation hence forms the basis of many numerical physical models, although it is not without its shortcomings. Firstly, it governs only linear relationships in wave propagation phenomena. Situations often arise whereby system structures introduce non-linear behaviours. A typical example might be a struck cymbal (or plate) in which one mode of vibration serves to excite another. Such non-linear behaviours are also often met at the intersection of flow conditions and acoustics, as explored in section 2.6. Numerical modelling of such a system demands a more complete description of the system physics than that provided by the wave equation alone. This more basic representation of system physics is also encountered in the second limitation of the wave equation, in that real world losses are not incorporated. Transfers of energy to different forms during mechanical vibration will always lead to energy losses through damping. This describes why a given motion cannot be perpetual in real-world cases. The formulations of the wave equation here are lossless, suggesting that in a system dependent on them alone the total energy would be constant. This is clearly inconsistent with the real-world case, although (depending on the circumstances) the assumption of a lossless wave propagation need not constitute a significant source of error. In simple physical models, inserting appropriate damping behaviours at the boundaries of a system is often sufficient to ensure an adequate representation of overall behaviour.

2.6 Aero-Acoustics

While acoustics is typically considered separately to the study of aerodynamics, the two are particularly closely linked. Section 2.1 has defined air flow as the hidden DC condition of acoustics. While linear systems analysis suggests the two might remain decoupled, there are numerous circumstances under which interactions occur.

Flow

Flow represents a net movement in particles of a medium. Acoustics is generally distinct from flow in that it operates by the principle of entropic equalisation and can exist with, or without a flow condition. While flow and acoustics could be considered separate domains, flow can in fact strongly affect (and generate) acoustic waveforms. The interaction between flow and various geometrical features can give rise to rapid fluctuations in pressure, occurring at a frequency so as to enter the acoustic domain. Such flow-feature interactions are often harnessed to create acoustic sources through complex mechanisms such as the reed source of an oboe or the embouchure of a flute. The human voice is not dissimilar in that an airflow from the lungs is used to generate audible vibration at the vocal folds. In this case, the source functions through the Bernoulli principle. A flow between two surfaces induces an adductive force, drawing the two surfaces together until contact whence the adductive force is eliminated and the two surfaces separate due to their tendency to regain an equilibrium position. This process repeats at frequencies within the human hearing range, creating a rich, harmonic spectrum. In some cases the flow itself can serve as a source, the most common mechanism perhaps presented by turbulence.

Turbulence

Turbulence is an aerodynamic phenomenon. It describes a condition where flow changes from being laminar (streams follow straight, smooth and predictable bearings) to chaotic (streams breakdown to form unpredictable,shifting patterns). This change brings about a largely non-deterministic pattern of vortices, creating localised regions of angular momentum. Turbulence can occur under many conditions, typically when the flow breaks down due to changes in the aerodynamic conditions or geometrical incidence. An example occurs when a flow is forced through a small aperture. The flow rate is forced to increase to a point where upon leaving the aperture the flow degenerates, as the rapid change of speed cannot be supported in laminar flow. Vortices are shed to support the decrease in wave speed, dispensing energy and thus

generating a chaotic region which further impedes laminar flow. As a concept, turbulence is strongly relevant to acoustics due to the ease with which it can manifest itself as an acoustic source. From the acoustic domain it can be seen to generate wide-band noise. Extremely common in speech, it is the foundation of frication, where air flow from the lungs is forced through a deliberate constriction in the vocal apparatus. The best models of turbulence characterise its behaviour in the aerodynamic domain [21]. In the acoustic domain, our major concern is that the effect of turbulence is reproduced. This can be as simple as providing a (potentially shaped) noise source.

Summary

In this chapter the domain of acoustics and the behaviours that define it have been introduced. Consideration has been paid to the processes inherent to wave reflection, and how these behaviours might lead to resonance. The mathematical approximation of cylindrical resonators has been introduced, particularly with regard to the conditions at resonant mode frequencies, and complications in their numerical representation. Finally, the wave equation has been introduced and in Appendix A its derivation is provided.

Voice production depends on a combination of flexible acoustic filtering and aeroacoustic source mechanisms. Representation of such acoustic behaviours can be straightforward using the wave equation, while similar mathematical approximation of aerodynamic processes can be demanding. Since vocal fold vibration or turbulent frication, being the main sources of acoustic excitation, fall outside the scope of this study, effort will be placed solely on acoustic simulation of the vocal tract and comparable structures. This assumes that the aerodynamic mechanisms can be considered as linearly decoupled sources of acoustic excitation.

The dynamic nature of voice production demands a means of simulation capable of time-domain manipulation and operation. Having defined the domain of acoustics here, Chapter 3 continues to consider how it might be represented using a time-domain numerical model.

Chapter 3

Time Domain Acoustics Simulation

Introduction

In Chapter 2 the fundamentals of acoustics were introduced. This chapter considers how a numerical model of such acoustical behaviour might be developed. Sections 3.1, 3.2 and 3.3 demonstrate finite-difference and wave-digital approaches to the approximation of the wave equation. In section 3.4 the means for developing geometrically analogous models by multi-dimensional connectivity are explored. Interfacing of the two different approaches is then explored in section 3.5, allowing the benefits of each to be exploited. Section 3.6 provides formulations for numerical approximation to acoustic behaviour at a boundary, introducing inherent complications with respective treatments. The possibility of non-constant sampling grid density through domain-decomposition is introduced in section 3.7, followed by a consideration of dynamic numerical models in section 3.8. The need for proper treatment of the input and excitation mechanism itself is explained in 3.9. Finally, in sections 3.10 and 3.11 the computational costs and numerical error inherent to this simulation process are considered.

3.1 The Wave Equation

The wave equation (introduced in section 2.5) describes the propagation of sound pressure (or velocity) across variables of time (t) and space (∇). By directly solving the wave equation for a particular set of initial conditions, it is possible to state the sound pressure for any point in a free field at any instant.

For a digital implementation the wave equation must first be discretised, entailing decomposition of the continuous equation into an operation over *discrete* changes in space and time. There are two common approaches to this discretisation. The method chosen has a fundamental impact on the nature of the resulting numerical method [22]. Both schemes result in an algorithm resolved over a spatio-temporal sampling grid. This facilitates the generation of geometrically analogous multi-dimensional networks, or *meshes*, to simulate wave propagation in that medium.

In this discretised system, a relationship between the temporal and spatial sampling rates must be established. This defines the distance in Cartesian space represented by a single spatial step and by association the time interval represented by a single temporal step. In acoustics modelling, this relationship is often communicated by the temporal sampling rate, as this determines the sampling rate of the signal the numerical method will produce.

3.2 Finite-Difference Modelling

Finite-difference discretisation uses straightforward mathematical operations to directly discretise the wave equation. These are typically forward (3.1) and backward (3.2) difference operators applied in turn to approximate differential equations by trapezoidal integration [23].

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{3.1}$$

$$f'(x) = \lim_{h \to 0} \frac{f(x) - f(x-h)}{h} \tag{3.2}$$

Alternatively, a central difference operator can be twice applied to directly approximate the second derivative [23], as in (3.3).

$$f''(x) = \lim_{h \to 0} \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} \tag{3.3}$$

For discretisation of the partial differential expression of the wave equation, these difference operators can be stated in multi-variable forms, as demonstrated by the central difference operators of (3.4) for the 1D case, (3.5) for the 2D case and (3.6) for the 3D case.

$$f''_t(x,t) = \lim_{h \to 0} \frac{f(x,t+h) - 2f(x,t) + f(x,t-h)}{h^2} \tag{3.4}$$

$$f''_t(x,y,t) = \lim_{h \to 0} \frac{f(x,y,t+h) - 2f(x,y,t) + f(x,y,t-h)}{h^2} \tag{3.5}$$

$$f''_t(x,y,z,t) = \lim_{h \to 0} \frac{f(x,y,z,t+h) - 2f(x,y,z,t) + f(x,y,z,t-h)}{h^2} \tag{3.6}$$

Discretisation results in a discrete expression, resolved over a spatio-temporal grid of dimensionality matching the originating wave equation. Consider the one-dimensional lossless wave equation for sound pressure in Cartesian spatial variable x, given in (3.7).

$$\frac{\partial^2 p}{\partial t^2} = c^2 \frac{\partial^2 p}{\partial x^2} \tag{3.7}$$

The forward and backward difference operators expressed in (3.1) and (3.2) are defined for the limiting condition of a zero-width interval. When this interval is non-zero, the expressions instead represent an approximation to the derivative. For a practicable system real intervals are defined, Δt here for the case of a function in time and Δx for a function in Cartesian space. Approximation of the first derivative of the partial differential expression in (3.7) can hence be performed as in (3.8), where subscript $_x$ and superscript ^{t} denotes spatial and temporal indices respectively.

$$\frac{p_x^{t+1} - p_x^t}{\Delta t} = c^2 \frac{p_x^t - p_{x+1}^t}{\Delta x} \tag{3.8}$$

A successive approximation (3.2) can consequently be applied, providing an expression for the second derivative, as in (3.9).

$$\frac{(p_x^{t+1} - p_x^t) - (p_x^t - p_x^{t-1})}{\Delta t^2} = c^2 \frac{(p_{x+1}^t - p_x^t) - (p_x^t - p_{x-1}^t)}{\Delta x^2} \tag{3.9}$$

This expression can be rearranged, resulting in (3.10).

$$\frac{p_x^{t+1} - 2p_x^t + p_x^{t-1}}{\Delta t^2} = c^2 \frac{p_{x+1}^t - 2p_x^t + p_{x-1}^t}{\Delta x^2} \tag{3.10}$$

Consider the relationship between sampling intervals Δt and Δx. It is clear that over a single step in approximation, the distance covered by the wave (of speed c) in time Δt should not exceed the distance represented by a spatial sampling interval (Δx), as expressed in (3.11).

$$\frac{c\Delta t}{\Delta x} \leq 1 \tag{3.11}$$

This restraint is known as the Courant-Friedrichs-Lewy condition [24] and is described by a parameter known as the Courant number, denoted here by λ in (3.12).

$$\lambda = \frac{c\Delta t}{\Delta x} \tag{3.12}$$

The meaning of the Courant number is more obvious for multi-dimensional systems, as will shortly be seen. For now let $\lambda = 1$. This implies that the time taken for the wave to cross a spatial sampling interval is exactly that of the temporal sampling period. Applying (3.11), observe that the update equation of (3.10) can be significantly simplified.

$$p_x^{t+1} - 2p_x^t + p_x^{t-1} = p_{x+1}^t - 2p_x^t + p_{x-1}^t \tag{3.13}$$

$$p_x^{t+1} = p_{x+1}^t + p_{x-1}^t - p_x^{t-1} \tag{3.14}$$

A simple recursive statement is provided by (3.14), giving the next temporal sample as a function of its spatial neighbours and previous temporal value.

Observe that when $\lambda \neq 1$ the update equation becomes more complicated.

$$p_x^{t+1} = \lambda^2(p_{x+1}^t + p_{x-1}^t) + 2(1-\lambda^2)p_x^t) - p_x^{t-1} \tag{3.15}$$

Discrete approximations to the two- and three-dimensional lossless wave equations can be derived similarly. The continuous forms are provided in (3.16) and (3.17) for Cartesian spatial variables x,y,z. The respective approximations are then provided in (3.18) and (3.19).

$$\frac{\partial^2 p}{\partial t^2} = c^2\left(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2}\right) \tag{3.16}$$

$$\frac{\partial^2 p}{\partial t^2} = c^2\left(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} + \frac{\partial^2 p}{\partial z^2}\right) \tag{3.17}$$

$$p_{x,y}^{t+1} = \lambda^2(p_{x+1,y}^t + p_{x-1,y}^t + p_{x,y+1}^t + p_{x,y-1}^t) + 2(1-2\lambda^2)p_{x,y}^t - p_{x,y}^{t-1} \tag{3.18}$$

$$\begin{aligned} p_{x,y,z}^{t+1} =& \lambda^2(p_{x+1,y,z}^t + p_{x-1,y,z}^t + p_{x,y+1,z}^t + p_{x,y-1,z}^t + p_{x,y,z+1}^t + p_{x,y,z-1}^t) \\ &+ 2(1-3\lambda^2)p_{x,y,z}^t - p_{x,y,z}^{t-1} \end{aligned} \tag{3.19}$$

The first and perhaps most fundamental consideration in generating a network should be the sampling rate of the system. This determines the density of the mesh, by describing the geometrical distance described by each sampling step.

Recalling the definition of the Courant number as in (3.12), note that the wave propagation speed relative to the grid is different in one-, two- and three-dimensional systems. The Courant-Friedrichs-Lewy stability condition [24] provides upper limits for λ, to maintain system stability and physically conceivable wave speeds. These limits are:

- $\lambda \leq 1$ for one-dimensional systems
- $\lambda \leq \frac{1}{\sqrt{2}}$ for two-dimensional systems
- $\lambda \leq \frac{1}{\sqrt{3}}$ for three-dimensional systems

In one temporal time-step, these Courant numbers correspond to a point-to-point traversal of a 1D grid, a diagonal traversal of a 2D square grid and a corner-to-opposite-corner traversal of a 3D grid. These relationships are shown in Fig. 3.1.

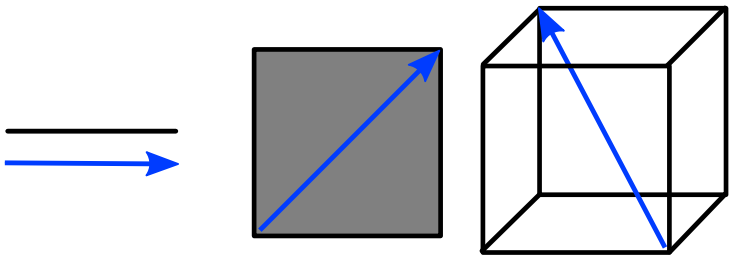


Figure 3.1: Ideal Wave Propagation Distances in a Single Time Step for Linear, Square and Rectilinear Grids

As with the one-dimensional case, choosing λ at the limit of the stability criteria causes the update equations of (3.18) and (3.19) to simplify.

By setting λ to the inverse root of the dimensionality of the system, substitute (3.20) into the expression relating temporal and spatial periods (3.11), to give equation (3.21), where N is the dimensionality of the system.

$$\lambda = \frac{1}{\sqrt{N}} \tag{3.20}$$

$$\frac{1}{\sqrt{N}} = \frac{c\Delta t}{\Delta x} \tag{3.21}$$

Through inversion of the temporal sampling interval Δt to give the system sample rate and rearranging, find:

$$f_s = \frac{c\sqrt{N}}{\Delta x} \tag{3.22}$$

where f_s is the system sampling rate and Δx gives the distance represented by each spatial sampling interval. In addition to simplifying the update equations, to minimise numerical dispersion (see section 3.10) the Courant number is often chosen at the limit itself [25].

In these finite-difference schemes, the values passed from unit to unit are measures of the system variables, in this case acoustic pressure or velocity. Since their behaviour is governed by Kirchhoff laws (as in circuit analysis) they are known as Kirchhoff variables, or K-variables [24, 26].

3.3 The Digital Waveguide Mesh

Section 2.4 introduced the concept of representing wave motion in a cylinder as a summation of two oppositely-travelling wave components. This idea is central to the d'Alembert solution of the wave equation. Mathematician Jean le Rond d'Alembert observed that any arbitrary twice-differentiable waveform, when simultaneously propagated in opposite directions (relative to the spatial variable) provides a satisfactory solution of the one-dimensional lossless wave equation [6]. The solution hence bears his name and the technique has become synonymous with wave digital representation of circuit elements. The d'Alembert solution can be elegantly demonstrated by factorisation of the wave-equation [27].

First, rearrange the one-dimensional wave equation of (3.7) to give (3.23).

$$\frac{\partial^2 p}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = 0 \tag{3.23}$$

While factorisation of such a partial differential expression is not strictly possible, separating the variables as in (3.24), where $p(x,t)$ is a pressure signal, helps to illuminate the underlying principles of the solution.

$$\left(\frac{\partial}{\partial x} + \frac{1}{c}\frac{\partial}{\partial t}\right)\left(\frac{\partial}{\partial x} - \frac{1}{c}\frac{\partial}{\partial t}\right) p(x,t) = 0 \tag{3.24}$$

Let this pressure signal $p(x,t)$ be composed of two arbitrary separately travelling signals, f_R and f_L according to (3.25).

$$p(x,t) = f_L(x+ct) + f_R(x-ct) \tag{3.25}$$

Substitution of this expression into (3.24) results in (3.26).

$$\left(\frac{\partial}{\partial x} + \frac{1}{c}\frac{\partial}{\partial t}\right)\left(\frac{\partial}{\partial x} - \frac{1}{c}\frac{\partial}{\partial t}\right)[f_L(x+ct) + f_R(x-ct)] = 0 \tag{3.26}$$

Multiplying out the elements of (3.26) consequently results in (3.27).

$$\frac{\partial^2 p}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = \frac{\partial^2 f_L(x+ct)}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 f_L(x+ct)}{\partial t^2} + \frac{\partial^2 f_R(x-ct)}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 f_R(x+ct)}{\partial t^2} \tag{3.27}$$

Observe now the effect of setting both f_R and f_L to the halved input signal defined in (3.28). Substituting this expression of the signals into (3.27) produces (3.29).

$$f_L = f_R = \frac{p(x,t)}{2} \tag{3.28}$$

$$\frac{1}{2}\frac{\partial^2 p}{\partial x^2} - \frac{1}{2c^2}\frac{\partial^2 p}{\partial t^2} + \frac{1}{2}\frac{\partial^2 p}{\partial x^2} - \frac{1}{2c^2}\frac{\partial^2 p}{\partial t^2} = \frac{\partial^2 p}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{3.29}$$

Equating (3.29) to the original 1D lossless wave equation demonstrates that $p(x,t)$ can be replaced by two components of half its value propagating in opposite directions. A more complete formulation of the d'Alembert solution is given in (3.30).

$$p(x,t) = \frac{1}{2}f(x-ct) + \frac{1}{2}f(x+ct) + \frac{1}{2c}\int_{x-ct}^{x+ct} g(y)dy \tag{3.30}$$

Note the additional term here, $g(y)$ describing the initial velocity of the system. In many acoustic models this quantity can be assumed negligible.

The d'Alembert discretisation leads to a particularly elegant implementation in the digital domain, establishing opposite travelling waves using parallel delay lines, as shown in Fig. 3.2. For the system to be realisable, all delay durations must be an integer multiple of a common unit delay [24]. This sug-

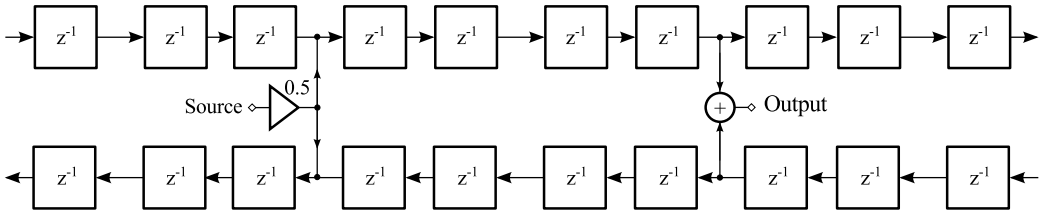


Figure 3.2: Example digital implementation of the d'Alembert travelling-wave solution to the lossless one-dimensional wave equation, demonstrating parallel delay line pair

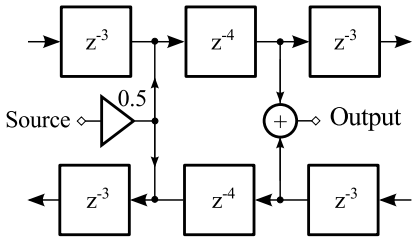


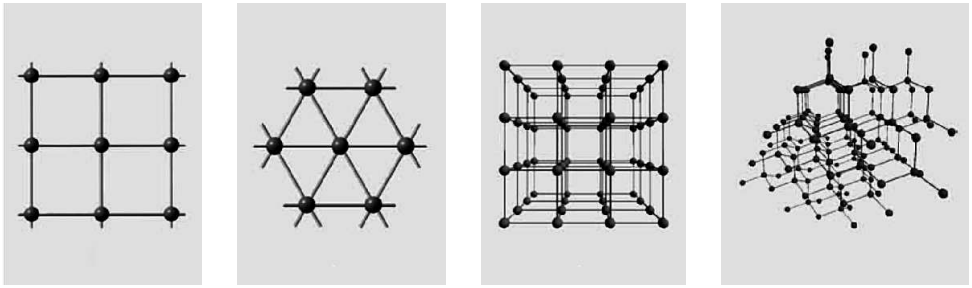
Figure 3.3: Reduced digital implementation of the d'Alembert travelling-wave solution to the one-dimensional wave equation by concatenation of series delay units

gests that the delay chain must be isotropic however consecutive units can be commuted, simplifying the hardware implementation as in Fig. 3.3 The stages in this one-dimensional network (of successive integer spatial variables) are known as digital waveguides [27].

Various units can be used to interface digital waveguides, to reproduce impedance-based scattering, reflection or for interface with different digital modelling units. These are known individually as Digital Waveguide Nodes [24] and are explored in section 3.4. The connection of DWNs in higher dimensionality arrangements provides an implicit solution to the appropriate higher dimensionality wave equation. This is in contrast to the Kirchhoff grid, for which a separate derivation must be made for each dimensionality. The multi-dimensional arrangement of DWNs is known as a Digital Waveguide Mesh (DWM). The DWM allows the connection of any number of waveguides to a node as long as the update equations are updated correspondingly and the topology is *regular*.

The topology of a mesh describes its spatial connectivity, as in the exam-

ples of Fig. 3.4.



(a) 4-Port 2D Rectilinear (b) 6-Port 2D Triangular (c) 6-Port 3D Rectilinear (d) 4-Port 3D Tetrahedral

Figure 3.4: Common multi-port mesh topologies from [6]

Square (2D) and rectilinear (3D) grids are widely used, each representing an implicit finite-difference solution and hence easy implementation. Triangular (2D) and tetrahedral (3D) topologies are popular where fitting a grid to irregular boundaries is a concern [24]. Different topologies of course also impart changing computational demand [28]. The sampling mesh developed is nominally uniform (although mesh sampling rates can be altered across domains and interfaced as explored in section 3.7. An exact statement of the limit in frequency domain validity is non-trivial, since it is related to factors such as geometrical fitting (in that the mesh may not allow an exact geometrical representation) and mesh topology. An additional limit is imposed on square and rectilinear grids. It can be observed that traversal between any two points on such a grid is made by either an odd or even summation of spatial intervals. This halves the effective resolution of the sampling grid, presenting two lower resolution interwoven networks and consequently halving the overall temporal Nyquist limit. Further, numerical accuracy is limited by dispersion error. This is the propensity of a sampling grid to present a frequency- and angle-dependent phase velocity to propagating components. Numerical dispersion error can be a significant form of error in physical models, and is hence considered in depth in section 3.10.

While Kirchhoff networks propagate the system variable itself, the variable that propagates in a DWM is a wave based representation of the be-

haviour of each individual waveguide. For this reason they are widely known as wave-variables or W-variables [26]. The crucial difference separating W- from K- variables is that wave variables do not directly describe the system behaviour. They are as Mullen describes, a hypothetical consideration to facilitate propagation [16]. To extract a meaningful value the contributions from the parallel delay lines must first be summed.

3.4 Scattering

Scattering entails the processes by which wave components are reflected and/or transmitted at the interface of individual propagating units. The Kirchhoff update equation encapsulates an explicit treatment of square/rectilinear scattering for the case of uniform acoustic impedance. The wave-variable case demands introduction of scattering units to allow the arrangement of digital waveguides in a regular connective topology (of any dimensionality) to produce a geometrical analogue. These networks can represent homogeneous or non-homogeneous media depending on whether acoustic impedance is explicitly included in the corresponding update equations. For both variable types a completely uniform acoustic impedance would correspond to a homogeneous medium. For non-constant acoustic impedance the nature of the grid formulation must change to represent this condition. The most intuitive example of intra-medium scattering is perhaps a one-dimensional chain representation of concatenated cylinders. The treatment of the acoustics of cylinders in section 2.4 introduced the nature of this problem, demonstrating the wave-variable case whereby opposite travelling components are summed. This showed a discontinuity in acoustic impedance to cause reflection and transmission of wave components from an interface, according to coefficients describing the nature of the change.

Implementation of scattering is quite different in Kirchhoff- and Wave-formulations, although the processes are fundamentally equivalent. Wave formulations use a measure of all incoming travelling-wave components and the respective acoustic impedances through which they arrive to calculate the appropriate node pressure and the corresponding outgoing component

for each connection. A Kirchhoff formulation can be derived from the wave case, arriving at a more direct treatment of neighbouring pressure values and corresponding acoustic impedances, defined over previous time steps.

Conditions of uniform acoustic impedance represent a simplified case of more general scattering formulations. For this reason scattering units for non-homogeneous media will first be considered, preceding a demonstration of the simplification afforded by impedance homogeneity to the Wave-variable case.

3.4.1 Non-Homogeneous Media

One-dimensional scattering in the wave domain is fundamental to the ability of digital waveguide chains to reproduce higher dimensionality behaviour.

The basic structure of the scattering port is illustrated in Fig. 3.5.

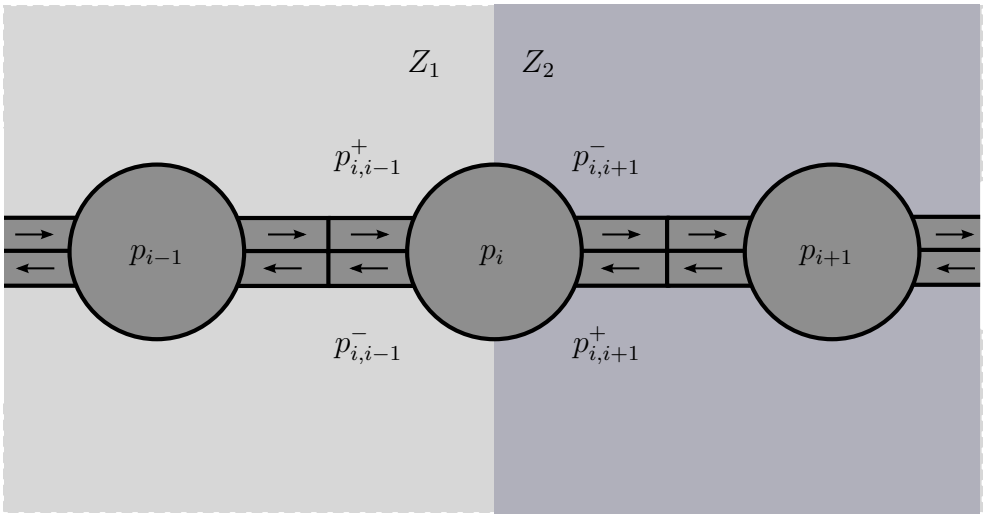


Figure 3.5: Chain of Digital Waveguide Nodes across a Specific Acoustic Impedance Discontinuity

Lossless scattering of wave components requires the direct application of continuity laws for the interface:

- The pressure on either side of the interface is equal to the sum of the incoming and outgoing pressure components on that side.
- The pressures against either side of the interface are equal.

- The total velocity at the interface is zero

Based on these continuity laws, there are two stages to scattering:

1. Calculate the pressure at the interface.
2. Calculate the outgoing variables as a function of the incoming components and the interface pressure.

To calculate the junction pressure begin with the third continuity law, that the total velocity at the interface is zero. To satisfy this condition the velocities either side must sum to zero as in (3.31) where u is the velocity component, superscripts $^+$ and $^-$ describe whether the component is incoming or outgoing and the subscripts describe the connection over which the component travels, with the closer node given first.

$$u^+_{i,i-1} + u^+_{i,i+1} = u^-_{i,i-1} + u^-_{i,i+1} \tag{3.31}$$

Rewriting and factorisation of (3.31) then results in (3.32) where Y_1 and Y_2 are the acoustic admittances either side of the interface.

$$Y_1(p^+_{i,i-1} - p^-_{i,i-1}) = -Y_2(p^+_{i,i+1} - p^-_{i,i+1}) \tag{3.32}$$

By combining the first and third continuity rules an expression for the junction pressure p_i can be made, as in (3.33).

$$p^+_{i,i-1} + p^-_{i,i-1} = p^+_{i,i+1} + p^-_{i,i+1} = p_i \tag{3.33}$$

This expression can then be used to eliminate outgoing components from (3.32), leaving (3.34).

$$-p_iY_1 + 2p^+_{i,i-1}Y_1 - p_iY_2 + 2p^+_{i,i+1}Y_2 = 0 \tag{3.34}$$

By gathering like terms, complete the expression for pressure at the interface, as in (3.35).

$$p_i = \frac{2(p^+_{i,i-1}Y_1 + p^+_{i,i+1}Y_2)}{Y_1 + Y_2} \tag{3.35}$$

Calculating the outgoing components is then trivial, by combination of (3.35) and (3.33) to give (3.37) and (3.38). These expressions define the outgoing components at the interface in terms of the incoming components and neighbouring acoustic impedances alone.

$$p^-_{i,i+1} = p_i - p^+_{i,i+1} \tag{3.36}$$

$$p^-_{i,i-1} = \frac{Y_1 - Y_2}{Y_1 + Y_2} p^+_{i,i-1} + \frac{2Y_2}{Y_1 + Y_2} p^+_{i,i+1} \tag{3.37}$$

$$p^-_{i,i+1} = \frac{2Y_1}{Y_1 + Y_2} p^+_{i,i-1} - \frac{Y_1 - Y_2}{Y_1 + Y_2} p^+_{i,i+1} \tag{3.38}$$

In section 2.3 the reflection coefficient was defined as a function of two acoustic impedances, as repeated in (3.39) and stated for acoustic admittance in (3.40).

$$r = \frac{Z_{i+1} - Z_i}{Z_i + Z_{i+1}} \tag{3.39}$$

$$= \frac{Y_i - Y_{i+1}}{Y_i + Y_{i+1}} \tag{3.40}$$

Substituting these definitions into (3.37) and (3.38) provides the effective scattering equations for the system in (3.41) and (3.42).

$$p^-_{i,i-1} = r p^+_{i,i-1} + (1 - r) p^+_{i,i+1} \tag{3.41}$$

$$p^-_{i,i+1} = (1 + r) p^+_{i,i+1} - r p^+_{i,i+1} \tag{3.42}$$

Fig. 3.6 shows the detailed structure of the digital implementation of these update equations. The role of the reflection coefficient is clear.

Smith demonstrates that the scattering unit can be further minimised as per Fig. 3.7 [27], in this case reducing the number of multiply operations required through substitution of (3.43).

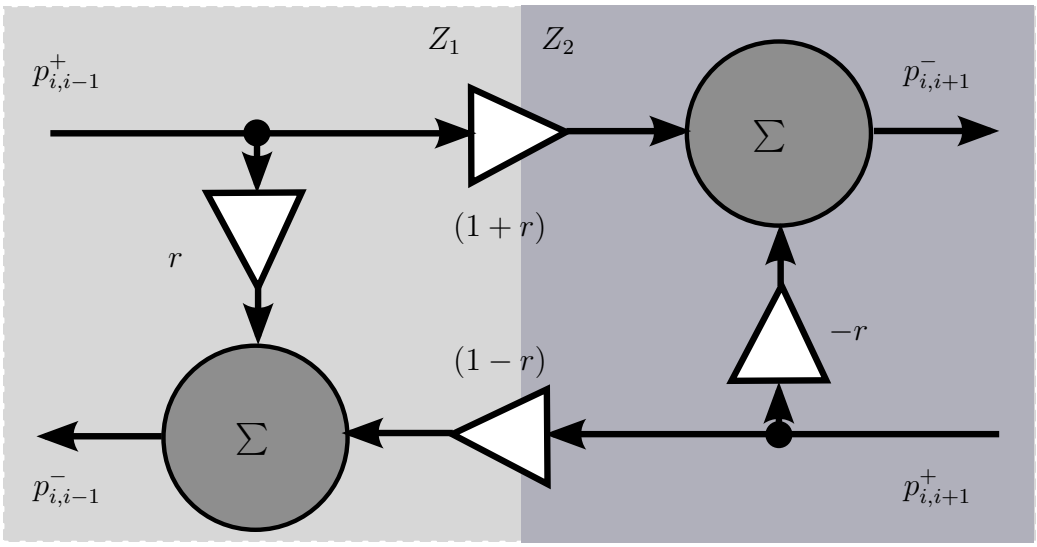


Figure 3.6: Digital waveguide scattering across an admittance discontinuity

$$
\begin{aligned}
w &= r(p^+_{i,i-1} - p^+_{i,i+1}) \\
p^-_{i,i-1} &= p^+_{i,i+1} + w \\
p^-_{i,i+1} &= p^+_{i,i-1} + w
\end{aligned}
\tag{3.43}
$$

One-dimensional wave-variable scattering is central to the operation of the Kelly-Lochbaum model, as explored in section 5.6.2. Higher dimensionality modelling using the digital waveguide mesh implies an intuitive extension of connectivity. Figs. 3.7 and 3.6 show the case for a one dimensional system, whereby wave variables from two surrounding nodes are taken into account and appropriately weighted by their respective acoustic admittances. Fig. 3.8 shows the case for scattering in a two-dimensional system, using a square mesh. Each node is indexed by its x and y coordinates, and the admittance of each of the n connections described by Y_n.

The resulting scattering equation is a version of (3.35) weighted for 4 connections, as in (3.44) where $p^-_{x+n,y+m}$ represents the outgoing travelling wave component of the node at index $x+n$,$y+m$, towards the node at x, y.

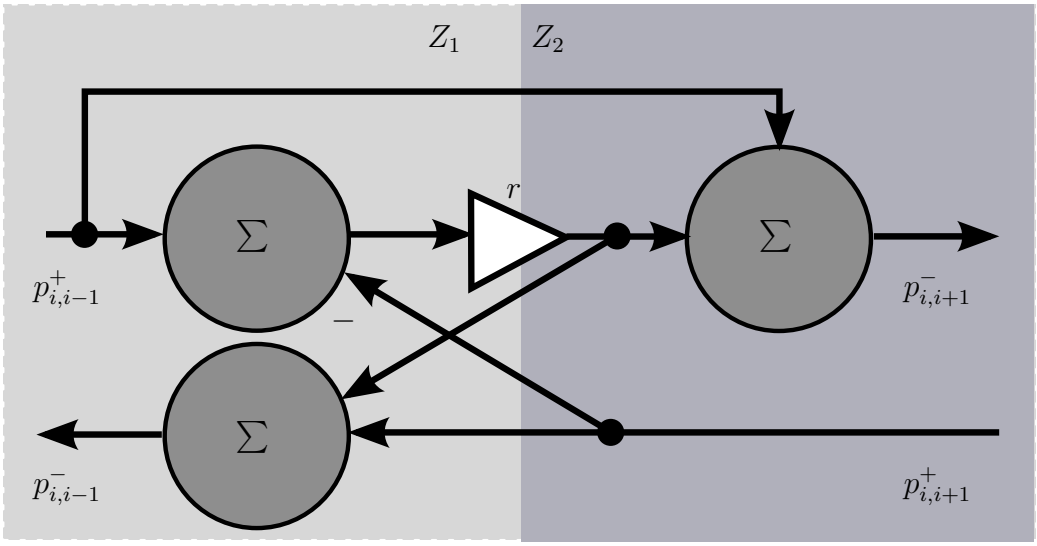


Figure 3.7: Minimised Digital Waveguide Scattering Across an Impedance Discontinuity

$$p_{x,y} = 2\frac{p^-_{x,y+1}Y_1 + p^-_{x+1,y}Y_2 + p^-_{x,y-1}Y_3 + p^-_{x-1,y}Y_4}{\sum_{j=1}^{4} Y_j} \tag{3.44}$$

This formulation can be standardised for any consequent connectivity, as in (3.45). Here p_i is the scattering node junction pressure, N is the number of connections, $p^+_{i,j}$ is the pressure component arriving at the junction i from the jth node and Y_j is the acoustic admittance of this connection.

$$p_i = 2\frac{\sum_{j=1}^{N} Y_j p^+_{i,j}}{\sum_{j=1}^{N} Y_j} \tag{3.45}$$

3.4.2 Homogeneous Media

For the case of homogeneous media, the uniform acoustic impedance generates greatly simplified update equations. Since $Y_1 = Y_2 = \ldots = Y_i$ where i is the connection index, (3.45) can be simplified to give (3.46).

$$p_i = \frac{2}{N}\sum_{j=1}^{N} p^+_{i,j} \tag{3.46}$$

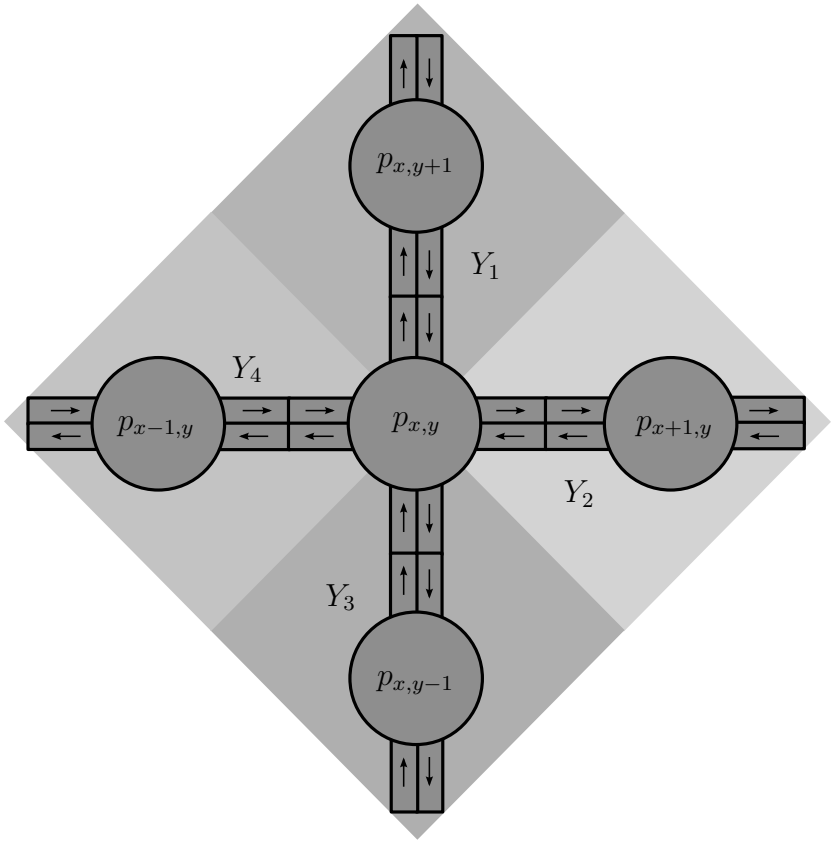


Figure 3.8: Digital waveguide scattering for a 2D rectilinear topology in a non-homogeneous medium

This can significantly reduce computation times for appropriate media.

3.4.3 Kirchhoff Equivalence

The Kirchhoff variable equivalents of these scattering formulations can be produced by substitution of the general wave update equation (3.47), describing the transfer of wave components from one unit to the next into the scattering equations.

$$p^{+}_{i,j} = z^{-1} p^{-}_{j,i} \tag{3.47}$$

Beginning with the non-homogeneous scattering equation, develop a series of equivalent expressions for node pressure update, as in (3.48), (3.49) and

(3.50).

$$p_i = 2\frac{\sum_{j=1}^{N} Y_j \left(z^{-1} p_{j,i}^{-}\right)}{\sum_{j=1}^{N} Y_j} \tag{3.48}$$

$$= 2\frac{\sum_{j=1}^{N} Y_j \left(z^{-1} \left(p_j - p_{j,i}^{+}\right)\right)}{\sum_{j=1}^{N} Y_j} \tag{3.49}$$

$$= 2\frac{\sum_{j=1}^{N} Y_j \left(z^{-1} \left(p_j - \left(z^{-1} p_{i,j}^{-}\right)\right)\right)}{\sum_{i=1}^{N} Y_j} \tag{3.50}$$

By splitting the summation of (3.50) to separate 1 and 2 sample time steps (3.51) can be defined.

$$p_i = 2\frac{\sum_{j=1}^{N} Y_j p_j z^{-1}}{\sum_{j=1}^{N} Y_j} - 2\frac{\sum_{j=1}^{N} Y_j p_{i,j}^{-} z^{-2}}{\sum_{j=1}^{N} Y_j} \tag{3.51}$$

Observing the relationship between (3.45) and the negative component of (3.51), the junction pressure of a non-homogeneous network can then be defined for the Kirchhoff variable alone, as in (3.52).

$$p_i = 2\left(\frac{\sum_{j=1}^{N} Y_j p_j z^{-1}}{\sum_{j=1}^{N} Y_j} - p_i z^{-2}\right) \tag{3.52}$$

A similar derivation can be performed for the homogeneous case, resulting in (3.53).

$$p_i = \frac{2}{N}\sum_{j=1}^{N} p_j z^{-1} - p_i z^{-2} \tag{3.53}$$

3.5 Mixed Models

Numerical modelling sits astride various fields of research outside acoustics. Both wave-digital and finite-difference approaches have seen extensive development and boast individual strengths and weaknesses. Developments in either methodology often complements the other. Wave-digital for example has seen the development of the wave-digital filter (WDF), describing frequency-sensitive RLC circuits as lumped elements [24] via the bilinear transform [27]. Finite-difference methods meanwhile have experienced a more explicitly mathematical development, resulting, for example in the frequency dependent locally reacting surface [25]. The treatment of numerical dispersion error in the finite-difference formulation is also well understood. While the variables in each approach strictly represent different things, they can be interfaced and used together in a hybrid numerical model. To combine K- and W- variables in such a system a translation must be performed from one format to the other. This is handled by a specialised scattering unit known as a KW-Connector or a KW-Pipe.

3.5.1 The KW Pipe

The fundamental difference between wave and Kirchhoff-based implementations is that the wave version has two parallel delay lines, supporting the two separately travelling wave components (as explored in section 3.3). The Kirchhoff equivalent instead has two delay units embedded in the update expression for each node (as indicated by the two delay components in (3.53). Effectively the memory in a K-system is in the node whereas in a W-system it is in the connections (waveguides).

To facilitate variable translation, Karjalainen/Erkut consider which components are described by transmission across homogeneous wave-to-wave and Kirchhoff-to-Kirchhoff connections respectively [22]. They consequently formulate a transfer matrix for the K-W coupling element, as (3.54).

$$\begin{bmatrix} p_2^+ \\ z^{-1}p_2 \end{bmatrix} = \begin{bmatrix} 1 & -z^{-2} \\ 1 & (1-z^2) \end{bmatrix} \begin{bmatrix} z^{-1}p_1 \\ p_2^- \end{bmatrix} \tag{3.54}$$

This transformation can be implemented as shown in Fig. 3.9, coupling a waveguide node with a Kirchhoff equivalent.

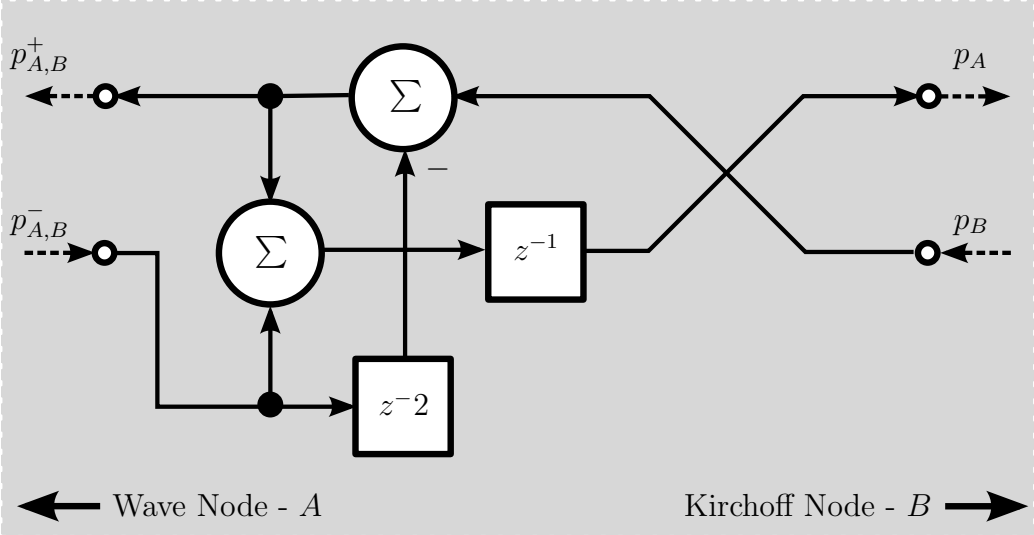


Figure 3.9: The KW Pipe - after [7]

The KW-Pipe provides a welcome opportunity to use either Kirchhoff or Wave variables in the conditions of their optimum application and interface the two where necessary. A single numerical model need not subscribe exclusively to a single domain of solution. For example, the dynamic digital waveguide mesh (see section 3.8) offers a highly stable means of changing the shape of a numerical model in real-time. Using KW-Pipes this system could then readily incorporate Kirchhoff-based boundary formulations.

3.6 Boundary Modelling

For numerical geometrical analogues, accurate and physically meaningful termination of scattering networks is fundamental to their performance. The characteristics of reflection (as per section 2.3) are not always straightforward and attention should be paid to the level of accuracy deemed necessary for satisfactory reproduction of reflective behaviours.

Fundamental digital waveguide units are able to approximate simple frequency-independent reflection at minimal computational cost. Frequency-dependent

reflection by contrast has traditionally presented a greater challenge. Wave digital filters have been used to provide an adequate representation of frequency-dependent reflection [27], with Kirchhoff-Wave hybrid techniques [7, 29, 24] developed to allow the inclusion of these wave-based units in Kirchhoff formulations.

Kowalczyk and van Walstijn have recently introduced a new Kirchhoff variable based frequency-dependent boundary formulation which has overcome an ambiguity in setting the Courant value in 1D connections to boundaries in multi-dimensional meshes [25, 30]. The new formulation offers computationally efficient frequency dependent and independent reflections and is explored further in section 3.6.2. Shelley has addressed diffusing reflections in the DWM [17], introducing a boundary formulation that approximates diffuse behaviour with reasonable effectiveness.

Most boundary formulations begin with a continuous equation for the boundary. This may take into account appropriate fundamental rules for the conservation of momentum and mass, and take properties such as diffusion and non-specular behaviour into account. Once discretised these equations can then be used as standard units within the scattering grid (using KW-couplers if appropriate).

In the case of the most simple DWN/DWM formulations, a simple boundary connection can be as simple as a standard scattering unit, either at the limit of transmission/reflection (transmission coefficients $\rightarrow 0$) or using the case where the transmitted component has an outward path but no return path (to be explored in section 3.6.1). This case is known as the standard one-connection boundary.

3.6.1 The One-Connection Boundary

Regardless of the dimensionality of the sampling grid in which it resides, the simple one-connection boundary is always a one-dimensional construct based on a standard variable scattering junction. While this represents a significant approximation it also leads to a particularly simple implementation. The first step is a discrete formulation of the boundary behaviour and the second

a substitution of this boundary behaviour into an appropriate scattering scheme.

Begin by considering the finite-difference derivation of a boundary junction. The discrete one-dimensional Kirchhoff scattering formulation is restated in (3.55).

$$\frac{p_x^{t+1} - 2p_x^t + p_x^{t-1}}{\Delta t^2} = c^2 \frac{p_{x+1}^t - 2p_x^t + p_{x-1}^t}{\Delta x^2} \tag{3.55}$$

By setting the Courant number to the one-dimensional limit ($\lambda = 1$) the reduced expression of (3.56) is obtained.

$$p_x^{t+1} = p_{x-1}^t + p_{x+1}^t - p_x^{t-1} \tag{3.56}$$

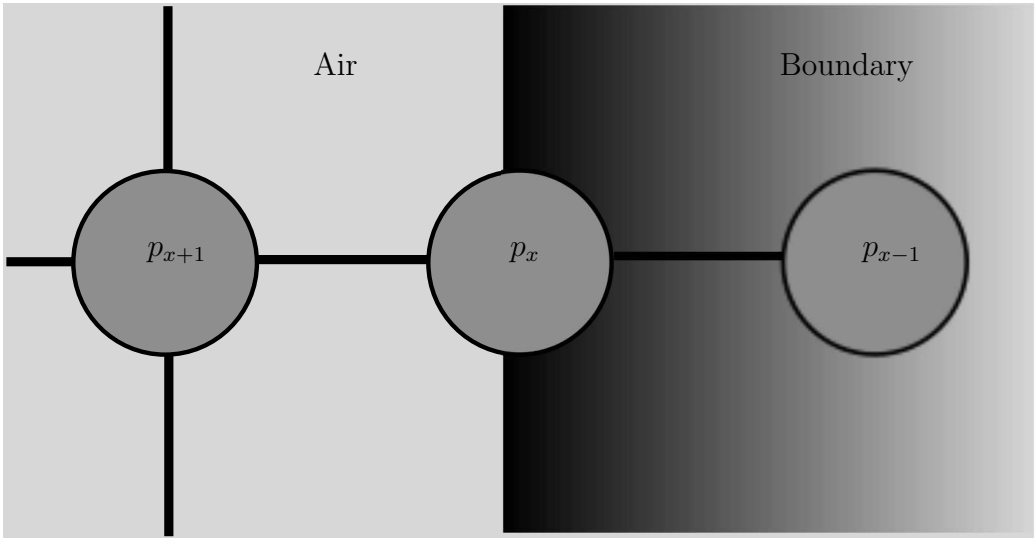


Figure 3.10: One-Connection Boundary Junction in Kirchoff Variables

Consider now Fig. 3.10. An understanding is required of the pressure at position p_{x-1}^t, within the boundary medium. Kowalczyk and van Walstijn call this 'dummy' node a ghost point [30].

As with the derivation of scattering, begin with a statement of the conservation of momentum, (3.57).

$$\frac{\partial p_x^t}{\partial x} = -\rho \frac{\partial u_x^t}{\partial t} \tag{3.57}$$

By translation through the acoustic impedance Z, rewrite (3.57) in terms of pressure alone, as in (3.58).

$$\frac{\partial p_x^t}{\partial t} = \frac{-Z}{\rho}\frac{\partial p_x^t}{\partial x} \tag{3.58}$$

Here define (3.59) to represent the normalised specific acoustic impedance of the boundary, G [31].

$$G = \frac{\rho c}{Z} \tag{3.59}$$

$$\frac{\partial p_x^t}{\partial t} = \frac{-c}{G}\frac{\partial p_x^t}{\partial x} \tag{3.60}$$

A generalised expression for the pressure at a boundary interface in terms of incident pressure and specific acoustic impedance is hence provided in (3.60). This can be easily discretised by applying the centred difference operator (3.3), resulting in (3.61).

$$\frac{p_x^{t+1} - p_x^{t-1}}{2T} = \frac{-c}{G}\left(\frac{p_{x+1}^t - p_{x-1}^t}{2X}\right) \tag{3.61}$$

Note that the ghost point here is $p(n-1, t)$, and hence (3.61) can be rewritten as (3.62).

$$p_{x-1}^t = p_{x+1}^t - \frac{XG}{cT}\left(p_x^{t+1} - p_x^{t-1}\right) \tag{3.62}$$

By reintroducing the Courant number, λ (3.12), find the expression in (3.63).

$$p_{x-1}^t = p_{x+1}^t - \frac{G}{\lambda}\left(p_x^{t+1} - p_x^{t-1}\right) \tag{3.63}$$

Substitution of this definition of the ghost point into the one-dimensional update equation (3.15) produces the expression for boundary pressure in (3.64), where the ghost point has been eliminated.

$$(1 - G\lambda)p_x^{t+1} = 2\lambda^2 p_{x+1}^t + (G\lambda - 1)p_x^{t-1} + 2(1 - \lambda^2)p_x^t \tag{3.64}$$

For normal incidence the boundary specific acoustic impedance G can be defined as (3.65) [30].

$$G = \frac{1-r}{1+r} \tag{3.65}$$

By setting the Courant number at the limit for a one-dimensional connection ($\lambda = 1$), substitute (3.65) into (3.64) and simplify the result to yield (3.66).

$$p_x^{t+1} = (1+r)p_{x+1}^t - rp_x^{t-1} \tag{3.66}$$

This is the Kirchhoff-variable update equation for any one-connection boundary, of reflective coefficient r.

Fig. 3.11 shows the equivalent wave-variable one-connection boundary. This is a special case of the wave scattering junction (see section 3.4) whereby the nominated boundary unit has an outgoing component toward a nominal ghost point in the boundary medium, but no corresponding incoming connection. Any wave component propagated into the boundary is therefore lost.

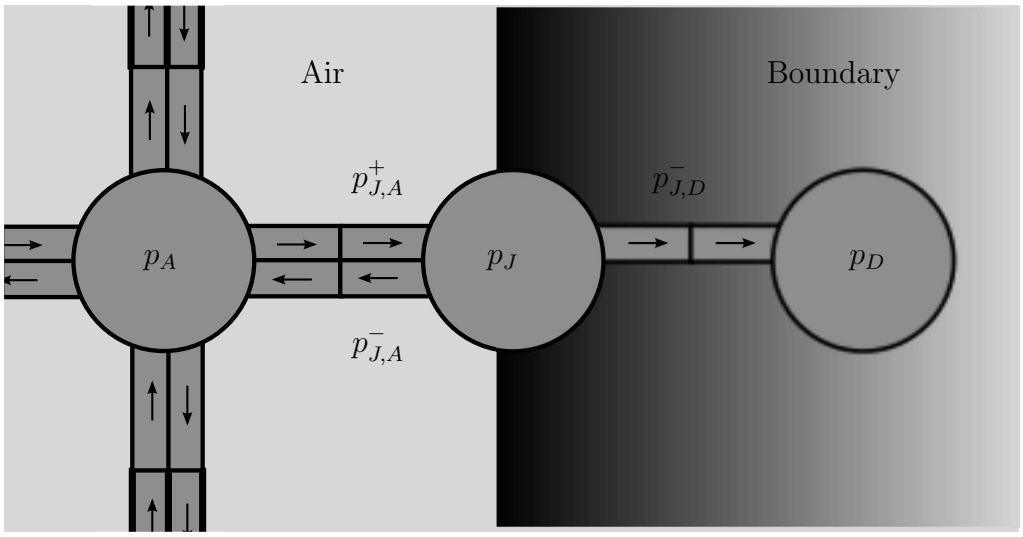


Figure 3.11: One-Connection Boundary Junction in Wave Variables

From section 2.3 it is clear to state that the outgoing wave component at

the junction can be determined by (3.67), where $p^-_{J,A}$ is the wave component returning from the junction and $p^+_{J,A}$ is the incident wave component. The reflection coefficient r can be positive or negative to represent a perfectly phase preserving or inverting reflection respectively.

$$p^-_{J,A} = r p^+_{J,A} \tag{3.67}$$

If necessary, the pressure at the interface itself (p_J) can be found by (3.35). Since the ghost point does not return an outgoing wave component this equation can be simplified as (3.68).

$$p_J = (1 + r) p^+_{J,A} \tag{3.68}$$

Similarly, the pressure radiated through the boundary medium can be found by (3.69).

$$p^-_{B,D} = (1 - r) p^+_{J,A} \tag{3.69}$$

The methodologies used in both boundary formulations are rooted in scattering techniques, hence the reflection coefficient remains a measure of the discontinuity in acoustic impedance, as per (2.20). A reflection coefficient of 1 therefore represents a frequency-independent, pure-real, constant and completely phase-preserving reflection. Likewise, a reflection coefficient of -1 represents a frequency independent pure-real and constant phase-inverting reflection. All values between will effect some degree of absorption, by transmitting a component which is not returned.

These boundary formulations are designed for conditions of normal incidence. This brings the accuracy of the boundary into question for non-specular reflection. The application of Huygen's wavefront principle is helpful here. Where a non-planar wave is incident on this manner of one-dimensional boundary a wavefront can be considered to be reconstructed in reflection by each individual boundary element. A more pressing concern is the choice of Courant number for the Kirchhoff formulation, as explored in section 3.6.2.

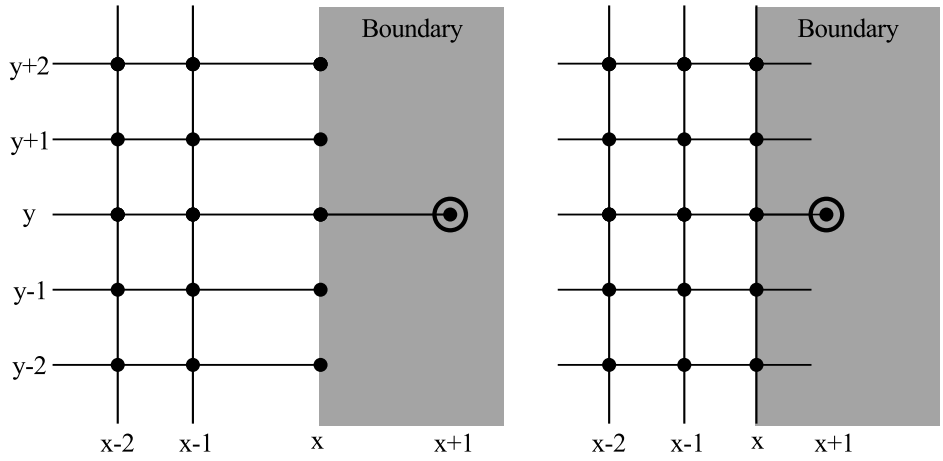
3.6.2 The Locally Reacting Wall

In section 3.2 the nature of the Courant number was discussed, particularly with regard to changing the dimensionality of the sampling grid. In the derivation of the Kirchhoff one-connection boundary the Courant number is set to $\lambda = 1$. This is perhaps understandable, since the nature of the connection is one-dimensional. One-connection boundaries are not however limited to one-dimensional grids. They can be (and are) widely used in two- and three-dimensional sampling grids. This introduces an ambiguity as to the spatial and temporal relationship represented by the boundary. One option is to set the Courant number so as to be appropriate for a one-dimensional connection, the other so as to remain consistent with the body of the mesh. Either approach will result in misrepresentation of the spatial/temporal relationship between the boundary and sampling grid as a whole. Kowalczyk and van Walstijn addressed this ambiguity with a Kirchhoff boundary termed the Locally Reacting Wall [25]. This formulation couples the ghost point to an update equation based on a derivation of the wave equation appropriate to the dimensionality of the body of the sampling network.

A two-dimensional mesh utilising a one-dimensional one-connection boundary is shown in Fig. 3.12a. Observe the disparity in spatial sampling interval between the boundary connections and the body of the mesh, introduced by the use of a one-dimensional Courant number for the boundary formulation and a two-dimensional Courant number for the body.

The locally reacting wall equivalent is shown in Fig. 3.12b. In this case the boundary formulation uses the same Courant number as the body of the mesh, leading to a consistent spatial sampling interval.

To implement the locally reacting wall boundary formulation the one-dimensional boundary update equation (3.63) is substituted for its equivalent in the discretised form of the two-dimensional wave equation (3.70), provided in (3.71). For the purposes of this demonstration consider index $(x + 1, y)$ to represent the ghost point, and G to represent the normal specific acoustic impedance of the boundary. The result of this substitution is given in (3.72).



(a) 2D Mesh - 1D Connection (b) 2D Mesh - LRW Connection

Figure 3.12: Comparison of one-dimensional and locally reacting wall boundary formulations in a two-dimensional rectilinear mesh

$$\frac{\partial^2 p}{\partial t^2} = c^2 \left(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} \right) \tag{3.70}$$

$$p_{x,y}^{t+1} = \lambda^2 (p_{x+1,y}^t + p_{x-1,y}^t + p_{x,y+1}^t + p_{x,y-1}^t) + 2(1 - 2\lambda^2) p_{x,y}^t - p_{x,y}^{t-1} \tag{3.71}$$

$$\begin{aligned} p_{x,y}^{t+1} = \lambda^2 (& \left[p_{x-1,y}^t - \frac{G}{\lambda} (p_{x,y}^{t+1} - p_{x,y}^{t-1}) \right] \\ & + p_{x-1,y}^t + p_{x,y+1}^t + p_{x,y-1}^t) + 2(1 - 2\lambda^2) p_{x,y}^t - p_{x,y}^{t-1} \end{aligned} \tag{3.72}$$

Rearranging (3.72) results in the final discrete update equation for the boundary point given in (3.73).

$$p_{x,y}^{t+1} = \frac{\lambda^2 (p_{x-1,y}^t + p_{x,y-1}^t + p_{x,y+1}^t) + 2(1 - 2\lambda^2) p_{x,y}^t + (\lambda G - 1) p_{x,y}^{t-1}}{1 + \lambda G} \tag{3.73}$$

In this case a Courant value appropriate to the two-dimensional mesh can be used for both the boundary connection and the body of the grid. This

removes the ambiguity introduced by the application of a one-dimensional boundary connection to a two-dimensional body.

In the case of a corner, there will be at least two ghost points. In these cases the boundary formulation (3.63) is substituted for each and every ghost point using different values of normal specific acoustic impedance if necessary.

3.7 Domain Decomposition

It was explained in section 3.2 that the spatial sampling interval represented by each unit in both wave and Kirchhoff modelling must be constant, leading to completely isotropic spatio-temporal sampling grids. For certain container geometries this is not an issue (for instance the square two-dimensional mesh provides an intuitive fit to a rectangular room). For other geometries it is possible that there will be both open and constricted regions. This makes the choice of system sampling rate complicated. A nominal number of units should be positioned across a geometry to reproduce resonant modes of that wavelength. It might appear sensible to adjust the system sampling rate so as to provide an adequate fit across the smallest dimension of interest, however this can often result in an unacceptably high number of units in the body of the geometry.

Domain decomposition describes a means of decomposing a sampling network into separate interfaced mesh regions with different sampling rates (and hence different spatial sampling intervals). This is not a new concept, mathematically rigourous descriptions for acoustics applications are provided by Bamberger, Glowinski and Tran [32] and Bilbao [24], but its implementation in the Digital Waveguide Mesh is only now beginning to be formalised, particularly in recent work by Kim and Scavone [8].

Bilbao approaches the problem by generating scattering units at the interface, resolving a rigourous finite-difference approach [24]. Kim and Scavone meanwhile pursue a technique intended for solving finite-difference problems in electro-magnetics, based on an overlap formulation [8]. This formulation operates by inserting a very narrow buffer region between regions (*subdomains*) of changing mesh density. Within this buffer region wave variables

are interpolated or averaged to translate them to the next, effectively independent mesh of higher or lower resolution respectively.

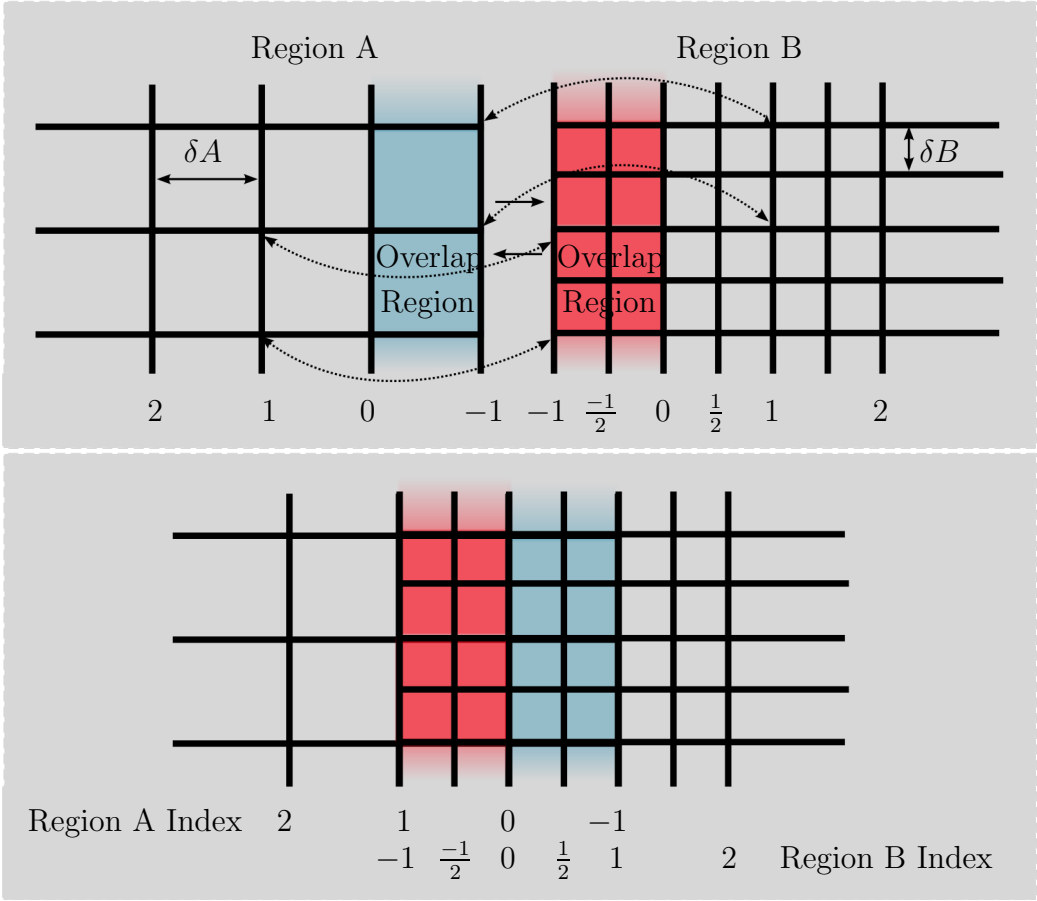


Figure 3.13: Interfacing Domains of Changing Waveguide Mesh Density by Overlap Method - after [8]

Fig. 3.13 demonstrates the nature of the overlap mechanism, with the leftmost grid (region A) of half the density of that on the right (region B). The equations to translate A to B are trivial, given by Kim/Scavone as (3.74) and (3.75) where superscript ^{A} denotes a node present in region A, superscript $^{A,+}$ describes the outgoing wave component of that node and subscripts $_{x,y}$ give the position index specific to each grid.

$$p^{A}_{-1,y}(t) = p^{B}_{1,y}(t) \tag{3.74}$$
$$p^{A,+}_{-1,y}(t) = p^{B,+}_{1,y}(t). \tag{3.75}$$

The node pressure in region A is described by (3.74), as a function of that in region B while (3.75) describes the outgoing wave component of that node after translation.

Transfer from region B to region A is more complicated since a linear interpolation is required to connect nodes with no obvious equivalent. For those with direct equivalents the update equations are given in (3.76) for node pressure and (3.77) for the outgoing wave component.

$$p^{B}_{-1,y/2}(t) = p^{A}_{1,y/2}(t) \tag{3.76}$$
$$p^{B,+}_{-1,y/2}(t) = p^{A,+}_{1,y/2}(t) \tag{3.77}$$

For the nodes without direct equivalents (those effectively interpolating the grid) the updates are given in (3.78) and (3.79) for the respective cases of node pressure and the outgoing wave component.

$$p^{B}_{-1,y/2}(t) = \frac{1}{2}\left(p^{A}_{1,(y-1)/2}(t) + p^{A}_{1,(y+1)/2}(t)\right) \tag{3.78}$$
$$p^{B,+}_{-1,y/2}(t) = \frac{1}{2}\left(p^{A,+}_{1,(y-1)/2}(t) + p^{A,+}_{1,(y+1)/2}(t)\right) \tag{3.79}$$

While domain decomposition presents an attractive proposition it can introduce problems of its own. The simple formulations presented by Kim and Scavone do not present a numerically transparent interface between the two regions. Effectively the two regions constitute differing characteristic impedances.

Fig. 3.14 shows an incremental simulation of two square grids of equal size, based on the 2D square digital waveguide mesh. The first (leftmost in visualisation) is 400×400 nodes and the second (rightmost in visualisation)

is 800×800 nodes. The domains are interfaced by the overlap method as per [8], utilising a 2-unit overlap. A dirac impulse is injected (see section 3.9) in the centre of the lower density mesh and the output from each point mapped onto a blue-black-red colormap for visualisation.

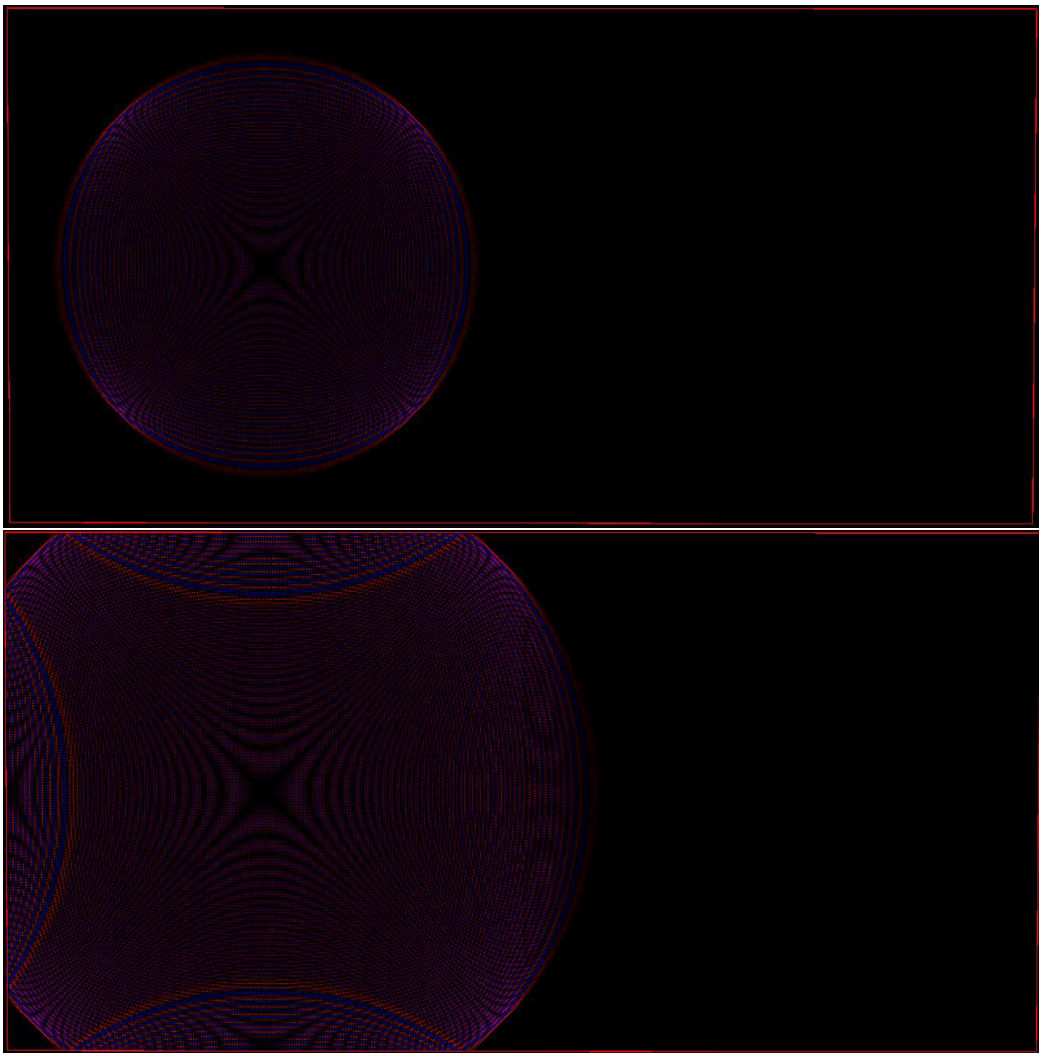


Figure 3.14: Simulation of two interfaced square grids of different mesh density using the overlap method. Note the wave component reflection exhibited at the boundary

As the wavefront passes through the overlap region a component reflection can be faintly observed. This apparent change in acoustic impedance damages the geometric analogue presented by the network and hence limits

its application. For now domain decomposition remains attractive, but its application would require careful consideration with regard to the artefacting it introduces.

3.8 The Dynamic Digital Waveguide Mesh

While a static physical model of a room is completely acceptable, there are many applications for which a dynamically changing acoustic model are desirable. These include models of dynamic processes such as voice production and physical modelling of certain musical instruments. In one-dimensional chains of DWNs, the characteristic acoustic impedance at each scattering junction can be changed without concern for numerical stability. This can be very effective and has been used to considerable effect in physical models of the voice, such as the Kelly-Lochbaum model (see section 5.6.2).

For a multi-dimensional Digital Waveguide mesh, the approach takes advantage of the role of acoustic admittance in the multi-port waveguide scattering formulation of (3.45) . All (simple) DWM models depend on this admittance-based scattering, whether this means an effectively infinite discontinuity at a solid boundary, or a more straightforward transmit/reflect combination due to a change in specific acoustic impedance. By manipulating Y_i in (3.45) it is possible to change the properties of the mesh from step to step with a reasonable degree of freedom. It is however important that due consideration is paid to whether the system remains physically meaningful. Moving an zero admittance to a new DWN to represent a contracting geometrical boundary shift can leave wave components effectively outside the boundary, yet still propagating either in free space or in a region of homogeneous acoustic impedance. Similarly, expanding the boundaries again can re-absorb these components. In these cases the accuracy of the physical model is called into question.

Significant developments towards dynamic modelling of the vocal tract were made by Mullen using a 2D DWM [16], overcoming these inherent issues by cosine-mapping discontinuities in acoustic impedance [33]. His work in this area is explored in more detail in section 5.7.

3.9 Excitation

Excitation is fundamental to the result of a simulation. Firstly, it cannot be overlooked that the resonant response of a system is largely dependent on source and receiver positioning. Resonance frequencies are largely determined by the combined response of the system and source, hence moving a source (or receiver) by a small distance can significantly affect the achieved response. Secondly, it should be considered how the simulated source is representative of a real source within that acoustic environment. The source radiation pattern and directivity particularly have the potential to impact the manner in which acoustic energy is delivered to the system and this should not be overlooked in simulation.

While the broad-spectrum of Dirac function excitation is appealing, the impact of aliasing must be carefully considered. In a system distributed both temporally and spatially, the maximum non-aliased frequency for injection is not clearly apparent. While the time domain is governed by the Shannon/Nyquist sampling theorem (and in practice to twice beneath this), the spatial distribution itself imparts a frequency-domain limit. As in the time-domain, a number of points in a spatial distribution are required for proper representation of a given wavelength. This will be a function of the spatial sampling interval. The matter is confused further by variant mesh topologies and non-constant phase velocity, as explored in section 3.10.

A practical solution is to restrict the excitation bandwidth to a fraction of the temporal sampling frequency. This can be as simple as injecting a low-pass filtered impulse - a windowed sinc function in this case, whose zero crossings correspond to the cut-off frequency of the band under simulation and with as wide a support as is possible within the scope of the simulation.

The manner in which the source is coupled to the grid also demands consideration. There are three common source mechanisms, the hard source, additive source and transparent source [34]. The application and correctness of each is largely dependent on the nature of the system under simulation, and the source behaviours the system aims to reproduce.

The hard source is the most straightforward of all the source mechanisms.

In this approach a particular input node is manipulated so as to be rigidly held at the values provided for input. Effectively the standard node update equation is overridden and the node pressure set to a desired value regardless of neighbouring values. This method is straightforward and computationally cheap, however it introduces significant changes to the nature of the sampling grid. The hard source itself becomes a part of the geometry under test rather than a separate source mechanism. While this might be appropriate for approximating source behaviours in certain physical models it is most often not for the excitation of models of room acoustics and in this case the vocal tract. In these cases our interest is in reproducing the acoustic behaviour of the geometry under simulation without explicit interference other than the source geometry. One option to reduce the impact of the hard source is to release the source mechanism after injection of a suitable (impulsive) waveform, although this restricts the length of the injected waveform which is undesirable in some cases [34].

The additive source is a simple development of the hard source whereby the input value supplements the existing variable field, rather than explicitly driving it. This yields a far less intrusive source mechanism, presenting a vanishing mechanical impedance and allowing source waveforms of arbitrary length.

The transparent source provides a means by which the very effect of the source and its interaction with the sampling grid is removed [34]. It is intuitive that any source mechanism will present its own impulse response, dependent on the mesh topology, sampling rate and formulation. This can be obtained by finding the response of an effective free-field to the intended input where the receiver is at the same position as the source. Since no propagated component should return to the source, the response obtained approximates that of the source mechanism itself and its coupling with the sampling grid. Once the response of a source-grid coupling to a given input is known, the transparent source can be implemented by subtracting the response from the input before injection. This ensures that any effects caused by source-grid interaction can be entirely removed from the simulation. The recorded response at any other position during simulation is then a measurement of the

response of the geometry under test with the impact of the source mechanism entirely negated.

Hard, additive and transparent source mechanisms can be implemented in either Kirchhoff or Wave formulations. In the wave formulation variable sources can be created by supplementing/replacing the incoming components at a scattering unit before the junction pressure and outgoing components are calculated. In the Kirchhoff formulation the junction pressures can be manipulated directly.

3.10 Numerical Dispersion Error

While numerical dispersion error certainly affects the absolute accuracy of the results of a simulation, a proper understanding of its causes and effects allows its influence to be acknowledged and accurate results (within implied criteria) to be achieved.

A medium is dispersive if it allows different frequencies to propagate at different speeds, implying a frequency dependent phase velocity. Air is such a medium, yet in the derivation of the lossless wave equation the assumption of frequency independence is made. This means the dispersive effect of air as a medium is not reproduced, hence the manifestation of frequency separation is not observed. This is indeed a form of dispersion error, but its effect can be reintroduced by adding passive lossy elements to scattering formulations [35]. This section considers instead a form of dispersion error introduced by the use of spatio-temporal sampling grids, specified by its particular definition as *numerical* dispersion error.

The sampling grids used in numerical simulation are a spatial sampling of a continuous acoustic field, in the same way a sampled signal is a discrete representation of its continuous counterpart. The ability of this grid to reproduce a field reliably is a function of the mesh topology, wave frequency and the shape of the wavefront. Other than through use of an infinitely dense grid, there is no way that any discrete sampling network can reproduce the behaviours of this wavefield entirely accurately. Simulation of any grid will hence introduce a frequency- and direction-dependent numerical dispersion

error. Different mesh topologies demonstrate different dispersive behaviours and in the rectilinear grid the Courant number chosen bears a significant influence.

The Courant number defines the relationship between the spatial and temporal steps used in the process of discretisation, as explored in the derivation of the Kirchhoff grid in section 3.2. Relating it directly to a particular grid of predefined density, it is possible to consider it to be the fraction of a complete grid interval (i.e. line, square, cube) that is traversed in one iteration. Where the Courant number is set at the limit as per (3.20), a wave component will travel the maximum distance within each node over each time step. This means that waves propagating in this exact path should experience zero dispersion error, regardless of frequency. Where the angle of wave propagation differs, a frequency-dependent dispersion will be introduced as the effective path length changes.

A broad description of the effect of numerical dispersion error would be that high frequencies will have a lower phase velocity than low frequencies [36]. The implications of this for numerical modelling are far reaching. Where a component's phase velocity differs from the expected value, frequency shifting of consequential resonant modes is likely to be observed. The shifting is frequency sensitive, lower frequencies tend to experience less error than higher frequencies. Indeed at higher frequencies and for oblique resonant behaviours, dispersion can affect the frequency response in a manner which is particularly difficult to delineate from the geometry itself.

Numerical dispersion error can be reduced to an arbitrary degree by increasing the density of the mesh. Whatever topology is used, an increase in density will result in a more accurate reproduction of the continuous acoustic field, and shifting of the frequency domain error curve away from the band of interest. A similar effect is achieved by the use of interpolated DWM schemes, whereby any number of nodes are introduced inside the existing grid and the original node values are calculated as a function of these new nodes [36, 37]. To treat predictable shifts in resonant modes at lower frequencies, efforts have been made towards correcting changes in phase velocity by frequency warping filters [37]. This process is intensive and hence does not

lend itself to realtime operation.

Different mesh topologies provide different means of discretising continuous fields and therefore produce different dispersive behaviours. Von Neumann analysis offers a mathematically rigourous means of analysing this behaviour [38], by which it is possible to express the relationship between the effective numerical wave speed and the ideal wave speed [36].

For the case of square, triangular and hexagonal two-dimensional networks Fontana and Rocchesso describe the ratio of the propagation speed to the ideal speed [35]. They achieve this by combining (3.80) and (3.81). The spatial phase shift for a single time sample is given in (3.80) as a function of spatial frequency and (3.81) describes the phase shift experienced over an arbitrary time interval where the Courant number is set at the limit.

$$\Delta\varphi_g(\xi_x, \xi_y) = -\frac{1}{\alpha_g} arctan \frac{\sqrt{4 - b_g^2}}{b_g} \tag{3.80}$$

$$\Delta\varphi = -2\pi D\xi \tag{3.81}$$

In both cases $\Delta\varphi$ describes the phase shift and b_g is a topology-specific function. In this two-dimensional case ξ_x and ξ_y describe the spatial frequencies, defining the frequency of a signal in terms of distance, instead of time. Given without subscript, ξ describes a spatial frequency vector as per $\xi = \sqrt{\xi_x^2 + \xi_y^2}$. A DC condition is represented by $\xi = 0$. Spatial frequency can be related to temporal frequency f in a square/rectilinear sampling network by $\xi = \frac{f}{f_s}$ where f_s is the temporal sampling rate of the system. Variable α describes the number of time steps between successive scattering nodes. It is equal to 1 for both square and triangular 2D topologies, but 2 for the hexagonal mesh [35]. Finally, D in (3.81) describes the optimum spatial sampling interval, which is the distance covered by a wave travelling at ideal speed in an idealised medium within a single time interval.

The spatial Nyquist limit is defined as:

$$(\xi_x, \xi_y) : (|\xi_x| < 0.5, |\xi_x| < 0.5) \tag{3.82}$$

$$(\xi_x, \xi_y) : \left(|\xi_x| < \frac{1}{2D}, |\xi_x| < \frac{1}{2D}\right) \tag{3.83}$$

The dispersion factor, defined as the ratio of the propagation speed in the mesh to that of an idealised medium is described by (3.84) as a function of spatial frequency.

$$k_g(\xi_x, \xi_y) = \frac{1}{2\pi\alpha_g D\xi} arctan \frac{\sqrt{4 - b_g^2}}{b_g} \tag{3.84}$$

The topology specific function is given for the square mesh in (3.85) and triangular mesh in (3.86). The function for the hexagonal 2D topology is available in [35].

$$b_{g,square,2D} = cos(2\pi D\xi_x) + cos(2\pi D\xi_y) \tag{3.85}$$

$$\begin{aligned} b_{g,triangular,2D} =& \frac{2}{3} cos(2\pi D\xi_x) + \frac{2}{3} cos\left(2\pi D\left[\frac{1}{2}\xi_x + \frac{\sqrt{3}}{2}\xi_y\right]\right) \\ &+ \frac{2}{3} cos\left(2\pi D\left[\frac{1}{2}\xi_x - \frac{\sqrt{3}}{2}\xi_y\right]\right) \end{aligned} \tag{3.86}$$

The dispersion factors for square and triangular two-dimensional grids are plotted against spatial frequency in figures 3.15 and 3.16 respectively. These are obtained by plotting (3.84) for topology functions (3.85) and (3.86), where $D = \alpha_g = 1$.

It is clear that for diagonal wave transmission in a square mesh dispersion is minimised, whereas off-axis behaviour is compromised by decreasing phase velocity at increasing frequencies. It is noted that the optimum wave propagation speed is $\frac{1}{\sqrt{2}}$, which is consistent with the Courant value chosen. The triangular mesh topology provides a more homogeneous dispersive behaviour in Fig. 3.16, but does not offer a single dispersion-free path as its square

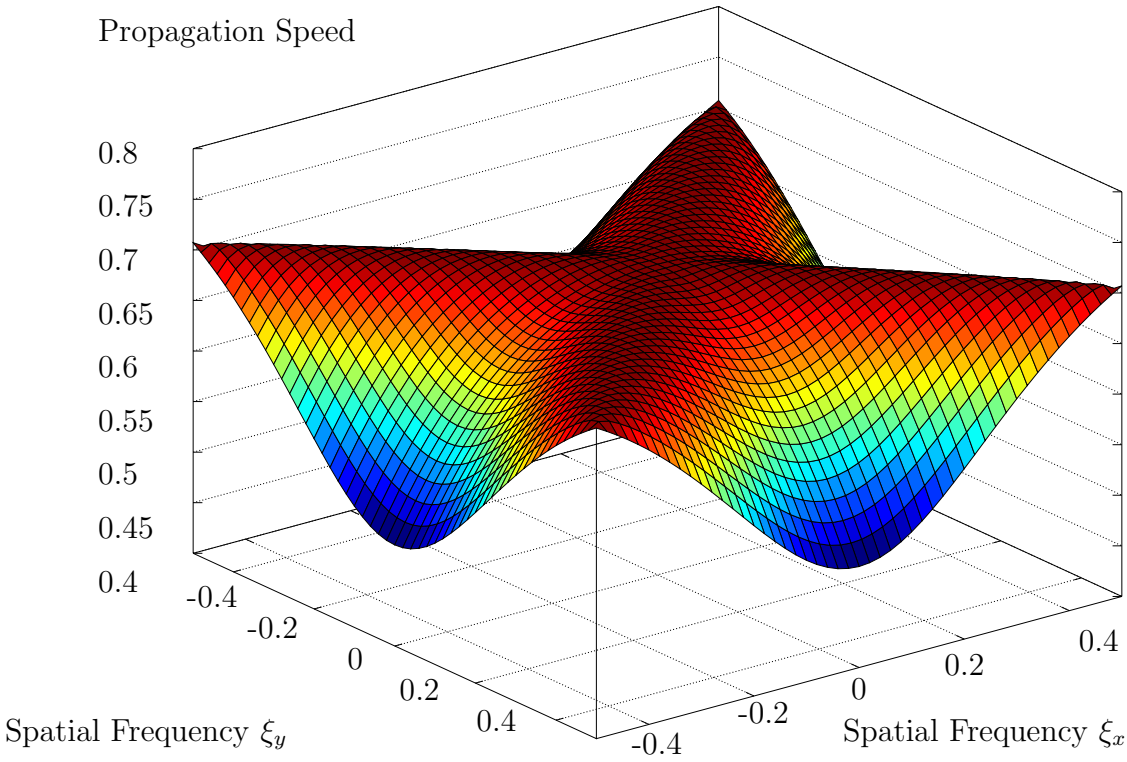


Figure 3.15: Dispersion Factor for a Square DWM as a Function of Two-Dimensional Spatial Frequencies

counterpart does.

Visualising the dispersion factor of a three-dimensional field poses a slightly different challenge. In [37] Savioja and Välimäki provide a function for the dispersion factor for spatial frequencies ξ_x, ξ_y and ξ_z, given here in (3.87).

$$k(\xi_x, \xi_y, \xi_z) = \frac{\sqrt{3}}{2\pi\xi} arctan \frac{\sqrt{4 - b_g^2}}{b_g} \tag{3.87}$$

The topology function b_g for a 3D rectilinear grid is given by $b_{g,rect,3D}$ as (3.88).

$$b_{g,rect,3D}(\xi_x, \xi_y, \xi_z) = \frac{2}{3} \left(cos(2\pi D\xi_x) + cos(2\pi D\xi_y) + cos(2\pi D\xi_z)\right) \tag{3.88}$$

This behaviour is plotted for intermittent slices of uniform ξ_z in Fig. 3.17.

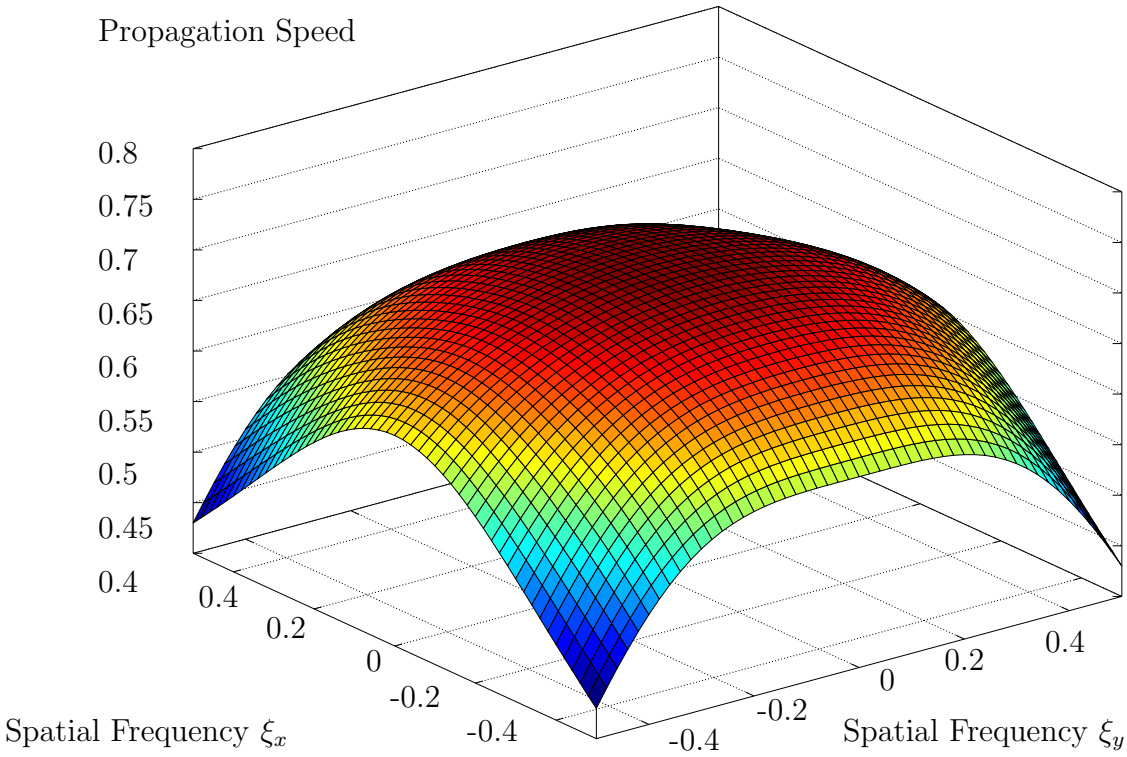


Figure 3.16: Dispersion Factor for a Triangular DWM as a Function of Two-Dimensional Spatial Frequencies

Subfigure 3.17a shows the condition for $\xi_z = -0.5$ (at the Nyquist limit). It is clear that the optimum wave speed occurs when all spatial frequencies are at their limits (i.e. $|\xi_x| = |\xi_y| = |\xi_z| = 0.5$). This corresponds to the condition of diagonal wave travel relative to the grid. This condition is further displayed in Figs. 3.17b and 3.17c. In Fig. 3.17d is effectively at the vertical centre of the established three-dimensional axes of spatial frequency. Note that the dispersion factor for $\xi_z = 0$ is not equivalent to that of a 2D square grid (Fig. 3.15. This should be unsurprising since in a 3D grid the optimum wave speed occurs diagonal to the cube formed by the grid, rather than diagonal to the square formed by a 2D grid.

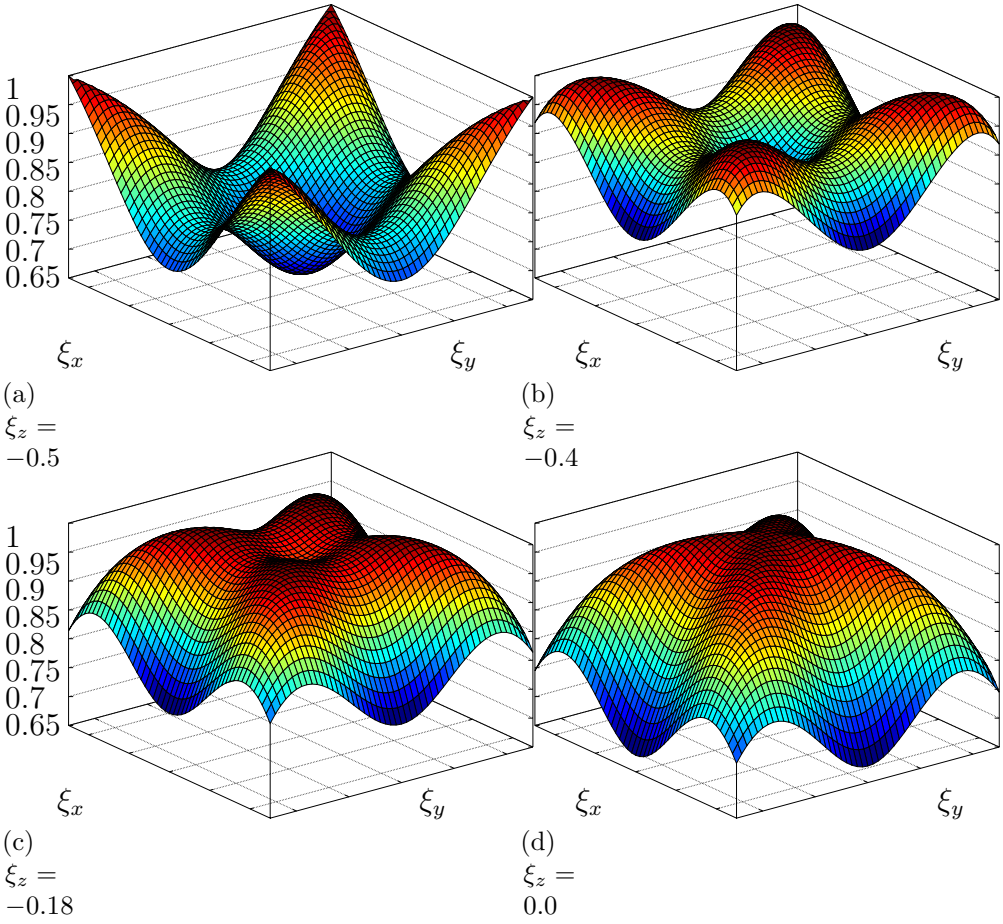


Figure 3.17: Dispersion Factor for a 3D Rectilinear DWM as a Function of Three-Dimensional Spatial Frequencies

3.11 Comparative Computational Cost

The smaller the geometry modelled, the smaller the spatial sampling interval required to adequately represent the space. This small spatial interval demands a similarly small temporal interval and therefore a high system sampling rate by (3.12). In the case of the vocal tract this is especially the case, with the combination of geometric constrictions and larger resonant spaces combining to demand a mesh that is both small and dense. It is important that despite these demands, the simulation is computationally *feasible*.

Feasibility in non-realtime numerical modelling comprises the ability to produce an impulse response in a sensible amount of time. Within the scope

of this project it also constrains computation to a standard desktop computer. These constraints can be challenging in two ways. Firstly, any computing system will impose a physical limit on the number of addresses that can be assigned in memory. Since in a typical implementation a single node will be given an individual data structure, this will restrict the number of nodes that can be generated in a simulation. A system running on a 32-bit operating system allows a theoretical limit of approximately 4.3×10^9 individual addresses, but even within the most optimised system these will not all be accessible by an individual process. To give this upper limit some context, it would represent a three-dimensional grid of approximately $1626 \times 1626 \times 1626$ nodes. At $192kHz$ this corresponds to a spatial sampling interval of approximately $3mm$, and the entire grid would span roughly $5m^3$. Such a high number of nodes will also generate a significant computational demand in terms of clock cycles. Under the conservative estimate of 6 double precision floating-point operations per node (6 additions, 1 multiplication), computing the equivalent of one second's worth of output for the above system at $192kHz$ would entail $6 \times (4.3 \times 10^9) \times (192 \times 10^3) = 4.9 \times 10^{15}$ individual computations, or approximate 5 petaflops. Assuming a CPU running at $3GHz$ can complete each operation in two cycles, one simulated second would require $\sim 9.9 \times 10^{15}$ cycles, completing in approximately 920 hours, or ∼38 days. This simulation is hence not computationally feasible. It is important to note that the values used in this estimation are a very rough approximation and can be considered particularly conservative. Simulation can very quickly become extremely demanding, hence it is crucial that due care is paid to minimising the data structures and iterated processes used.

Implementing a sampling grid using either variable type introduces two fundamental challenges: memory management and variable processing.

The challenge in memory management (beyond fitting the model into memory) is in finding the best solution for making each node aware of its neighbours. The optimum solution would be to have a large homogeneous array however outside perfectly rectilinear geometries this is not possible. There are two common practical approaches to this challenge. One is to maintain pointers inside the nodes with the addresses of each neighbour,

requiring $2N$ pointers per node and $2N \times 4$ bytes on a 32-bit machine. This approach is attractive as it allows each unit (and indeed the entire structure) to be completely self-contained. Calling a function in each node can allow the mesh to update itself entirely autonomously. Its major disadvantage is that connections must be duplicated, for example Node A has a pointer to Node B and Node B has a reciprocal pointer to Node A. The second approach is to maintain a connectivity graph away from the node data structures themselves. In this implementation each node contains only the memory required for data storage and variable update is handled by a central servicing process. The advantage of this approach is that each node is very small in terms of memory consumption. Its disadvantage is in the number of backward-and-forward calls and lookups that is required by the central servicing process to resolve the connectivity.

The second major challenge is that of providing an optimised processing stage. In the wave-variable case the node should maintain incoming and outgoing components for each connection. Assuming the use of double precision floating point on a 32-bit machine this will correspond to $2 \times 2N \times 8$ bytes, with an additional double(+8 bytes) to hold the current node pressure. In the Kirchhoff case all that is required is two doubles (16 bytes per node under our current assumptions), for the current and $[t-2]$ time-steps. Each variable type requires two stages of processing, an 'update' stage and a 'shuffle' stage. During the update stage the node pressure is calculated in both variables, according to the discrete expression (Kirchhoff) or component summation (wave). During the shuffle stage the outgoing wave variables are swapped to their neighbours via pointers (corresponding to a unit delay) and the Kirchhoff variables are stepped back in time.

The relationship between the system sampling rate and the number of nodes rises as a very approximate order of the dimensionality, as per (3.89) where n_{nodes} represents the number of nodes in the system and N is the system dimensionality.

$$n_{nodes}^{new} \approx \left(\frac{f_s^{new}}{f_s^{old}} \right)^N n_{nodes}^{old} \tag{3.89}$$

The developing trend in microprocessor development is no longer towards increasing clock speeds, but increasing parallelisation. The application of GPGPUs (General Purpose computing on Graphics Processing Units) has risen in popularity and developments have been made towards their application in acoustics modelling. While these approaches see significant acceleration in simulations they are also exposed to hardware-specific memory limitations [39].

Summary

In this section the fundamentals of multi-dimensional time-domain numerical modelling of acoustics have been introduced, along with derivations, and descriptions of particular formulations for approximating boundary behaviours. Techniques demonstrating particular potential for enhancement of the model are considered, including the dynamic and domain-decomposed digital waveguide mesh. Considerable attention has been paid to demonstrating the nature of numerical dispersion within the numerical method and factors affecting the feasibility of implementing the system.

It has been demonstrated that the digital waveguide mesh has the proven capacity for acoustic simulation of non-static structures, through effective shape changes implemented using impedance mapping. For voice simulation, such a capability is fundamental to the dynamic manipulation of vocal tract configurations through adjustment of the numerical model. The digital waveguide mesh will hence be pursued as the means of acoustic simulation, although impedance-mapping in the three-dimensional mesh falls outside the scope of this thesis.

Since this study largely constitutes an assessment of the suitability of this methodology to reproduction of voice, formulations will be kept as straightforward as possible to allow development of an appropriate benchmark simulation platform. To this end, wave-variable formulations alone will be implemented, omitting the potential computational savings afforded by incorporating Kirchoff formulations via the KW pipe. It is considered that the computational load of such a system in either formulation is initially too

large for real-time operation, hence a single variable formulation is preferred. The locally-reacting wall is omitted in favour of the simple one-connection boundary, which will introduce errors due to the effective change in dimensionality of the simulation, but is consistent with results produced in prior work. Domain decomposition is not incorporated at this stage, but provides an exciting, and perhaps necessary, avenue for future development.

It is these systems which will be used to construct a time-domain model. Having considered how they can be implemented, the next step is to gather an understanding of the nature of the problem presented by accurate simulation of the acoustics of the human voice.

Chapter 4

The Human Voice

Introduction

In Chapter 2 the mathematics of acoustics was introduced, particularly with regard to simple cylindrical systems. Following this, Chapter 3 explored methods for time-domain numerical acoustics simulation. In this chapter the anatomy and function of the voice is introduced, with consideration paid to the acoustic processes it encompasses. In terms of simulation of the voice, this represents our problem domain.

The human voice production system provides our most fundamental means of communication, functioning as a highly flexible acoustic instrument. A resulting speech waveform is an amalgam of information, which can be interpreted across various layers of abstraction. Ternström identifies this as reminiscent of a communications protocol stack, as per Table 4.1.

At each level of the waveform a different stream of information can be extracted based on modulation of some baseline quantity, which itself divulges information about the speaker [40]. While the vocal anatomy takes the role of the transducer, it responds and reacts to inputs from all layers. The focus of this Chapter is in defining the voice as an acoustic transducer.

Layer	Transmission	Content
Context	Society - Language - Genre	
Sender	Body - Mood - Personality - Situation	
Message	Content	*Thought*
Script	Verbal - Prosodic - Expressive - Extralinguistic	*Language*
Symbols	Phoneme - Timing - Pitch - Loudness - Timbre	*Speech*
Transducer	Articulation - Phonation - Respiration	*Voice*
Physical	Acoustic Waveform	*Sound*

Table 4.1: Effective Protocol Stack for Voice Communication after [1]

4.1 The Vocal Anatomy

The vocal anatomy can be loosely separated into the vocal tract and nasal tract. The vocal tract here entails the acoustic pathway from glottis to lips, incorporating the oral cavity, pharynx and supra-glottal larynx. The oropharynx is occasionally mentioned, referring to the region coupling the oral cavity to the pharynx. The nasal tract meanwhile describes the acoustic pathway from the nares (nostrils) to the velar coupling with the vocal tract, via the nasal cavity and coupled sinuses. The sub-glottal structure is not explicitly considered part of the vocal tract, with its coupling dependent on vocal fold closure [41, 42].

In this section orientation is described using appropriate anatomical terms relative to the subject, as per Fig. 4.1.

The vocal tract begins at the lips, providing a flexible aperture to the oral cavity, the base of which is provided by the mandible (jaw). The jaw is able to move in a superior/inferior direction at an angle subtended to the base of the skull. It is also able to move in the anterior/posterior direction by a much smaller degree.

The tongue is an isovolumetric muscle, rooted in the anterior pharynx and extending into the oral cavity [43]. It can be considered to have four parts, the tip, blade, front and back [4], proceeding in a posterior direction as per Fig. 4.2. Primarily used to manipulate food, the tongue serves an important secondary purpose in manipulating the shape of the oral cavity during phonation. The hard palate is the ceiling of the oral cavity and also the

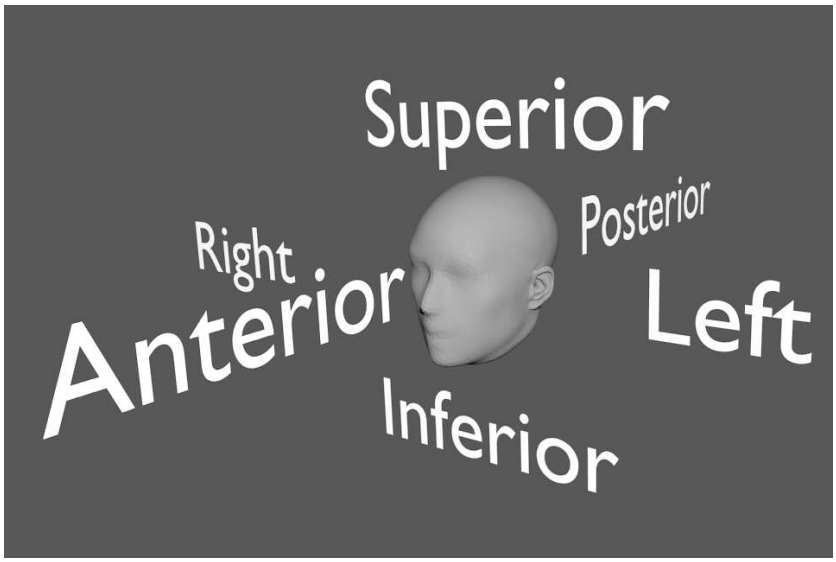


Figure 4.1: Terms for Orientation

floor of the nasal cavity. The tongue can form contact with the hard palate at all positions. At the rear of the hard palate is the flexible soft palate (velum) which controls coupling of the oropharynx with the nasal cavity. The port is typically sealed against the tongue back during mastication to allow simultaneous breathing through the nasal passage [43]. It can also manipulate the degree of coupling to the nasal cavity during phonation.

Above the velum is the nasal cavity, providing the primary path for inhalation/exhalation. The path begins at the nares (nostrils) of the nose, following a protracted path through the nasal conchae (which clean and moisten the incoming airstream) before arriving at the velum. Surrounding the nasal cavity are various other cavities, collectively known as the paranasal sinuses. These are coupled by very narrow apertures (ostia) to the nasal cavity [43]. The bilateral maxillary sinuses and ethmoidal sinuses lie either side of the inferior and superior concha respectively, as per Fig. 4.3. The maxillary sinuses are large pyramoidal hollows, each connected by ostia of 2-4mm in the superior, posterior nasal tract wall. The ethmoidal sinuses consist of a varying number of air cells and hence can have several ostia. The sphenoidal sinus is behind the superior concha, connected by ostia behind each. The frontal sinus is directly above the nasal bone, anterior to, and slightly above

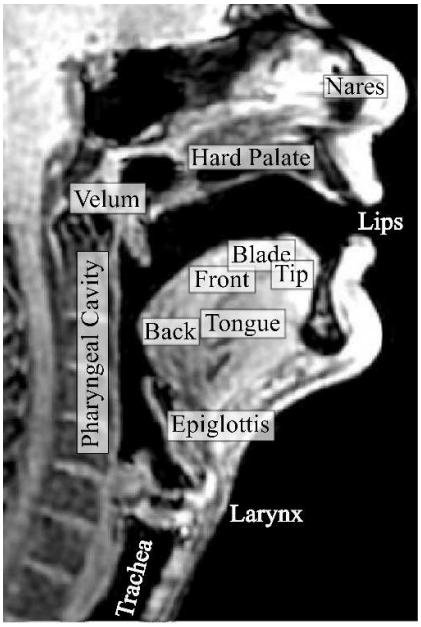


Figure 4.2: Midsagittal Section of the Adult Male Head with Annotated Vocal Anatomy

the superior conchae. The sphenoidal and frontal sinuses are both split by a septum, which Sinnatamby observes is rarely along the midline, leading to asymmetry in each [43].

At the root of the tongue is the epiglottis, which extends upwards into the oropharynx during phonation, as per Fig. 4.2. A membrane known as the aryepiglottic fold surrounds the path towards the larynx, folding down with the epiglottis during swallowing to seal the airway. Surrounding this fold is the thyroid cartilage, prominence in which is widely known as the 'Adam's apple'. The vocal folds (or cords) are the superior tips of the cricothyroid ligament, suspended over the larynx between the arytenoids (two hinged cartilages) and the back of the thyroid cartilage. The arytenoids, in combination with larger laryngeal movements, serve to adjust the position and tension in the vocal folds. During breathing the arytenoids are in a fully abducted position so as to not interfere with airflow. The surfaces of the folds are covered by the soft vestibular folds (sometimes called the 'false' vocal folds).

Below the larynx is the trachea, a ribbed cylindrical cavity composed of

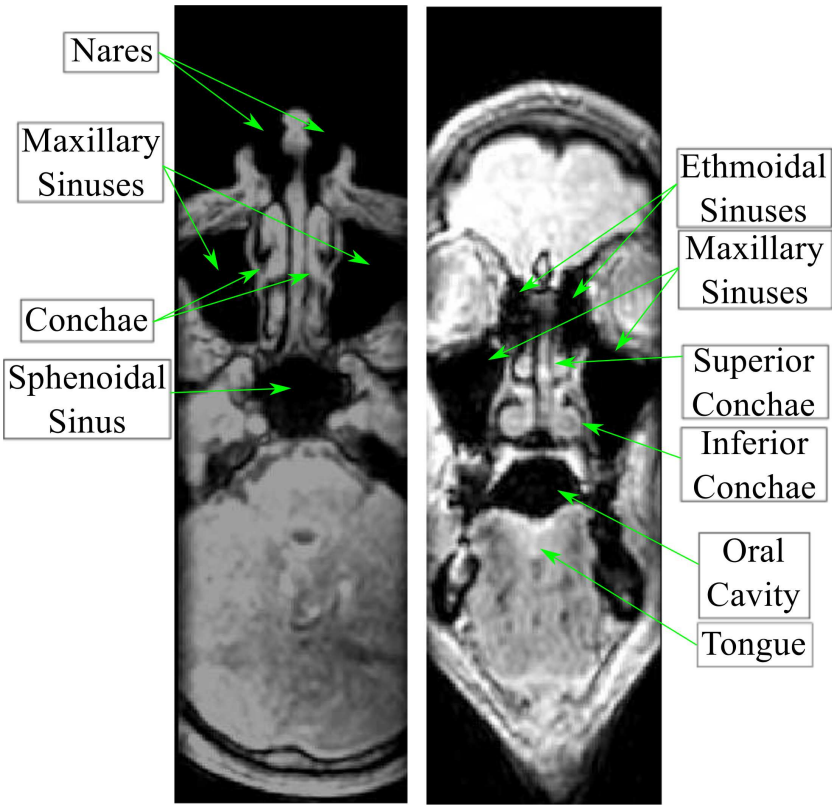


Figure 4.3: Axial (left) and Coronal (Right) Views of the Adult Male Head with Annotated Nasal Anatomy

15-20 rings of cartilage. Approximately 10cm long and 2cm in diameter, it can expand to a length of 15cm during full inhalation and is also capable of expanding and contracting significantly in diameter [43]. The trachea is coupled to the left and right main bronchi (of approximate lengths 5 and 2.5cm respectively), which couple with the lungs via the lobal and segmental bronchi. The walls of the trachea are lined with mucus and cilia.

4.2 The Sound Source

During voiced phonation the vocal folds are adducted, leading to constriction of the glottis. The resulting increase in flow between the folds leads to further adduction by the Bernoulli effect, until contact, whereupon the airway is completely obstructed. This obstruction causes a pressure drop above

the folds and a consequent pressure differential. Since the adductive force is no longer generated the pressure differential causes the folds to separate past their original equilibrium point before consequent constriction, beginning a harmonic motion. This behaviour is often called the glottal cycle, characterised by the closed quotient and open quotient. The closed quotient describes the fraction of a cycle for which the folds are closed, with the open quotient likewise for an open condition. Fig. 4.4 shows a typical glottal cycle in terms of vocal fold contact area as measured by electroglottography (see section 4.8).

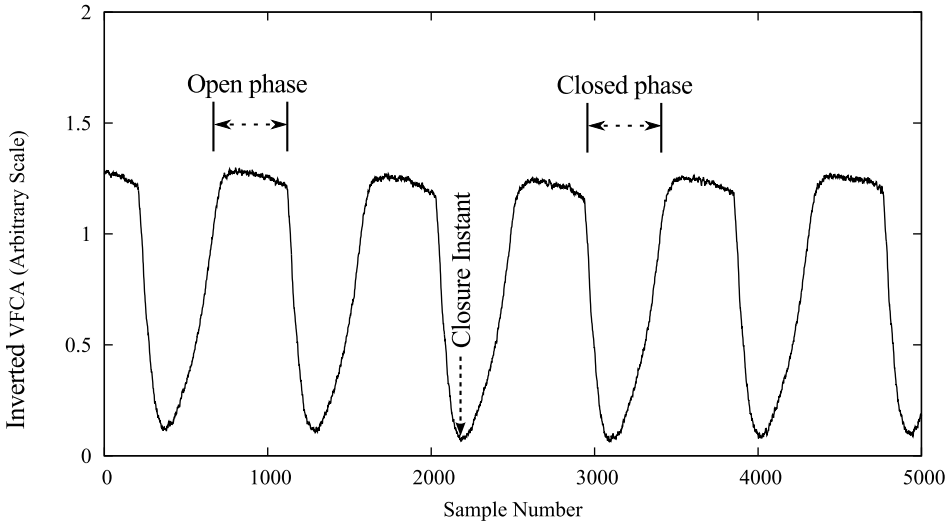


Figure 4.4: Inverted Estimate of Vocal Fold Contact Area Measured using EGG for Adult Male Tenor Voicing /ɜː/ at 220Hz

During voiced speech the glottal cycle produces a strong fundamental of average value 120Hz for men and 210Hz for women [44]. This fundamental is joined by a rich harmonic series, decaying at approximately -12dB/Oct in spoken voice and -9dB/Oct in sung voice [4].

4.3 Sound Modifiers

The vocal tract is considered to describe the oral acoustic propagation path above the larynx, encompassing the pharyngeal and oral cavities. The contri-

bution of the nasal cavity to phonation is typically treated separately. The vocal tract has a particularly characteristic acoustic response to excitation by vocal fold vibration. Its response is largely dependent on its shape, which can be manipulated by articulation of the vocal apparatus. The tongue, assisted by the mobility of the mandible provides a highly flexible means of altering the shape of the vocal tract. Movement of the lips and velum also contribute to articulation in the vocal tract geometry. The larynx is capable of slight vertical displacement and manipulation of the cross-sectional profile of the pharynx is possible.

4.3.1 Vowels

Vowels are the fundamental building blocks of voice-borne information delivery, constituting the smallest differentiable phonetic units that can be produced. Fant's acoustic theory of speech production describes formants as the final, dominant frequency peaks produced after filtering of a voiced or unvoiced sound source by the acoustic transfer function of the vocal tract [45]. In this theory the vocal tract and sound source are seen as linearly separable acoustic functions. Vowel determination depends upon the frequency and to a lesser degree bandwidth of the first two resonant peaks (formants) in the vowel spectrum [2]. The presence of a formant is dependent on an effective frequency-domain sampling by the rich harmonic series of the source function. Table 4.2 gives an overview of average measured formant frequencies for a range of vowels.

The shape of the vocal tract is the fundamental contributor to its transfer function (vocal tract transfer function VTTF), a shape directly manipulated by the vocal articulators. Correlating the VTTF with corresponding articulation is an ongoing challenge in speech research, and is addressed to a limited extent by acoustic phonetics.

4.3.2 Acoustic Phonetics

Acoustic phonetics is a field of research concerned primarily with the acoustics of voice production and consequent phonetic quality. It allows the physics

	Men			Women			Children		
Vowel	F1	F2	F3	F1	F2	F3	F1	F2	F3
/i/	270	2300	3000	300	2800	3300	370	3200	3700
/ɪ/	400	2000	2550	430	2500	3100	530	2750	3600
/ɛ/	530	1850	2500	600	2350	3000	700	2600	3550
/æ/	660	1700	2400	860	2050	2850	1000	2300	3300
/ɑ/	730	1100	2450	850	1200	2800	1030	1350	3200
/ɔ/	570	850	2400	590	900	2700	680	1050	3200
/ʊ/	440	1000	2250	470	1150	2700	560	1400	3300
/u/	300	850	2250	370	950	2650	430	1150	3250
/ʌ/	640	1200	2400	760	1400	2800	850	1600	3350
/ɜ/	490	1350	1700	500	1650	1950	560	1650	2150

Table 4.2: Average Formant Frequencies for Men, Women and Children after [2]

of voice production to be explored and for consideration of the acoustic-articulatory inversion of speech waveforms. The latter has applications in voice compression, voice morphing, synthesis and speech recognition. Traditional acoustic phonetics replaces the vocal tract with an analogue of a more simply appreciable response, be it mechanical or mathematical.

In the vocal tract one-dimensional axial resonant behaviour is dominant, responsible for formant reproduction as a function of the changing cross-sectional area of the vocal tract and the terminations at either end [46]. These behaviours can hence be reproduced with a one-dimensional chain of waveguides, such as in the case of the Kelly-Lochbaum model. The most simple acoustical representation of the vocal tract (at least for a closed glottal condition) is the quarter-wave cylindrical resonator. This simple uniform cylinder is most closely matched to IPA vowel /ɜ/, corresponding to a predominantly neutral vocal tract position. Performing a one-dimensional approximation of the behaviour for this arrangement is straightforward, as demonstrated in section 2.4. Consider now a closer approximation to the vocal tract, whereby it is instead represented by the concatenation of two cylinders of distinct

cross-sectional area. The first (closed at the far end) represents the pharyngeal condition and the second the oral condition, where both are free to change length and diameter in correspondence with reasonable anatomical limits. The resonant frequencies of such a system are not represented by superposition of each constituent cylinder's response, since coupling affects the behaviour of each [46, 47]. This coupling can be taken into account by considering the reactivity presented by one cylinder to the other, as per (4.1), where A_n and l_n are the cross-sectional area and length of cylinder n respectively and k is angular wavenumber. For frequencies satisfying the condition of (4.1), cancellation of the reactive components leads to resonance.

$$-\frac{\rho c}{A_1}\cot(kl_1) + \frac{\rho c}{A_2}\tan(kl_2) = 0 \tag{4.1}$$

The test case, for two cylinders of identical cross-sectional area and length is plotted in Fig. 4.5. It can be seen that all reasonable, non-trivial resolutions of (4.1) occur at frequencies matching those of a simple cylindrical quarter-wave resonator with pure-real terminations.

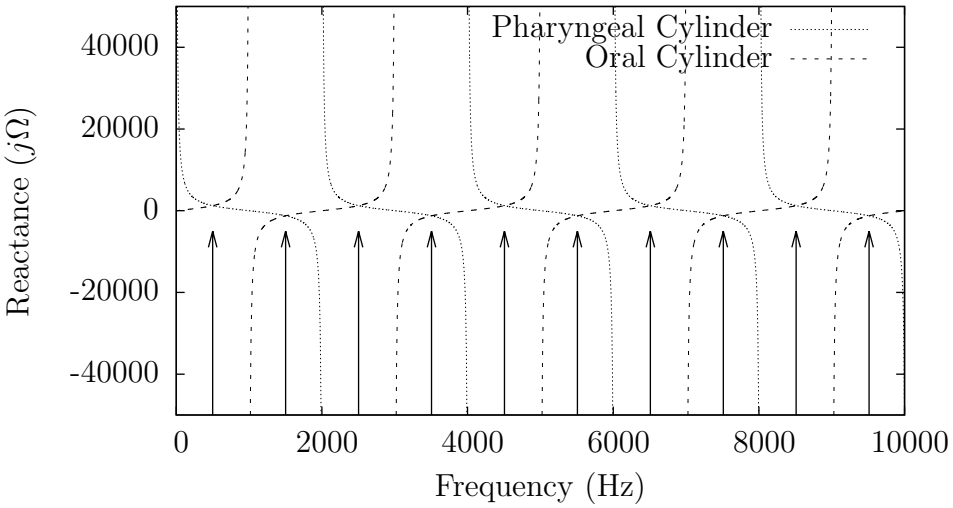


Figure 4.5: Reactivity in Two-Cylinder System of Identical Lengths and Cross-Sectional Areas

Next, consider a two-cylinder analogue of the vowel /ɑ/, featuring a constricted pharyngeal section propagating into a more expansive oral sec-

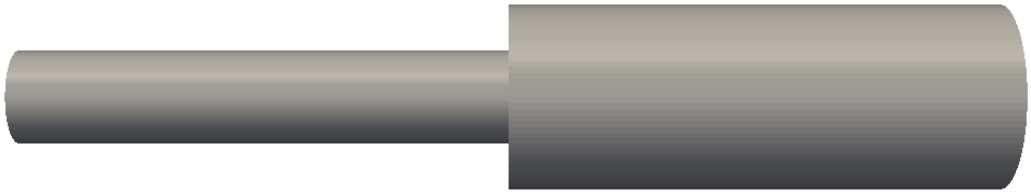


Figure 4.6: Concatenated Cylindrical Analogue of /ɑ/ - Dimensions $0.085 \times 0.008 \rightarrow 0.085 \times 0.016$

tion, as per Fig. 4.6. In this case, the pharyngeal cylinder has dimensions $0.085 \times 0.008m$ and the oral cylinder $0.085 \times 0.016m$. Fig. 4.7 demonstrates the conditions for which (4.1) is satisfied, leading to formant transitions consistent with movement between vowel configurations /ɜ/ and /ɑ/ as per Table 4.2.

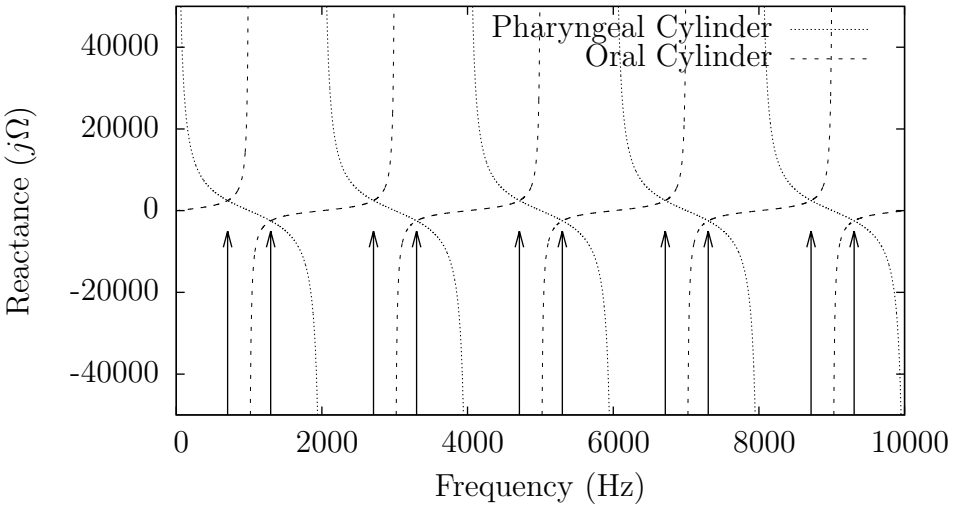


Figure 4.7: Reactivity in Two-Cylinder System of Dimensions $0.085 \times 0.008 \rightarrow 0.085 \times 0.016$ as a Mechanical Analogue of /a/

Pursuing this model slightly further can illuminate an interesting feature of vocal tract modelling. Two vocal tract analogues are shown in Fig. 4.8. Fig. 4.8a shows a short, constricted pharyngeal cylinder propagating into an extended, more expansive oral cylinder. Fig. 4.8b by contrast shows a long, narrow pharyngeal cylinder coupled to a short, expansive oral cylinder. These are both representative of conceivable vocal tract configurations. Ap-

(a) $0.0425 \times 0.008 \rightarrow 0.1275 \times 0.016$

(b) $0.1275 \times 0.008 \rightarrow 0.0425 \times 0.016$

Figure 4.8: Concatenated Cylindrical Analogues Demonstrating Resonant Non-Uniqueness

proximation of their resonant mode frequencies by (4.1) is shown in Figs. 4.9 and 4.10 respectively. Despite each demonstrating overtly different reactive trends, resonance is seen to occur at exactly the same frequencies in both. While (4.1) of course represents a significant approximation, this suggests that radically different vocal tract configurations can produce similar resonant behaviours at low frequencies where the one-dimensional assumption holds. This condition is described as acoustic-articulatory non-uniqueness, where the acoustic to articulatory mapping is not on a one-to-one basis [45].

Linguists relate vowels to articulation by the vowel space, as per Fig. 4.11 which maps phones based on the lengthwise point of constriction in the vocal tract and the degree of vertical constriction. A relationship can be established between vertical and horizontal transitions in the vowel space and formant shifts, although a correlation with absolute articulator positions cannot be drawn, since multiple sympathetic articulations can contribute to the same overall effect. Some efforts have been made to correlate movements in the vowel space with constriction in the vocal tract through perturbation theory, by which formant movements are predicted based on knowledge of the point of constriction and the likely position of nodes and anti-nodes in air pressure [48, 49]. While informative, perturbation theory is based on assumptions too

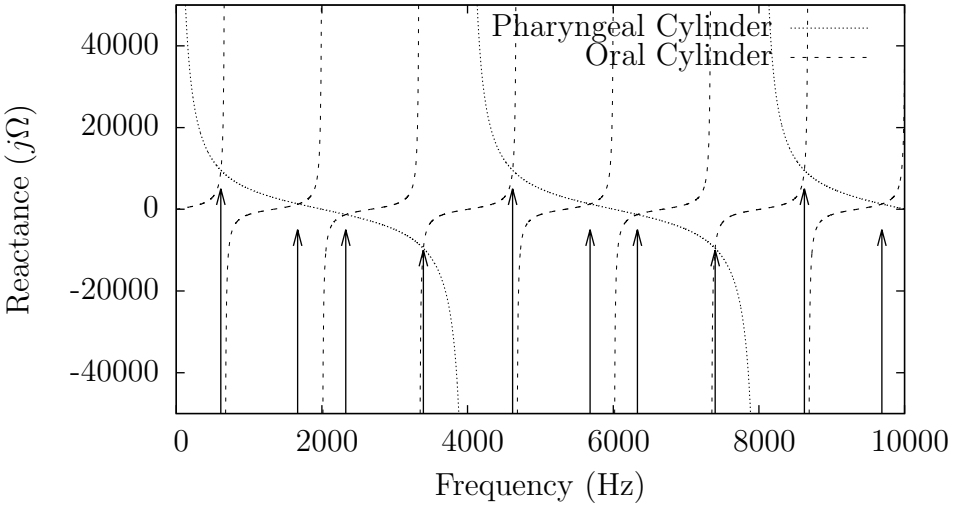


Figure 4.9: Reactivity in Two-Cylinder System of Dimensions $0.0425 \times 0.008 \rightarrow 0.1275 \times 0.016$

broad to be of genuine application to human speakers [50]. Indeed, any one-dimensional cylindrical representation presents a significant simplification of the vocal tract geometry.

By increasing the number of cylinders used to represent the vocal tract the geometrical accuracy of the analogue can be increased, although its accuracy will always be constrained to beneath the lowest frequency tangential resonant characteristic which for some configurations can be as low as 3kHz. Beyond a small number of cylinders, mathematical approximation of the system also quickly becomes non-trivial as the behaviour of each cylinder becomes dependent on all coupled combinations of its neighbouring cylinders [51, 52]. A preferable approach in this case is a scattering-based model of the cylindrical analogue, as explored in section 5.6.2. Chiba and Kajiyama observed that the VTTF could be better approximated using a series of continuous mechanical analogues, as in Fig. 4.12 [53]. These analogues represent the cross-sectional area changes in the vocal tract, approximated from early mid-sagittal X-ray imaging. They can be seen to produce gradual constrictions and expansions in the acoustic path, which is better representative of the articulations required for changing phonetic quality in vowel reproduction

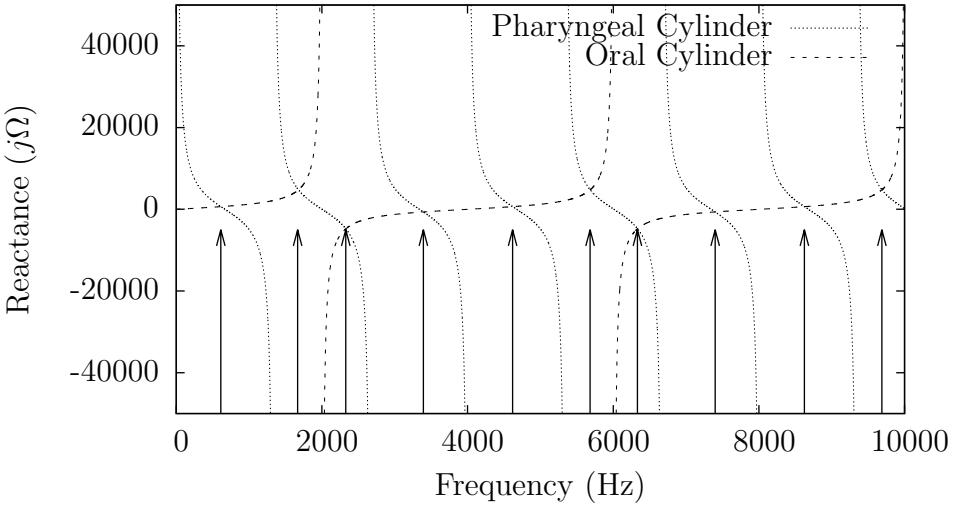


Figure 4.10: Reactivity in Two-Cylinder System of Dimensions $0.1275 \times 0.008 \rightarrow 0.0425 \times 0.016$

[9].

4.4 Frication and Plosives

Frication

Where air flow is forced through a constriction in the vocal tract the flow can become turbulent (see section 2.6). This turbulence can function as an additional noise-like sound source termed frication. This is the primary sound source in unvoiced speech, which does not feature the spectral slope of voiced speech. It is possible to generate *mixed* sounds, which use a combination of vocal fold vibration and frication as an acoustic source. Fricatives are identified by the point at which they are generated. Fig. 4.13 shows typical places of constriction, representing dental, labio-dental, alveolar, and palato-alveolar fricatives. For glottal fricatives the place of constriction is the glottis itself, hence the vocal tract configuration is not explicitly responsible for sound generation.

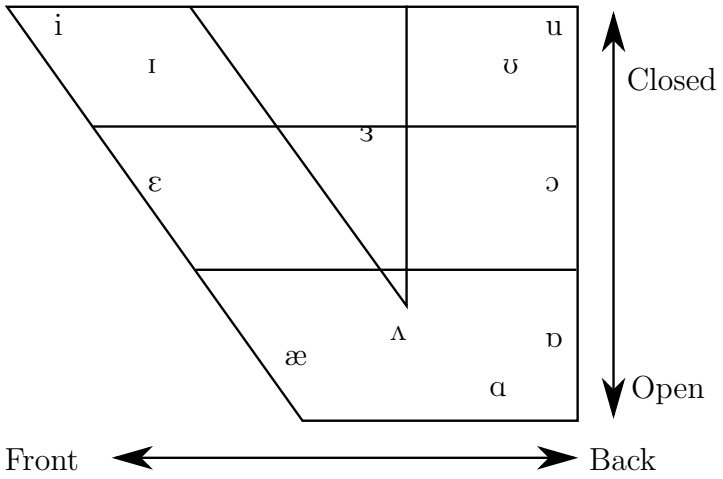


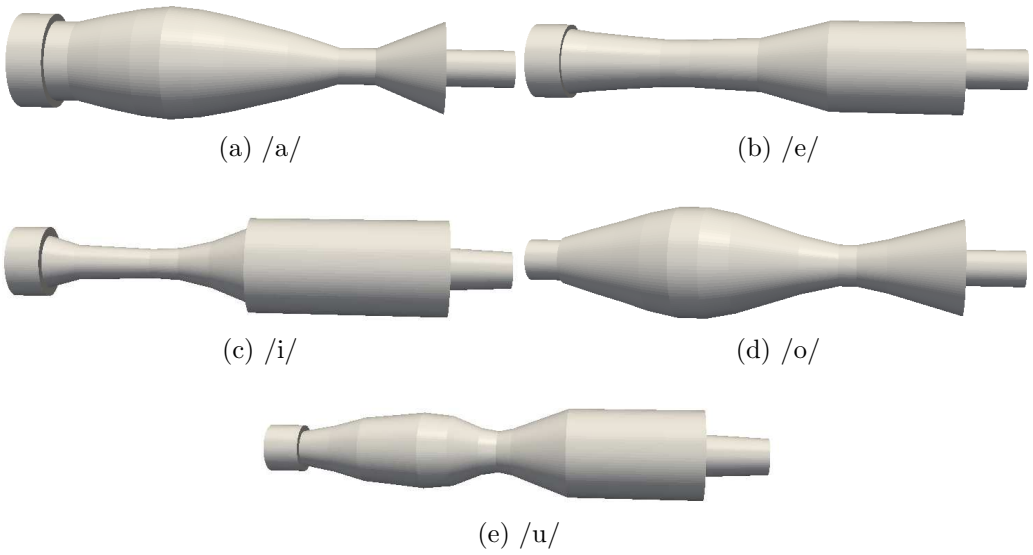
Figure 4.11: Vowel space for IPA vowel symbols, after [4]

Plosives

Complete obstruction of the vocal tract under flow conditions causes a significant pressure differential across the constriction. Sudden relaxation of this obstruction causes a pressure burst, the centre frequency of which is related to the point of constriction [4]. As with fricatives, plosives are identified by the point of constriction, as shown for bilabial, alveolar and velar plosives in Fig. 4.14.

4.5 Lip Radiation

At the lips there is a significant discontinuity in acoustic impedance between the oral cavity and the surrounding air, at which wave components will be reflected and transmitted into the free field accordingly. This change in acoustic impedance is largely a function of lip position, which can be manipulated to provide a variable filtering effect. This filtering is known as lip radiation and is often described by the complex radiation impedance at the mouth [54]. Lip protrusion during speech (or indeed singing) also causes an elongation of the vocal tract, changing its resonant and radiative behaviour [55]. The complex radiation impedance can be approximated by an infinitely flanged cylinder, giving a magnitude which increases with frequency at a gradient



(a) /a/ (b) /e/

(c) /i/ (d) /o/

(e) /u/

Figure 4.12: Mechanical vocal tract analogue geometries for five Japanese vowels after [9]. The 'mouth' is on the left and the 'glottis' to the right of each model

dependent on the size and shape of the aperture [11].

4.6 The Nasal Cavity

The nasal tract presents a secondary propagation path for the voice source, which can be coupled and decoupled from the vocal tract at the velum. While the vocal tract is considered to provide the fundamental filtering function in source-filter theory of the voice [45], the nasal tract provides an additional filtering effect. It is also fundamental to generating 'nasal' phones, whereby the vocal tract is occluded by articulator contact leaving the nasal tract to function as the sole propagational path and the vocal tract a terminated branch. Nasal phones are referenced by the point of occlusion in the vocal tract, as demonstrated for bilabial, alveolar and velar nasals in Fig. 4.15.

When coupled, the nasal tract contributes significantly to the overall transfer function, introducing additional poles and zeros [56, 57, 58]. Coupling of the sinuses and paranasal cavities too has been shown to lever an influence on the spectra of nasal phones, and interestingly non-symmetry

between the two nasal passages has been seen to introduce extra pole-zero pairs [59, 60]

4.7 Glottal Coupling

A common assumption made in vocal tract modelling is that of linear separability of the voice source and filtering functions. This assumption fails in several ways. The acoustic load presented at the larynx can affect the frequency of vocal fold vibration [42, 61]. Vocal fold opening/closure causes formant shifts by extension of the resonating structure [42]. Bandwidth modulation of the formants can be induced by changing damping conditions [62], and the subglottal structure itself can introduce additional formants [63]. To further complicate the nature of glottal coupling, each of these effects is modulated by the cross-sectional area of the glottis [63]. These coupling effects constitute a fundamental breakdown in the typical assumption of linear separability in the vocal source and tract. Accurate formant reproduction remains the most significant requirement in vocal tract modelling. By modelling the vocal tract in a condition of glottal closure a strong approximation of formant behaviours can be made with a one-dimensional analogue, since most coupling effects modulate formant bandwidths and frequencies about this condition [62]. While the response of a vocal tract model assuming linear separability can be considered correct for this closed glottal condition, it must be remembered that this will not necessarily be representative of natural phonation.

4.7.1 Subglottal Formants

The glottis is often modelled as being permanently closed, yielding an infinite acoustic impedance and completely disregarding sub-glottal behaviour. The vocal tract is coupled via the trachea to the lungs by a complicated bronchial tree [43]. This structure has a complicated resonant behaviour, often characterised by subglottal formants [63] or a complex characteristic glottal impedance [46], although sub- and supra-glottal interaction is non-

constant during voiced phonation. A static vocal tract model constructed for glottal closure is hence unlikely to produce a complete frequency domain representation outside this condition.

4.7.2 Piriform Fossa

The piriform fossa are two small pockets in the larynx, sitting in the piriform recesses behind the thyroid cartilage surrounding the larynx and the vocal ligament as in Fig. 4.16 [43]. Often neglected in speech production models they function as coupled branches either side of the voice source, introducing low frequency troughs and influencing early formant frequencies [64]. Study of the piriform fossa is complicated by their small dimensions, requiring accurate three-dimensional imaging.

4.8 Voice Measurement

There are various forms of empirical data, both dynamic and static, we can use to describe the human vocal system. The usual demands are for measurement of articulator positions, characterisation of source functionality and determination of cross-sectional area functions. Dynamic point-wise determination of articulator positioning is possible through electromagnetic articulography and electropalatography, although these provide only a very limited description of the overall vocal tract configuration. The most complete description of the vocal tract is collected through static imaging, traditionally obtained through X-Ray imaging or computed tomography, and more recently through magnetic resonance imaging. Dynamic models of the vocal tract often begin with a highly detailed static model which is consequently warped using dynamic articulator measurements [65].

Characterisation of voice source function is widely approached by electroglottography, by which a trace representative of vocal fold contact area can be obtained. While the trace cannot be considered to provide an accurate portrayal of air flow through the glottis, its rich harmonic series is representative of the acoustic waveforms generated by vocal fold vibration.

An example of electroglottography is shown in Fig. 4.4.

X-Ray imaging provided the earliest non-invasive insights into the vocal anatomy [45, 53], allowing for capture of a single plane of the body in any dimension. Typically used for mid-sagittal imaging of the vocal tract, the consequent data can be used to predict the cross-sectional area function of the vocal tract and analyse articulator positioning. The use of X-Ray imaging in targetted studies is no longer common due to ethical considerations raised by the harmful radiation dose delivered. This has also restricted the use of X-Ray computed tomography, which uses a rotating scanning system to rebuild an image in more than one plane.

4.8.1 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) has quickly established itself as an invaluable asset for speech research. Non-invasive and with no known long-term side effects it presents a means of developing highly accurate multi-dimensional models [3, 66]. MRI immerses the subject in a strong, constant magnetic field, induced by a super-cooled magnetic bore in the scanner itself. This field causes a sympathetic magnetic alignment of hydrogen atoms in the tissues of the body. A radio frequency (RF) pulse is transmitted by the scanner, knocking the magnetic alignment of each atom partially or completely onto a normal transverse plane. In returning to its equilibrium alignment (a process known as relaxation) a decaying RF echo is emitted by each particle at the same frequency as the initial RF pulse and is recorded by a receive coil in the scanner. RF echoes are spatially separated by generating short-time gradients in the magnetic field across the scan region. These gradients can be induced independently across all three axes by gradient coils in the scanner, allowing complete three-dimensional localisation [3]. Two features of the received echo are used to characterise the tissues from which they were emitted. The first is spin-lattice relaxation time (or T_1), describing the rate of recovery in the direction of magnetic alignment. The second is spin-spin relaxation time (or T_2), which describes the decay of the signal in the transverse plane [3]. Different tissue types will produce different T_1 and T_2 times

as a function of the proton density (PD) in those tissues, as summarised in Table 4.3.

Tissue Type	T_1 (ms)	T_2 (ms)	PD (%)	Examples
Fluid	$1500 - 2000$	$700 - 1200$	> 95	
Water-Based	$400 - 1200$	$40 - 200$	60-85	Muscle, brain, cartilage
Fat-Based	$100 - 150$	$10 - 100$	60-85	Fat, bone marrow

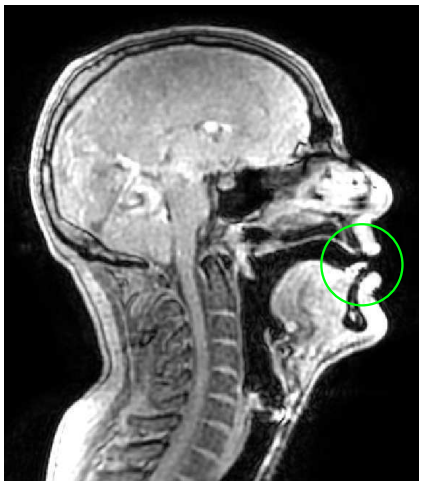
Table 4.3: Typical Tissue Types, and Characteristic Properties - after [3]

The sensitivity of the imaging to any of these three characteristic quantities is dependent on the nature of the scanning sequence. The contrast of a T_1 *weighted* image is largely dependent on the T_1 time of each tissue, likewise for T_2 weighted and PD weighted imaging. This sensitivity can be configured by adjusting the time between successive RF pulses (repetition time - TR) and the time from the RF pulse to its echo (or echo time - TE) [3]. Differing tissue types will then be observable as changes in the brightness of the imaging, allowing different tissue regions to be segmented by detecting contrast boundaries. Further examples of such MR imaging of the vocal tract are provided in Chapter 6.

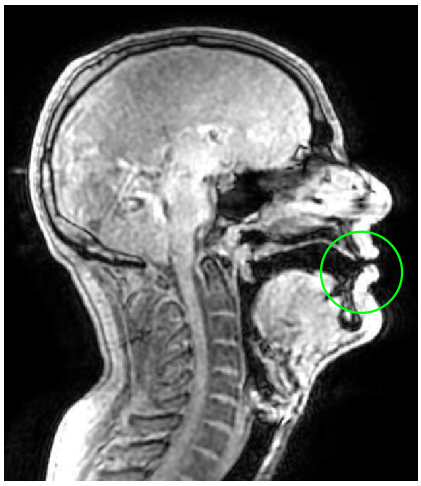
Summary

In this chapter the human voice has been introduced. The vocal anatomy has been explored, with particular consideration paid to its meaning for the acoustics of voice production. Common means of collecting data regarding the vocal anatomy have been explored. The human voice is an extremely complex instrument and even after decades of research resynthesising its behaviour represents a considerable, multifaceted undertaking. One of the most fundamental assumptions made in voice modelling is that of linear separability of the vocal tract and the vocal source. While interactions between the two have been identified, the assumption largely holds and significantly

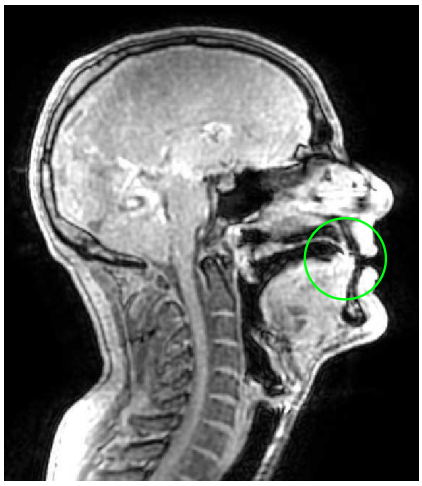
reduces the challenge of physical modelling the voice production system by separating the two functions. It is for this reason that the work described in this thesis will consider only the acoustic response of the vocal tract. In Chapter 5, different approaches to time-domain numerical modelling of the vocal tract (and the more general voice production system) will be introduced and assessed.



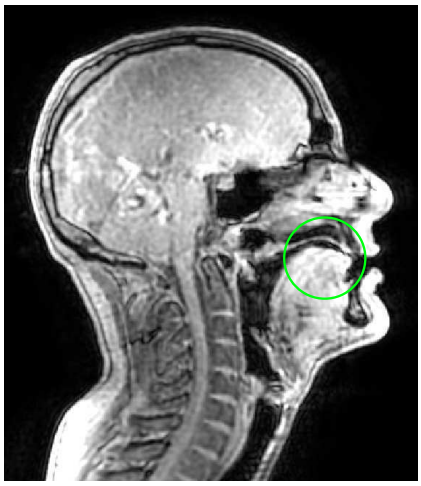
(a) Dental



(b) Labio-Dental

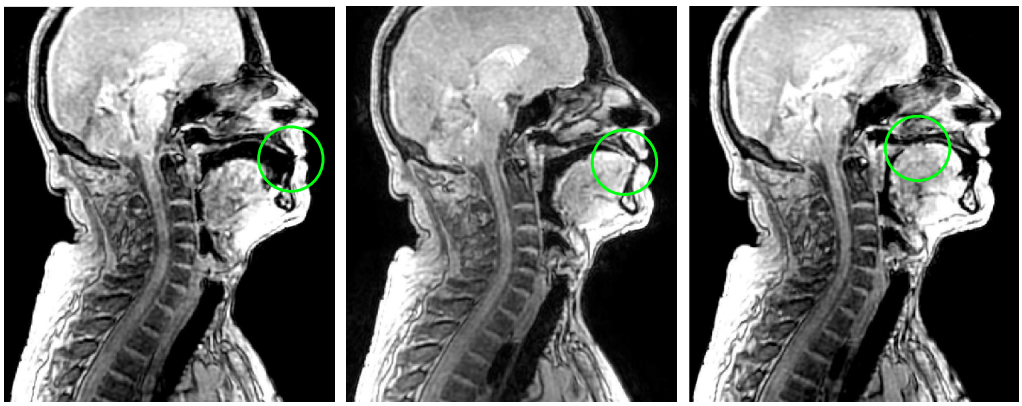


(c) Alveolar



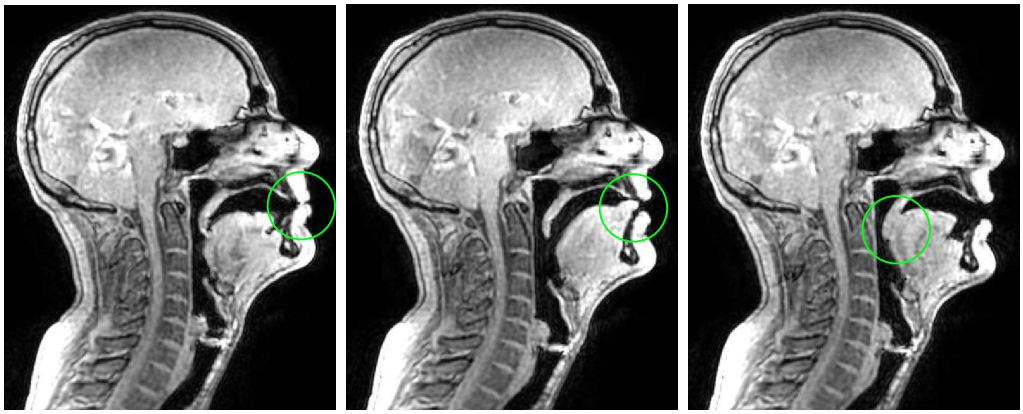
(d) Palato-Alveolar

Figure 4.13: Midsagittal Imaging of Common Fricative Vocal Tract Configurations for an Adult Male



(a) Bilabial (b) Alveolar (c) Velar

Figure 4.14: Midsagittal Imaging of Common Plosive Vocal Tract Configurations for an Adult Male



(a) Bilabial (b) Alveolar (c) Velar

Figure 4.15: Midsagittal Imaging of Common Nasal Vocal Tract Configurations for an Adult Male

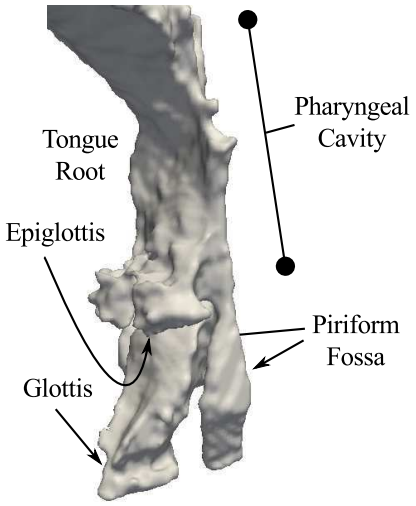


Figure 4.16: Graphical model of the pharyngeal cavity and larynx indicating the piriform fossa

Chapter 5

Voice Modelling

Introduction

Having introduced techniques for acoustic simulation in Chapter 3 and the nature of the voice as a problem domain in Chapter 4, this chapter combines the two by considering techniques for numerical simulation of the voice. It begins by revisiting the implications of source-filter separation in voice modelling paradigms, before investigating the acoustic contribution of the vocal tract and vocal source more closely. The manner in which geometrical data is used in voice synthesis is addressed. The design philosophy and theory behind physical models of voice production are described. Consideration is paid to how the digital waveguide mesh has been previously applied to voice simulation in a two-dimensional guise and the opportunities afforded by the dynamic digital waveguide mesh are introduced. The chapter concludes with an appraisal of comparable approaches to numerical simulation of voice production.

There are three fundamental requirements for a model of the voice, naturalness, accuracy and intelligibility, the relative importance of which depends on the priorities of the particular application. Natural voice synthesis, as described in section 1.2, is that in which the human listener can not deduce whether the output produced is a real human phonation or a simulated equivalent. Accurate modelling of the voice implies that synthesised audio

corresponds exactly to the performance of the physical acoustic behaviour of system the model is analogous to. It is expected that an accurate numerical simulation of the human vocal process will yield both intelligible and natural voice. A model can however produce accurate and/or natural voice without being completely accurate as demonstrated here for simple physical models. Intelligibility is a core requirement where the system is used for communication and essentially encompasses the ability of a model to reproduce essential phonetic features for interpretation of the voice waveform. The most fundamental and frequently targetted is formant reproduction, by which vowel identification is possible (as in section 4.3.1). For this reason, models of the voice are often examined in terms of their frequency-domain performance. It follows that when the geometry of a resonator is adequately described, its consequent acoustic behaviour will be closely reproduced. This is very much the aim of physical models, whereby simulation of the acoustic behaviour of the voice is approached by representation of the geometry of the vocal apparatus under the assumption of source-filter separation.

5.1 Source-Filter Separation

As mentioned briefly in section 4.3, Fant's acoustic theory of speech production delineates the voice production process into separate acoustic functions; the source and the filter [45]. The source in this case is the acoustic waveform generated at the vocal folds and the filter constitutes the acoustic vocal tract transfer function (VTTF). Source-filter separation is a useful concept as it breaks the challenges of physically modelled voice synthesis into two sub-domains, although it has been demonstrated that the VTTF and the source function are not exactly linearly separable, as explored in section 4.7.

5.2 The Vocal Tract

The acoustic behaviour of the voice is dominated by the frequency response of the vocal tract, as explored in section 4.3. This is often described in terms

of formant frequencies and bandwidths. Above 4kHz, resonant modes are not widely referred to as formants, since it becomes difficult to prescribe phonetic meaning to them. There are a number of other characteristics to be observed in the vocal frequency response, such as notches and pole-zero pairs attributable to the nasal tract and paranasal sinuses (section 4.6) and formant structures attributed to the resonant behaviour of the sub-glottal structure [63] (section 4.7). A complete physical acoustic model of the voice would reproduce all these characteristics, although in practice it is only the early formants that are accurately reproduced in existing physical models. This early resonant behaviour is largely dependant on the axial behaviour of the vocal tract - interpreted as its changing cross-sectional area. These can be effectively reproduced by simple concatenated cylindrical analogues, using perturbation theory and simple acoustic phonetics as demonstrated in section 4.3.2. Models of greater geometrical accuracy and higher dimensionality are hence expected to reproduce higher-order resonant modes with greater accuracy.

5.3 The Voice Source

As explored in section 4.2, the periodic vibration of the vocal folds give rise to a strong harmonic frequency component and a rich, spectrally decaying series of harmonics. Since the source mechanism is rooted in complex aerodynamics, means of implementing an appropriate model in the acoustic domain are not entirely obvious. This is complicated further by the nature of source-filter separation, in that the source model must be appropriate for injection to a heavily abstracted vocal tract model [67]. A common approach is to use a waveform indicative of the air flow through the glottis. A popular, parameterised example is the Liljencrants-Fant (LF) model, describing the glottal cycle (section 4.8) in terms of frequency, amplitude, and constants for growing and decaying exponentials [68, 69].

Another option is the injection of a real source waveform, although this can be extremely difficult and dangerous to obtain. A comparable alternative is injection of real EGG measurements (as in section 4.8). While EGG pro-

vides a measurement of vocal fold contact area instead of flow, the harmonic series of each is similar. EGG has the additional advantage of containing the slight cyclic variations intrinsic to human phonation. The strict periodicity in prescribed source models (such as the LF model) can be extremely damaging to perceived naturalness of the resulting synthesis.

5.4 Geometrical Data

5.4.1 X-Ray

X-Ray imaging provided the first wave of non-intrusive anatomical imaging of the voice, and contributed to improved understanding of the voice production process. Its most significant shortcoming in comparison to more modern techniques is that it only allows image capture in a single two-dimensional plane. While this shortcoming was addressed to some extent by computed tomography, an increased understanding of the health risks presented by X-Ray imaging has led to the restriction of its application in targetted studies due to ethical considerations.

The most commonplace application of X-Ray imaging in developing models of the vocal tract has been the extrapolated measurement of cross-sectional area functions from midsagittal imaging [53, 45]. The vocal tract is spanned at several key positions along its path and this distance used as the diameter of an equivalent circle. In some cases the curvature of the vocal tract is taken into account (developing a polar coordinate system for example [70]), although in most the vocal tract is effectively flattened (resulting in predictable formant shifts [71]). X-Ray data is also useful in the study of articulation, since changes in absolute articulator position can be observed between consecutive scans, if only in the midsagittal plane [45].

5.4.2 Magnetic Resonance Imaging

As explored in section 4.8.1, MRI presents an exciting opportunity for full three-dimensional imaging of the voice, allowing development of graphical

models of the vocal tract. These can be decomposed to generate accurate cross-sectional area functions (as required by simple physical models) or used to develop more geometrically complex models for reproduction of acoustic behaviours at higher frequencies. MRI also has applications in imaging the nasal tract, paranasal sinuses and sub-glottal structures (trachea, bronchi). While certainly possible, imaging of the source is more complicated due to both the intricacy of its geometry and its highly dynamic nature.

Processing of MR imaging with respect to the vocal anatomy presents some unique challenges. The first and perhaps most significant is that the teeth (solid, calcified structures) do not appear on scans, as demonstrated in Fig. 5.1. While it is possible to see the outline of the teeth in the gums and as imprints on the cheek and tongue, exact segmentation of the teeth is not possible.

Figure 5.1: Midsagittal Image for Adult Male Phonation of /iː/ - Demonstrating Absence of Teeth and Mandible

A number of approaches have been taken to tackle this issue, including the use of a film painted over the subject's teeth. While this permits segmentation of the teeth it also introduces significant artefacting in the imaging [72]. It also introduces another alien element to a scanning process where it is desirable to match as closely as possible the conditions of natural phonation. A preferable approach is the reintroduction of a model of

the teeth constructed from dental scans [73], although this too is non-trivial considering jaw mobility and the further ethical considerations introduced by X-Ray-based scanning procedures.

An additional challenge in MRI scanning is presented by the compromise between capture times and scan resolution. MRI scanning is widely used to develop highly detailed structural imaging of the brain, using scans that can take up to 20 minutes. For imaging of the human vocal process a separate scan is required for each articulatory configuration, which must remain still for the period of that scan. Shorter capture times result in less motion artefacting in imaging, but come at the cost of image resolution. An effective compromise is therefore sought.

Even for extremely high-resolution scans there is a practical limit to the structure sizes that can be obtained. For example, it can be difficult to discern the ostia connecting the nasal tract to the paranasal sinuses, since the walls surrounding the tract are extremely thin and as solid structures can not be easily identified on scans.

Imaging of the vocal source has been best investigated using purpose built scanning coils [74, 75]. By specifically targetting the larynx, the region of interest can be greatly reduced and scans developed that are better suited to its tissues and structure. Techniques for synchronisation of scan times have also been developed that allow successive images to be acquired at matching intervals of the glottal cycle [76].

5.5 Mechanical Models

The very earliest articulatory models of the vocal tract were elegant mechanisms, boasting the kind of physical analogue that system-models arduously aspire to in the digital domain. The best known is perhaps von Kempelen's speaking machine. This featured a set of bellows for the lungs, a reed-larynx, switchable nasal tubes and a deformable leather tube to represent the vocal tract. While these models could be used to simulate simple vowels, more complicated phonations are nearly impossible due to the unintuitive (and restrictive) interface. One particularly interesting modern application of me-

chanical modelling is in anthropomorphic robotics, an example of which is the Waseda Talker shown in Fig. 5.2.

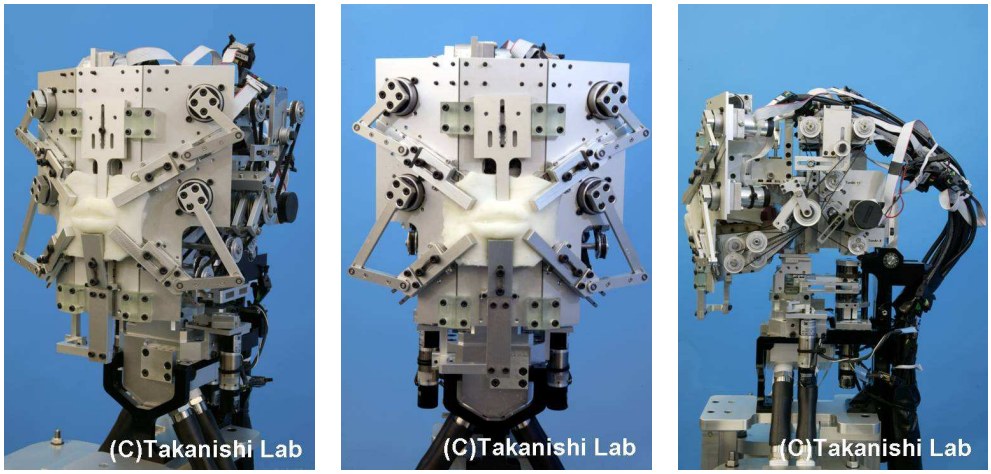


Figure 5.2: The Waseda Talker WT-7RII, from [10]

This type of model combines substitution of the vocal anatomy for mechanical counterparts with modern control techniques for their manipulation [77]. The mechanised control system circumvents the awkward interface of human controlled models but leads to completely deterministic movements. While the results of such biomechanical modelling for single phones can be convincing, replication of the intricate musculature of the human vocal system is extremely difficult, as is reproducing the acoustic and biological characteristics of its constituent elements.

Aside from biomechanical speech synthesis, mechanical models of the vocal anatomy have an important role in the validation of numerical models. Many such constructs have been developed for this purpose and can also be used to aid teaching and learning [9, 42, 78, 79]

5.6 Classic Articulatory Models

Articulatory methods encompass a range of approaches to modelling the voice, whereby the overall effect of the vocal tract (and in some cases other elements of the vocal anatomy) are imitated, as per Fig. 5.3.

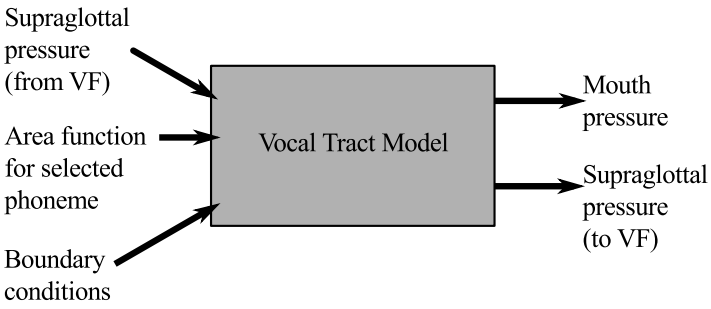


Figure 5.3: Functional Definition of a Vocal Tract Model, after [11]

The application of articulatory modelling to the voice is particularly exciting as it provides an intuitive representation of the voice production process (mirroring the vocal anatomy as a transducer - Table 4.1). This is advantageous as it allows control of voice synthesis on a platform directly related to articulation and phonetics, rather than relying on complex methods for speech waveform transformation and analysis.

As previously explored, concatenated cylindrical analogues provide an effective approximation of the vocal tract. This is particularly the case for low frequencies (incorporating early formants) where the axial acoustic performance of the analogue more closely matches that of the one-dimensional component of the vocal tract acoustic field. While desirable, direct mathematical determination of the acoustic performance of such a one-dimensional analogue is non-trivial beyond the concatenation of 2-3 cylinders. Even in the one-dimensional case, the acoustic response of concatenated cylinders of changing cross-sectional area begins to become complicated. In the case of a chain of cylinders the apparent impedance interface between cylinders is defined by an awkward recursive function of the acoustic impedance of those cylinders which surround it. Direct mathematical approximation of the system is possible [51, 52], but not straightforward. Many solutions employ a matrix, or transmission-line style model whereby each cylinder is represented by its characteristic acoustic impedance. Conditions for pressure and velocity continuity are established (as in section 2.3) and scattering is modelled

accordingly [41]. Rather than being mathematical definitions of the system these approaches result in a system which models the behaviour of the cylinder chain.

Such one-dimensional systems are based on the assumption of planar wave propagation, which does not always hold [5]. An illuminating case is that of two concatenated cylinders, with a step in cross-sectional area. Here a planar wave travels in first (narrow) cylinder until incidence at an interface to a cylinder of greater cross-sectional area. There is a 'step' in the cylindrical geometry, with corners which the propagating wave must instantly fill. This requires a violation of the principles of diffraction and constant wave speed. In fact, for the planar wavefront assumption to be satisfied here the wave speed would have to be instantly infinite. Blackstock resolves that this violation is tenable under the condition that the 'communication time' is much smaller than the wave period [5]. For the small geometries of the vocal tract this error can therefore be largely discounted.

5.6.1 Transmission Line Models

In a transmission line model each successive cylinder is represented by an LC circuit (with series inductance and shunt capacitance) as in Fig. 5.4 [12]. The inductance, L and capacitance, C of each are given by (5.1) and (5.2), where ρ is ambient air density, l is cylinder length, A is the cross-sectional area of each cylinder, c is the speed of sound and k is angular wavenumber [80].

$$L = \frac{\rho l}{kA} \tag{5.1}$$

$$C = \frac{kAl}{\rho c^2} \tag{5.2}$$

By series connection of these LC networks a transmission line model of the concatenated cylindrical analogue can be produced, as in Fig. 5.5.

Fig. 5.5 shows a noise rail used to selectively inject noise at various points of the analogue to approximate frication generated by constricted flow [81].

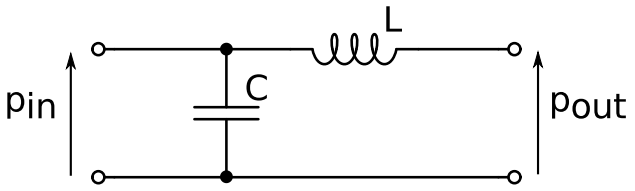


Figure 5.4: LC Circuit to Represent Single Cylindrical Component

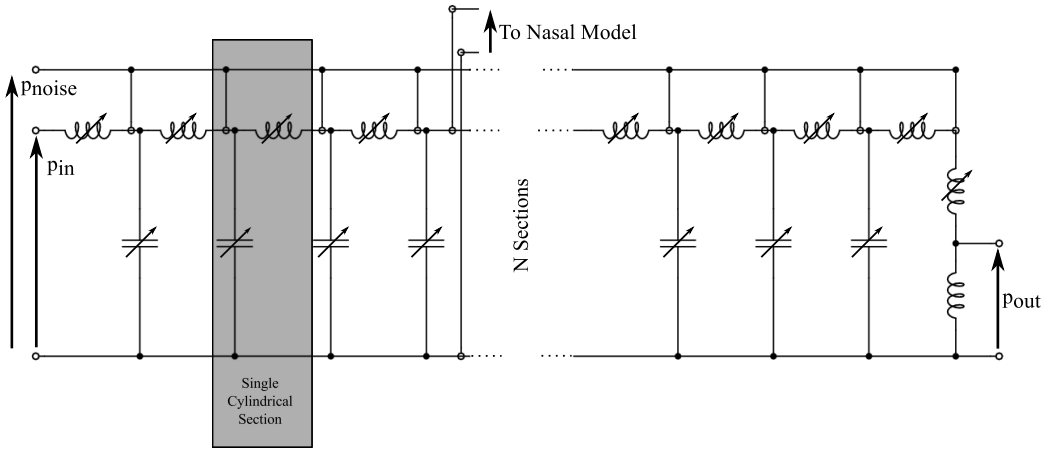


Figure 5.5: Complete Transmission Line Model of the Voice, after [12]

The transmission line is also tapped at the position of the velum to function as the input to a comparable nasal tract transmission line. Output is taken across a shunt inductance to approximate the complex radiation impedance.

5.6.2 The Kelly-Lochbaum Model

Where transmission line models propagate Kirchoff variables (as in section 3.1, the Kelly-Lochbaum (KL) model represents the same methodology for the case of wave digital variables [82].

The KL model represents a chain of concatenated cylinders as a one-dimensional wave variable based scattering network, providing an implicit representation of the one-dimensional lossless wave equation [82]. In this implementation parallel delay lines represent uniform cylinders, according to the d'Alembert solution explored in section 3.3. Discontinuities in characteristic acoustic impedance between cylinders of incommensurate cross-sectional area are simulated using the simple non-homogeneous scattering junction

derived in section 3.4 in Fig. 3.6. The implementation is demonstrated in Fig. 5.6, where Z_n represents the characteristic acoustic impedance of the n^{th} cylindrical element and $r_{x,y}$ represents the reflection coefficient defining wave scattering between elements x and y, according to the scattering formulations of section 3.4.

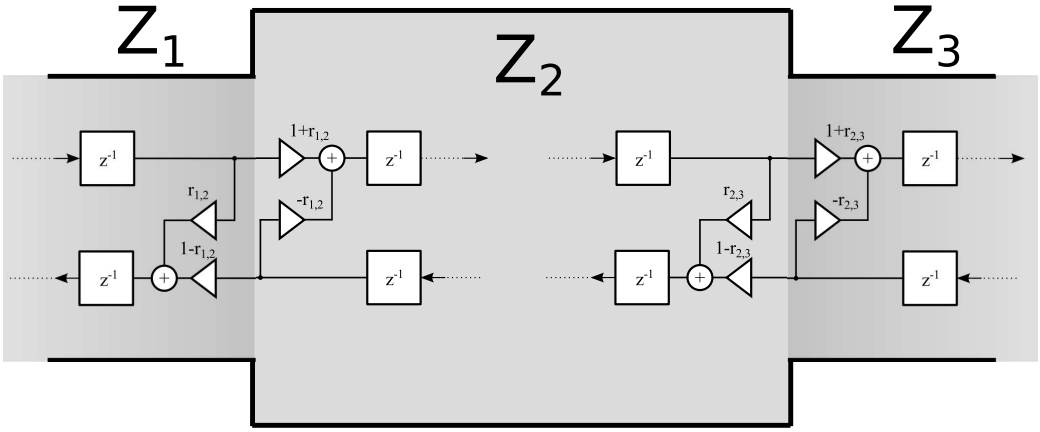


Figure 5.6: Scattering in the Kelly-Lochbaum Model at Acoustic Impedance Interfaces

Injection to the model is by halving the source pressure signal and adding to both delay lines. Similarly, a pressure signal is extracted by summing sympathetic travelling wave components from either delay lines. These functions are demonstrated in Fig. 5.7.

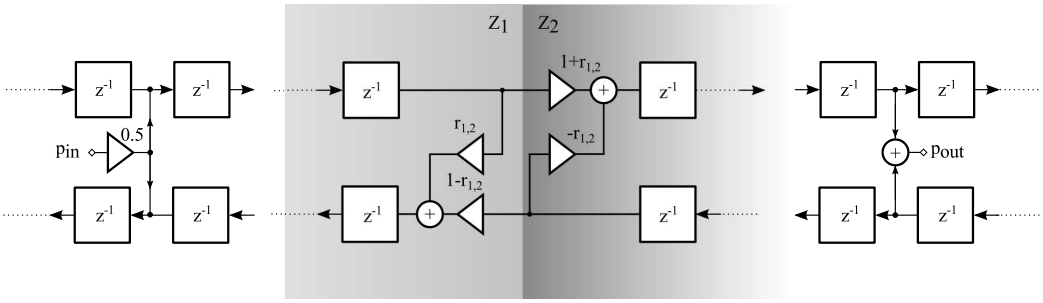


Figure 5.7: Implementation of Injection and Extraction in a Kelly-Lochbaum Model Across an Acoustic Impedance Interface

The model is finally terminated with boundary conditions representative of glottal and lip behaviours. These are represented as $r_{glottis}$ and r_{lip} in

Fig. 5.8, which presents a complete prototype implementation. While the configuration of Fig. 5.8 represents a two-cylinder model, any number of scattering units can be used to represent additional cylindrical units.

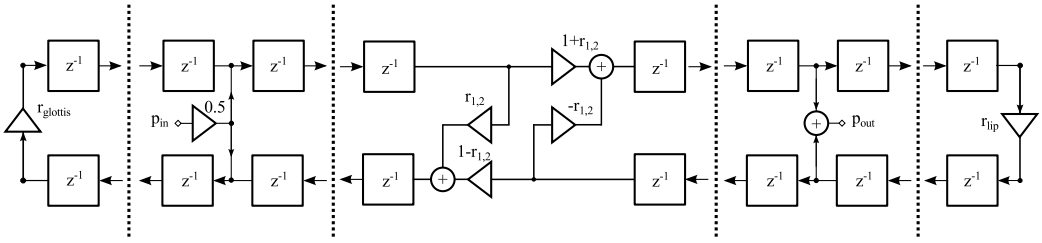


Figure 5.8: Constituent units of a lossless Kelly-Lochbaum model of the vocal tract

The major benefit of using the Kelly-Lochbaum model is its efficient, low cost implementation in the digital domain, as per Fig. 5.8. Within each cylindrical section delay units and losses can be commuted (although in some circumstances retaining a direct physical transformation is preferred to commutation) While it presents a very limited computational demand, the degree to which the model represents the vocal tract is still limited in that the geometrical representation is weak (as with any one-dimensional analogue). Axial acoustic behaviours are described, but tangential and oblique behaviours are not, leading to an incomplete representation of higher order modes. While only low frequency accuracy is required for reliable reproduction of early formants, higher frequency content is important for the perception of naturalness in synthesised voice (see section **??** and [83]). Correspondingly, while phones synthesised with such a one-dimensional model are clearly intelligible, they are not convincingly natural [16, 83].

A number of efforts have been made to increase the accuracy of the Kelly-Lochbaum model (and comparable transmission-line models) by taking further losses into account [62, 12, 81]. These include losses due to wave propagation through the glottis, wave propagation into the vocal tract walls, vibration of the vocal tract walls and sound radiation at the lips [11]. These losses are largely functions of either velocity or pressure and are represented by either series or shunt losses respectively. In the case of the transmission-line model, reproducing these losses can be as simple as introducing extra

LC networks. For the case of the Kelly-Lochbaum model, these losses are implemented by manipulating reflection coefficients (and consequently update equations) according to series and shunt loss factors, corresponding to wave digital filters [62].

5.6.3 The Digital Waveguide

The Kelly-Lochbaum model is essentially a chain of digital waveguides, predating their explicit formalisation [84, 82]. This similarity illuminates the exciting capacity of the network for an intuitive extension into higher dimensionality models (as introduced in section 3.4). By producing two- and three-dimensional digital waveguide meshes it is possible to directly reproduce tangential and oblique resonant behaviours of an appropriate geometrical model, and hence generate a more accurate reproduction of the vocal tract transfer function.

5.7 2D Modelling of the Vocal Tract

Considering the Kelly-Lochbaum model represents a one-dimensional numerical model of concatenated cylinders, Mullen took the next logical step in the development of these physical models and sought to increase the dimensionality of the model. This was performed by generating a two-dimensional membrane, of length and width corresponding to the length and diameter of each successive cylindrical element. The aim was to reproduce tangential resonant behaviours in additional to the axial behaviour and hence increase the lowest frequency for which the simulated system response can be considered accurate. The cylindrical elements of each were based on cross-sectional abstractions of MR imaging [85].

Use of a membrane adds an additional degree of freedom to the vocal tract model. Rather than being represented by scattering units, the discontinuities between cylindrical elements are explicitly reproduced by variations in membrane width. This provides scope for better representation of reflective/absorptive behaviours of the vocal tract wall and hence greater control

over formant bandwidths [86].

Since the vocal tract is of course not two-dimensional, such a model still constitutes only an approximate representation of the VTTF. While oblique resonant behaviour is not expected to be reproduced, it is found that a more significant shortcoming of the membrane model is exhibited by the nature of acoustic coupling in the cylindrical model.

5.7.1 Acoustic Coupling

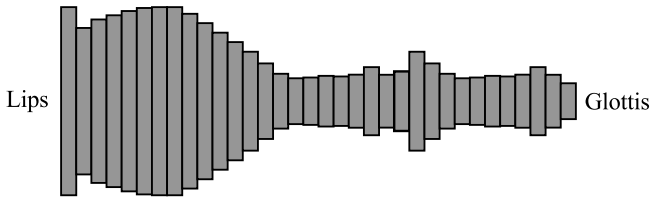
While representing a single bisecting plane of a uniform cylinder with a rectilinear membrane is acceptable, the nature of discontinuity in cross-sectional area introduces ambiguities regarding acoustic coupling. In section 2.4 it has been demonstrated that the characteristic acoustic impedance of a cylinder is a function of its cross-sectional area. In representing each cylinder as a membrane, this relationship no longer holds. Whereas a one-dimensional network uses scattering junctions to explicitly govern reflective/transmittive behaviour, multi-dimensional meshes rely on a correct specific representation of acoustic behaviours to reproduce the overall effect of such acoustic coupling. In the case of a 2D membrane representation, the scattering relationship supported is that of characteristic acoustic impedance as a function of diameter. Mullen circumvented this problem by squaring the effective diameter of each cylindrical section [16]. While this served to ensure one-dimensional (formant) behaviours are accurately reproduced, it means the tangential modes produced are the product of an incorrect geometrical configuration. The two-dimensional mesh essentially presents a choice between correct reproduction of one-dimensional behaviours and correct reproduction of tangential behaviours [47]. While it was observed that two-dimensional modelling of the vocal tract produced phones that were more natural than a one-dimensional equivalent, this perceived increase in naturalness might simply reflect the inclusion of high frequency content that is not purely representative of a uniform cylinder, as explored in section 8.2.3.

5.8 The Dynamic Digital Waveguide Mesh

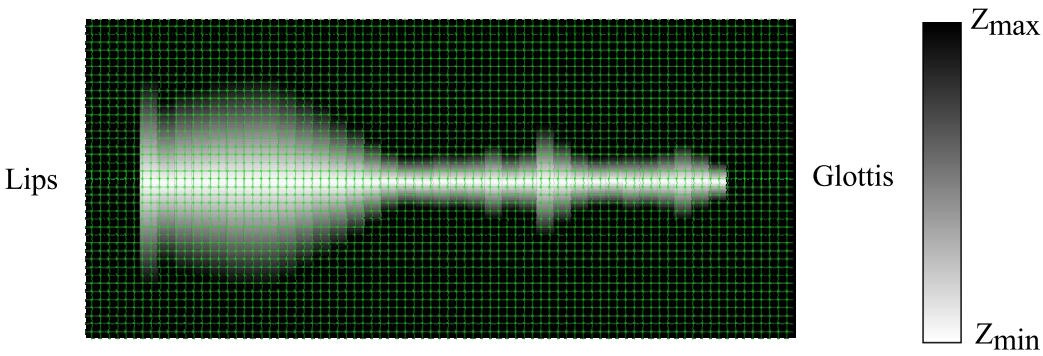
As an addition to the two-dimensional membrane model, Mullen developed a technique permitting real-time manipulation of the effective cylinder diameters. Dynamic extension of the membrane geometry itself is not feasible, often leading to physically inconceivable mesh behaviours. Instead a container mesh was generated, within which geometrical limits are represented by gradients in characteristic acoustic impedance. As explored in section 3.4, scattering in the digital waveguide mesh is governed by reflection/transmission coefficients determined from discontinuities in acoustic impedance. By dynamically changing the characteristic acoustic impedance represented by each scattering node, reflective behaviours can be artificially reproduced at any point in the containing mesh.

In developing a dynamic voice synthesiser, Mullen incurred the additional consideration of restricting computational demand for real-time operation. As explored in section 3.11, the overall computational demand of a DWM system is largely determined by the system sampling rate. It was found that operation of a full container membrane with adequately fine spatial sampling across all conceivable vocal tract configurations was not feasible in real time. Instead, a hybrid method was sought which combined the reproduction of axial resonant modes with an augmented representation of tangential resonant behaviour. To implement such a system Mullen employed a container mesh membrane which was no longer geometrically analogous in width/diameter, and instead used impedance-mapping to reproduce the effect of changing cross-sectional area functions. The raised-cosine impedance map for example applied a function of acoustic impedance across the width of the mesh, varying the weighting of the function appropriately for changing cross-sectional area. This relationship is demonstrated in Fig. 5.9 and 5.9b

The result of this implementation was an effective two-dimensional impedance-mapped dynamic digital waveguide mesh voice synthesiser, named Vocal-Tract.



(a) Concatenated Cylindrical Analogue



(b) Impedance Mapping

Figure 5.9: Linear Dynamic Impedance Mapping of a Concatenated Cylindrical Analogue to a 2D Digital Waveguide Mesh

5.8.1 VocalTract

VocalTract (shown in Fig. 5.10) features a series of vertical sliders, each representing cross-sectional area at regular intervals along the length of the vocal tract. These can each be continuously varied in real-time to manipulate the effect configuration of the vocal tract analogue. A choice of noise, or pitched synthesised source waveforms can be injected at the glottis position allowing the synthesis of monophthongs, diphthongs and glottal frication. It was also found that by closing and suddenly releasing certain area-function parameters (representing the lips, tongue front for example) plosives could be simulated.

The development of VocalTract was informative in many ways. Firstly it demonstrated that mesh-membrane representation of the vocal tract transfer function yields voice synthesis that it more natural than a simple one-dimensional analogue. This was attributed to the presence of non-axial higher frequency content, even though in the case of the impedance-mapped

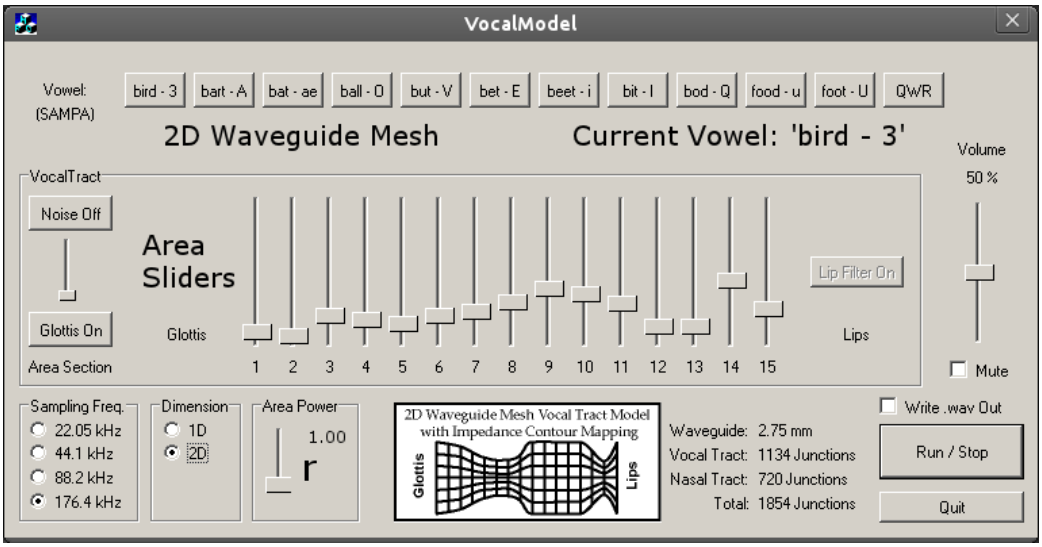


Figure 5.10: User interface for Mullen's real-time dynamic impedance-mapped two-dimensional digital waveguide mesh vocal tract simulation software - VocalTract

2D membrane the tangential geometry is not completely accurate [16]. It demonstrated that by injection of noise, frication can be simulated and also that by sudden release of complete constriction in a continuous dynamic model plosives can be simulated.

It was interesting that VocalTract could achieve these results despite having made several key abstractions from the voice production process. These include the omission of a nasal tract model and representation of the vocal tract as a completely straight path. The nasal tract is known to impact the frequency response of voice, functioning as an acoustic branch and second propagational path when coupled (section 4.6). Curvature of the vocal tract is also known to impact formant frequencies [71].

It's conceivable that VocalTract's impressive performance is due to its flexibility in creating desired results, rather than by functioning as an accurate anatomical analogue. For this reason, while it's an impressive tool, its application to the investigation of acoustic-articulatory correlation is limited as the articulators and indeed the vocal tract geometry itself are not directly instantiated. Perhaps the most important thing it does reveal is that the

path to more natural physical modelled voice synthesis indeed lies through higher dimensionality acoustic representation. The application of dynamic digital waveguide meshes also hints at the possibility of a fully-functional multi-dimensional articulation driven physical model of voice acoustics.

5.9 Other Numerical Models

The finite-difference approaches to discretisation of the wave equation introduced in section 3.1 are not the only means of modelling wave behaviour, instead representing only a subset of techniques in computational fluid dynamics simulation. Finite-element and finite-volume provide similar, if more mathematically rigorous approaches. These methodologies involves the definition of specific points (in the case of finite-element) or volumes (in the case of finite-volume) of interest in the problem domain, and the definition of numerical relationships between them. The numerical model is unconstrained in terms of complexity (and computational intensity). They can be used to represent compressible conditions, situations of net flow (effectively DC in the acoustic domain) and they can incorporate mathematical models of aerodynamic phenomena such as turbulence in efforts to completely describe the problem domain. Such fully integrated numerical models consequently introduce an extremely large computational demand, although this demand is of course a function of model complexity (and hence accuracy).

These approaches to numerical simulation have been used widely in voice research, with considerable success. They have been applied to simulation of the vocal tract transfer function [58, 87, 88, 89, 90] investigating coupling with the nasal tract [58], studying the radiation field at the lips and modelling vocal fold vibration [21, 91].

While undoubtedly accurate, such models do not necessarily present the best approach. It is quite conceivable that finite-difference based numerical acoustic models are capable of providing a correct reproduction of voice acoustics. It is additionally considered that such models are likely to present the smallest possible computational demand for natural reproduction of the vocal tract acoustics. While the capacity for explicit modelling of the detailed

aeroacoustic phenomena of voice is exciting, aside for modelling the voice it is yet to be demonstrated that reproduction of these features is *necessary* for natural voice synthesis. Mullen has demonstrated the operation of the dynamic digital waveguide mesh, for real-time continuous manipulation of vocal geometries [33]. This function is particularly attractive for the study of acoustic-articulatory correlation.

Summary

This chapter has introduced several approaches to physical modelling of the voice and the problems typically confronted. Further, key methods for acquiring geometrical data suitable for these models have been addressed. The Kelly-Lochbaum model has been defined and its extension to two-dimensional models explained.

The digital waveguide has an established history in voice modelling and while potentially more accurate frequency-domain approaches to acoustic modelling exist, the digital waveguide mesh still presents an attractive methodology for vocal tract simulation. Perhaps the most attractive facet of the digital waveguide is its capacity for time-domain operation and dynamic manipulation via impedance-mapping. Such a function is essential and indeed intuitive to the development of a dynamic voice synthesis system. Frequency-domain accuracy can be increased to an arbitrary level by increasing the mesh density, which in turn implies an increase in the computational demands of the simulation. Numerical dispersion error is seen to introduce only limited errors. The flexibility of wave-variable simulation is also appealing, allowing the use of different connective topologies and the integration of more complex or complete boundary and scattering methodologies including hybrid Kirchoff-Wave systems using the K/W pipe. The digital waveguide mesh has also shown exciting potential for the use of domain-decomposition to reduce the number of mesh points required, which might well provide a route towards real-time computation. Such simulations also have the potential for distributed parallel implementation, especially exploiting general purpose graphics processing units (GPGPUs).

The digital waveguide mesh hence offers an exciting avenue for physical modelling of the vocal tract. Chapter 6 demonstrates how these techniques have been extended in the course of this study, to produce three-dimensional acoustic models. The techniques for data collection and model development will also be introduced.

Chapter 6

3D DWM Simulation of the Voice

Introduction

Having considered existing techniques for time-domain modelling of the acoustics of the voice in chapter 5, this chapter continues to demonstrate how a three-dimensional digital waveguide mesh model of the voice can be developed.

There are clearly significant bodies of work in different areas of time-domain boundary modelling research. With the advent of K/W-Connectors, it is possible to combine these techniques as and when is appropriate. The key considerations are perhaps computational complexity and *necessity*. Scattering grids can very quickly become extremely computationally intense. While using all available technologies at all times is certainly an attractive proposition, it is likely to result in a numerical method that requires a significant amount of time to run. Whether the system is designed to work in real-time or simply reproduce an impulse response for offline convolution, excessive computational complexity has the potential to quickly render an implementation impractical. Conversely, the *ability* to reproduce all the nuances of reflective behaviour need not make it necessary. If an approximation of the reverberation time of a uniform room with completely flat walls is all that

is required, then excessive concern regarding the reproduction of diffusive behaviour, for example, is unnecessary. It is most important that the complexity of a model is appropriate to its specifications and is computationally conceivable.

The limited success together with obvious shortcomings of existing two-dimensional equivalents encourage the development of a full three-dimensional digital waveguide mesh representation of the vocal tract. While Kelly-Lochbaum modelling presents an extremely limited geometrical analogue and two-dimensional simulation is affected by ambiguities in the representation of characteristic acoustic impedance discontinuity, a three-dimensional mesh provides a more accurate geometrical solution. The two-dimensional DWM implies an extension of one-dimensional wave-digital scattering techniques, whereby formulations are adjusted to accommodate additional incoming wave components. The extension to a three-dimensional DWM is similarly straightforward, whereby update equations must account for six connections (with two on each axis). This simple change will increase computational demand, possibly beyond the conceivability of real-time operation. It is however possible to compute an impulse response representative of the vocal tract transfer function (VTTF) (or in the case of a coupled nasal tract the vocal/nasal tract transfer function (VNTTF)). The anticipated result of such an improved geometrical analogue is a more accurate representation of voice acoustics, resulting (after convolution with an appropriate source waveform) in more natural voice synthesis.

This chapter begins by addressing the challenges and issues presented by MRI capture of voice configurations, continuing to introduce the experimental and scan protocols developed. It then details the steps followed to transform the output of each scan into a graphical model and consequent numerical simulation. During this development treatment of the nasal tract introduces a number of challenges which are addressed and finally the means for development of derivative two-dimensional models are described.

6.1 Experimental Goal

For development of 3D models of the vocal tract in various articulatory guises, accurate imaging is required for each. For optimal results, the scans should minimise capture times (to reduce motion artefacts) while maximising the resulting resolution. Development and prototyping of the scan protocol is explored in section 6.2.2. The procedures followed to perform acoustic recording of the subjects are described in 6.2.1.

6.1.1 Validity of Data

An important consideration in the use of MR scans of the human head is whether the acquired imaging is truly representative of the vocal tract during normal phonation. The most obvious difference in phonation is that the subject is in a supine (horizontal) position. Several studies have addressed the difference between supine and standing phonation, typically with a focus on determining the validity of such MR imaging. Gravitation is seen to have a significant effect on articulation, identifying backward movement of the tongue and corresponding narrowing of the pharynx [92, 93] (at least in the midsagittal contour). Constriction of the pharynx may lead to reduced consistency of each sustained phonation as the air passage becomes small at the oropharynx, placing significant importance on the stability of the velum during sustained phonation to avoid significant modulation of the cross-sectional area. While the articulatory effect of supine phonation is acknowledged, the phonetic implications are thought to be minimal [93, 73] (perhaps aided by compensatory articulation). Engwall consequently investigated whether a face-down position (instead of the typical face-up orientation) would remove the articulatory artefacts of face-up supine phonation, instead finding an induced protrusion of the lips (in turn causing the cheeks to be pulled in) and forward movement of the epiglottis [92]. While it is fair to conclude that each orientation introduces undesired articulatory changes, the face-up supine position is preferred since there is more understanding of its articulatory consequences.

An additional, significant consideration in determining the adequacy of

MR imaging is the consistency and stability of successive sustained phonation. Consistency in this case refers to the ability of the subject to exactly reproduce the same articulation immediately before, during and after scanning. Stability here describes the ability of the subject to maintain each articulation exactly for the duration of the scan. The challenge of repeating phonations is significant, demanding articulatory consistency and identical source-filter interaction. It is clear that there is little reason to expect consistency in phonation [94] and also that over a longer course of time morphological changes can take place in the voice [95]. The manner of phonation is hence not 'permanent'. The likely lack of consistency is demonstrated by conditions of non-uniqueness in the articulatory-phonetic transformation, highlighting a speaker's ability to retain phonetic quality in changing voice quality. The awkward contextual influence of coarticulation is also raised. If phonetic quality can be considered consistent (where articulation can not) formant values might be considered in successive phonations as a basic measure of consistency. Short of successive MRI scans there can be little way of ensuring articulatory consistency. The relationship between geometrical manipulation and the articulatory vowel space (for early formants) is clear [96]. While the speaker might be trusted to reproduce formant frequencies with some accuracy, higher frequency resonant behaviour is very likely to change between any phonations.

Stability in scanning is crucial both in terms of maintaining phonetic quality and in preventing motion artefacts in the resulting imaging. Aalto identifies three influences on the accuracy of sustained articulation [97]:

- Gravity (Supine/Standing - as previously addressed)
- Lung volume
- Fatigue

The difficulties presented by finite lung volume are obvious. A subject cannot be expected to phonate continuously for longer than approximately 15s (although exceptions may of course be found). After this period oxygen debt may occur, leading to an increased rate of breathing and consequent

articulatory variation. The effect of periodically paused phonation is the introduction of multiple onset patterns, potentially carrying interacting formant and fundamental frequency changes as the subject settles to a target reference pitch and articulatory setting. Such changes in the articulatory setting during scanning cause motion blurring in the resulting imaging. Aalto identifies significant unintended retraction of the tongue tip in the midsagittal contour during phonation [97], which is perhaps unsurprising considering the gravitational effect of supine articulation. Engwall identified a number of other features of sustained phonation [92]. First amongst these was a reduced change in tongue contour for different vowels, combined with the gravitational effect of a narrowed pharynx. He also observed a contradicting hyper-articulation in some cases, particularly in generating a larger cavity in front of the tongue. It is interesting to consider these articulations as compensatory for pharyngeal narrowing. Perhaps most interesting, Engwall observes the use of tongue body height to counter a reduction in manipulation of the jaw. This is particularly applicable in the case of MR imaging since the head/neck coil features a bar across the base of the chin which may affect the lowest possible jaw position. While little can be done to prevent this affecting jaw height in some subjects, it is useful to be aware that the subject may use a changed tongue body height to counteract it and yield a similar acoustic output.

To avoid the introduction of multiple onset patterns during scanning scan the subjects will not be asked to repeatedly voice vowels, rather voicing the vowel for as long as reasonably possible then holding the articulation in an unvoiced condition for the remainder of the scan. It is considered that this will improve articulatory stability in each phonation.

An additional concern is whether sustained phonations are indicative of real, dynamic speech. While not approached here, there are various strategies employed to convert phonemes resynthesised using static imaging to a form suitable for dynamic voice [92].

6.1.2 Experimental Protocol

Within the scope of this study, models will be developed for a set of 11 standard vowels, 3 nasals and 5 fricatives, as in Table 6.1. Coarticulatory contexts are derived from the maritime-themed example words of [4]. While not exhaustively representative of human phonetic capabilities, they present an appropriate sampling of the vowel space and vocal capacity of the spoken English language.

Group	IPA	Co. Context	Description
Vowel	/iː/	*neap*	
	/ɪː/	*jib*	
	/ɛː/	*red*	
	/æː/	*anchor*	
	/ɑː/	*hard*	
	/ɒː/	*locker*	
	/ɔː/	*port*	
	/ʊː/	*foot*	
	/uː/	*food*	
	/ʌː/	*rudder*	
	/ɜː/	*stern*	
Nasal	/mː/	*mast*	Bilabial
	/nː/	*main*	Alveolar
	/ŋː/	*rigging*	Velar
Fricative	/θː/	*thwart*	Unvoiced Dental
	/fː/	*fog*	Unvoiced Labio-Dental
	/sː/	*sea*	Unvoiced Alveolar
	/ʃː/	*ship*	Unvoiced Palato-Alveolar
	/hː/	*heeling*	Glottal

Table 6.1: Experimental phones with coarticulatory contexts as per [4]

Reproduction of nasal phones is of course dependent on the development of accurate models of the nasal tract (and perhaps paranasal sinuses). Segmentation of the nasal tract presents inherent challenges, as explored in section 6.4. Simulation of fricatives meanwhile is dependent on turbulent aeroacoustic noise generation, which cannot be reproduced by the digital waveguide mesh. It is however anticipated that by injection of a noise source

at an appropriate position in the tract, a similar effect can be obtained.

Movement of the subject during scanning can result in motion artefacts, which are typically seen as blurring of contrast boundaries in the imaging. This effect is hard to avoid in long phonations, particularly through slight movement of the velum and epiglottis. Fig. 6.1 demonstrates regions exhibiting slight blurring caused by these movements. While largely inelective, it is hoped that highly trained singers and capable linguists are likely to demonstrate stronger vocal motor coordination and hence hold a more stationary articulatory configuration. For this reason, talented singers and linguists (in some cases both) of either gender are chosen as subjects. Participant details (and pseudonyms used throughout the study) are provided in Table 6.2.

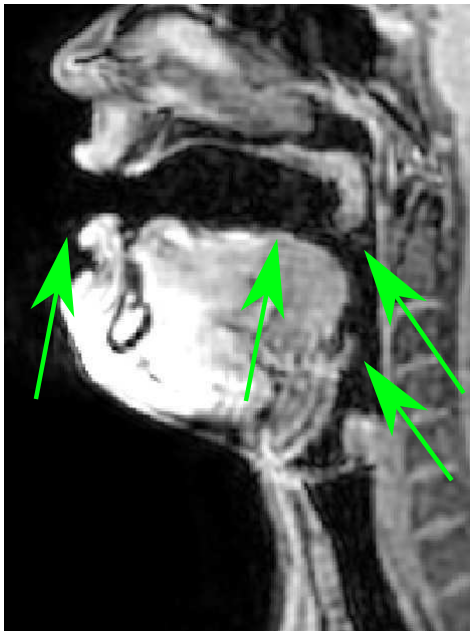


Figure 6.1: Motion artefacts (blurring) caused by inelective movement during scan period

Scans are made at the York Neuroimaging Centre (YNiC), using a General Electric 3.0T HDx Excite MRI Scanner. To use the facility an exhaustive ethical approval process is executed, including development of a suitable experimental protocol governing the processes to which participants are subjected.

Pseudonym	Details
Jack	Male, English first language, classically trained singer (tenor range), phonetically trained.
Jill	Female, German first language, phonetically trained, no singing background.
Jasmine	Female, English first language, phonetically trained, classically trained professional singer (mezzo-soprano range).
Jim	Male, English first language, classically trained singer (tenor range), no phonetic training.
Jeff	Male, French first language, classically trained singer (bass range), no phonetic training.

Table 6.2: Pseudonyms and backgrounds of subjects for MR imaging

Before scanning begins, the nature of the experiment is explained to each subject and a target pitch for spoken phonation is chosen (based on musical pitch the subject feels is closest to typical speech). Prior to the scanning session each subject is taken to a fully anechoic chamber and the following scanning process is repeated in both standing and supine orientations (section 6.2.1 to provide an audio benchmark.

After the first audio recording session the subject is taken immediately to the Neuroimaging centre. They are positioned in the scanner by the radiologist, who also provides each subject with foam earplugs and optical headphones for hearing protection and to provide an intercom facility. The chosen target pitch is played over the intercom as each scan is then executed in turn. Before each scan the subject is reminded of the target phonation and counted in by the radiologist. The subject is notified when each scan is complete.

After scanning is complete, the subject is returned to the anechoic chamber to repeat the audio recording exercise.

6.2 Method

6.2.1 Audio Capture

For objective validation and assessment of resynthesised vowels, it is important that benchmark audio is acquired corresponding directly to the vocal tract configurations captured during scanning. While it would be ideal to capture the sound produced by each subject during MR scanning, audio capture during the scan process is extremely difficult. Since the scanner bore is an extremely strong electromagnet, no ferromagnetic materials can be taken within a given radius of the device. While there are commercially available optical microphones [98], these still require potentially intrusive adaptive noise reduction techniques and are not currently available in the facility used. The design and construction of bespoke, predominantly mechanical devices for audio capture during magnetic resonance imaging has been addressed [99, 100, 101] but such devices introduce excessive demands in obtaining clearance for use in a medical facility and are more invasive than optical microphones. Even assuming a system for audio capture is available, the noise produced by the scanner during imaging is extremely high level, broadband and constantly changing. Noise cancellation and the inversion of a potentially awkward transducer configuration therefore presents an additional and significant engineering challenge [101, 100].

Instead of performing audio recording in-situ, the same phonations are instead captured immediately before and after scanning, utilising a protocol matching that of the scan procedure as closely as possible. The procedure is repeated for both standing and supine subject positions.

The subject is fitted with a headset mounted AKG CK77 omnidirectional lavalier microphone and a set of Audio-Technica ATH-M30 closed-back headphones. A recording of a typical MR scan process is played over the headphones, to mirror the occlusion of auditory feedback experienced during the scan procedure. A cue tone is added at the pitch chosen with the subject, further mirroring the experimental protocol. The same sequence of phonations is performed as during MR scanning, in the given co-articulatory context.

The subject is also asked to wear Laryngograph EGG electrodes, which will be later used as a linearly decoupled source for 3D simulation. Acoustic output (3cm from the lips) and the EGG trace are each recorded at 192kHz sampling rate using 32 bit resolution in Audacity via Firewire connection to an RME Fireface 800.

While primarily providing a benchmark for validation, the audio acquired also offers an opportunity for objective assessment of the continuity between standing and supine orientations, consistency in supine phonation and stability of sustained supine phonation. As discussed in section 6.1.1, the only anticipated metric for consistency between separated phonations is formant frequency. These are measured over the period of a scan by computation of power spectral densities for each plot. For a measurement of motor stability during sustained phonation, the tracking of formant trajectory standard deviation has been suggested [102], particularly focussing on the second formant. This is easily possible using simple linear prediction techniques, as demonstrated in section 7.5.

6.2.2 Scanning

A series of pilot scans were carried out with the intention of determining the optimal scan protocol for minimised capture time and maximal resolution. The final scan developed was a 3D fast gradient echo sequence, the details of which are given in Table 6.3. The of TE = 1.7s (echo time), TR = 4.8 (repetition time) and a 5 degree flip angle. Acquisition is isotropic 2mm in a 192x192 matrix. Output is then interpolated to 512x512 using 50% slice overlap giving an effective anisotropic output of 0.75x0.75x1mm. A stack of 80 images is produced in the midsagittal plane in approximately 16s.

During prototyping efforts were made to collect imaging of plosive configurations, however plosives are largely dynamic processes and are strongly coarticulation dependent. In the case of such phones it is considered that static imaging is not useful, however it is instead suggested that plosive imaging could be approached using dynamic MR imaging in future studies.

Static Protocol	
T_e	1.7ms
T_R	4.8ms
Flip Angle	5°
Bandwidth	±41.67Hz
FOV	$260mm^3$
Slice Width	2mm (50% separation)
Matrix	192 × 192

Table 6.3: MRI protocol developed for static vowel scanning

6.3 Model Development

The completed scans are delivered in NifTI file format, rolling all images and a header into a single file with a header, compressed using GZip. A set of DICOM (Digital Imaging and Communications in Imaging) compliant images are also provided.

The imaging delivered consists of a serious of 16-bit 512x512x80 anisotropic greyscale images (0.75x0.75x1mm resampled from 2mm isotropic imaging), whose voxel intensities describe tissue types as determined by the nature of the scan (section 4.8.1). Most medical imaging software allows the recomposition of this series of images to allow inspection on all three planes (sagittal, axial, coronal) as shown in Fig. 6.2 using the open source viewer FSLView.

Before simulation, the vocal tract model must move through a series of steps of abstraction from this initial data. The first is the development of a three-dimensional segmented graphical model of the vocal tract (and possibly nasal tract), by separating regions of interest in the imaging. In this case this means delineation of the vocal pathways from the surrounding anatomy. Segmentation is a common challenge in medical imaging, hence open source solutions are available, explored in section 6.3.1. The second step is the development of a uniform grid fitting inside the graphical model to represent a 3D rectilinear topology digital waveguide mesh, as explored in section 6.3.2. The final step of abstraction is the decomposition of this geometrically meaningful grid into a data structure well suited to efficient numerical simulation, explored in section 6.3.3.

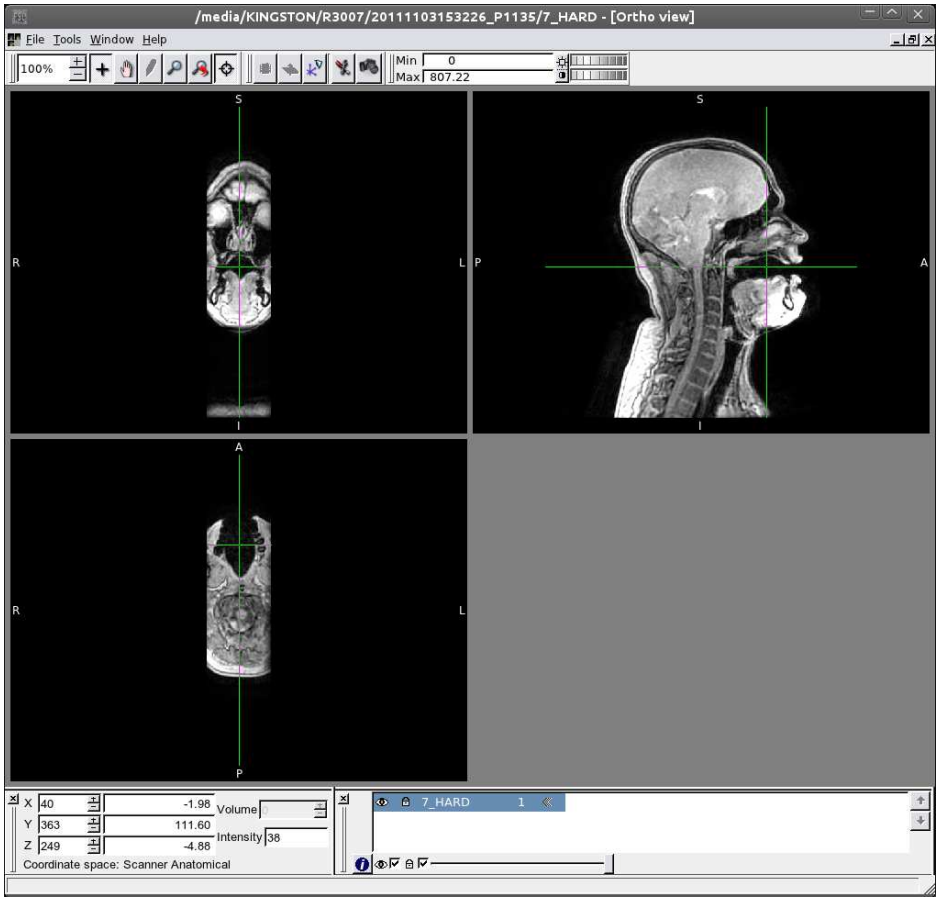


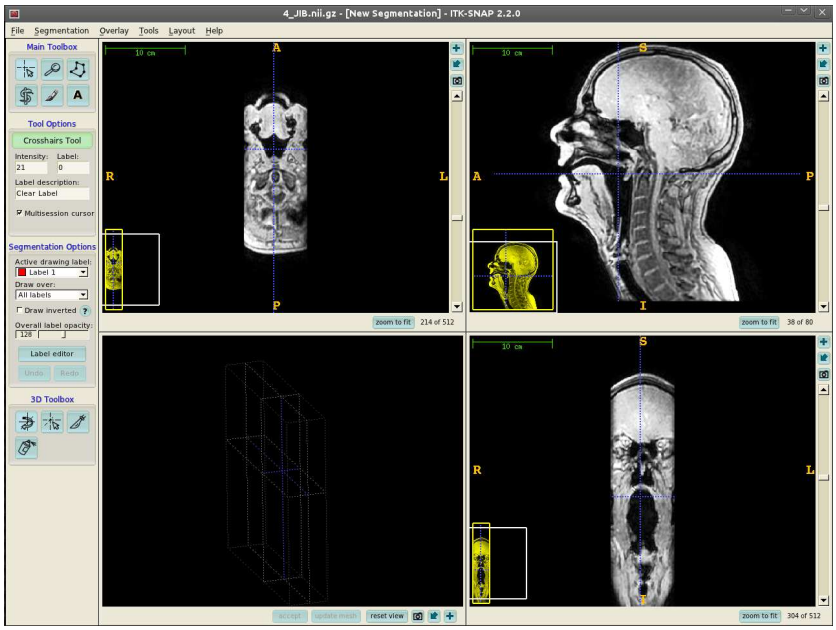
Figure 6.2: Inspecting MR imaging using FSLView, in Coronal (top left), Sagittal (top right) and Axial (bottom left) planes

Each stage in this abstraction process produces a deliverable that can be saved in a VTK (Visualisation ToolKit [103, 104, 105]) format. VTK is a freely available open-source graphical visualisation toolkit, widely used in research and medical imaging. Based on OpenGL, it can be used on all common platforms and used with several common programming languages (C++, Java, Tcl/Tk, Python). The popularity and accessibility of VTK, along with the widespread provision for its file formats in open source tools makes it a good choice for storage and manipulation of graphical data types, especially considering accessibility of the data for future studies.

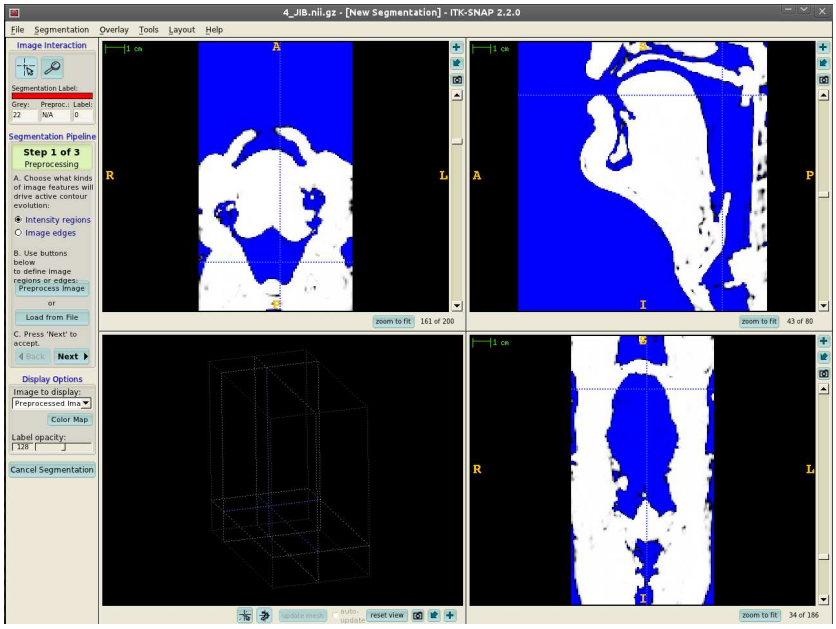
6.3.1 Segmentation

In the case of the vocal tract, segmentation describes separation of the air path from the vocal tract walls (including the lips, tongue, epiglottis, larynx). This task is made easier by the scan protocol (section 6.2.2), which is optimised for intensity variance between tissue and air. Development of a three-dimensional model from two-dimensional constituents is still non-trivial and while increased resolution in imaging permits a more detailed view of the pathway limits it also eliminates the possibility of piecewise segmentation by hand. Various approaches have been taken to automatic segmentation of the vocal tract from MR imaging, varying from simple regional intensity thresholding filters [106] to edge-detection driven spline fitting [73].

Here an automatic image-contour segmentation algorithm is used, provided as part of the freely available ITK-Snap tool for biomedical structure segmentation [107]. This algorithm is based on the concept of snake evolution, whereby a contained region expands or contracts according to sectional velocities determined by contours in regional image intensity. This process is demonstrated for the case of vocal tract segmentation in Figs. 6.3 through 6.6.

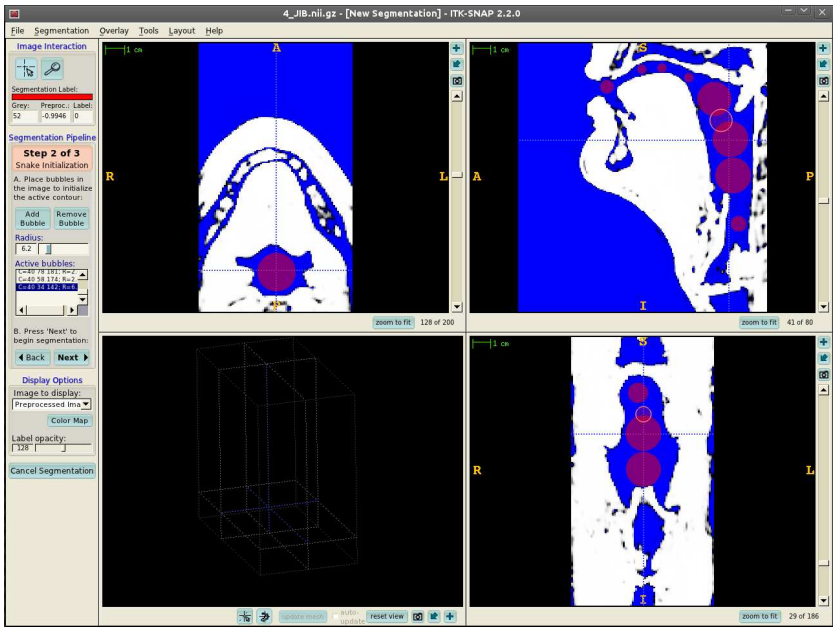


(a) Multi-Plane View in ITK-Snap

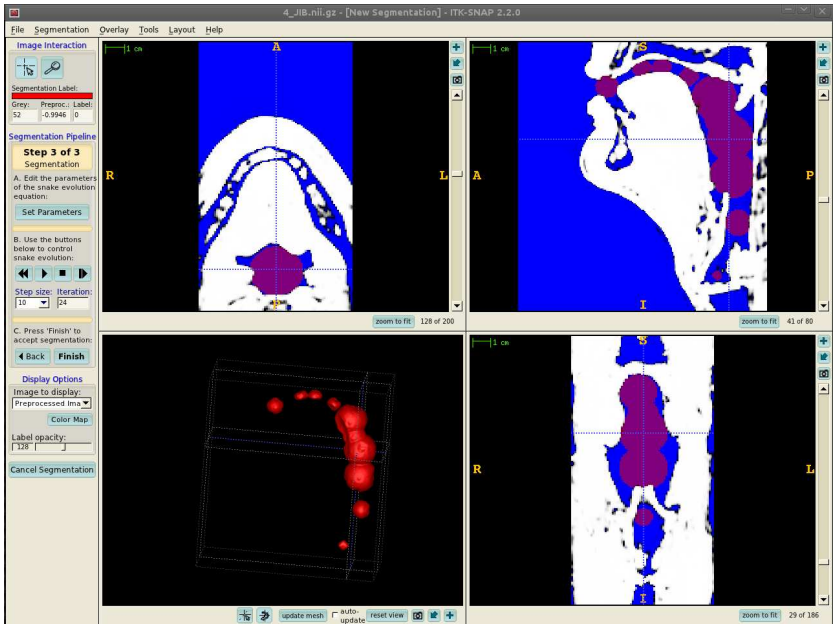


(b) Windowed regional selection

Figure 6.3: Windowing of MR imaging for adult male phonation of /ɪː/ to isolate vocal tract pathway

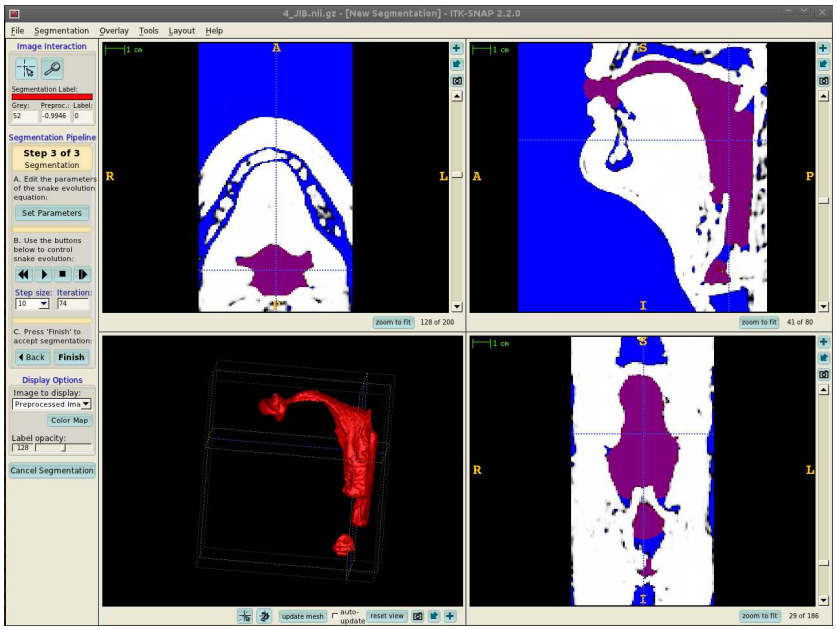


(a) Bubble positioning to initialise segmentation region

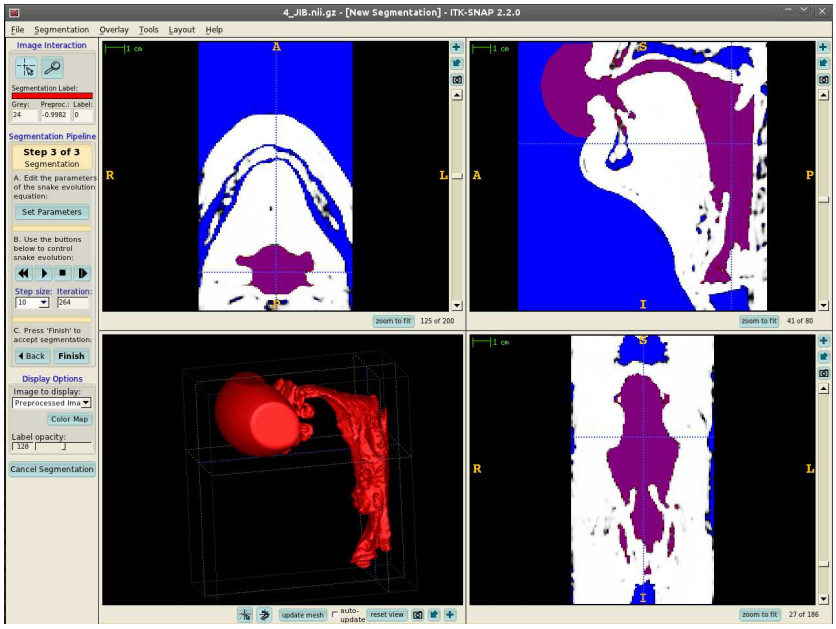


(b) Evolutionary growth of segmentation region according to intensity regions

Figure 6.4: Evolution of vocal tract segmentation for adult male phonation on /ɹː/



(a) Further evolution of segmentation region



(b) Final, uncorrected segmentation

Figure 6.5: Further evolution and completion of segmentation for adult male phonation on /ɪː/

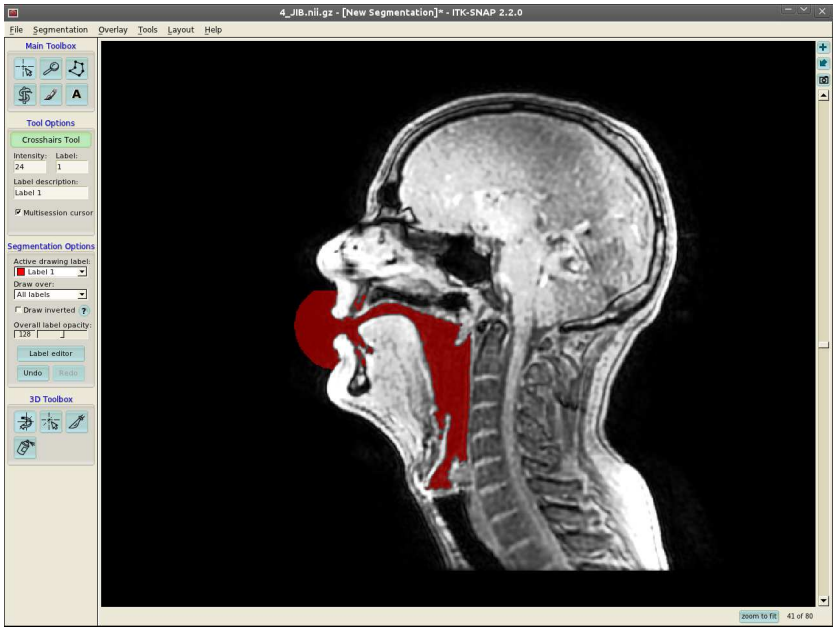


Figure 6.6: Mid-sagittal view of completed (uncorrected) segmentation

In Fig. 6.3a, the original data is loaded into ITK-Snap as a NifTI file. Fig. 6.3b shows the same imaging, for which a three-dimensional region of interested has been selected (the limits of each view) and the image intensities windowed to provide a clear delineation of free space (or indeed solid bone) and soft tissue.

To bootstrap the segmentation algorithm initial regions are required. These are created by positioning three-dimensional spheres, or *bubbles* inside the intensity region of interest, as demonstrated by Fig. 6.4a. In Fig. 6.4b the segmentation algorithm has begun, allowing the initial bubbles to expand into the intensity region but not across intensity contours. The segmented region is displayed in the lower left panel of ITK-Snap. This process continues through Fig. 6.5a to Fig. 6.5b, which shows the final condition of the segmentation.

Fig. 6.6 shows a mid-sagittal view of the completed segmentation, demonstrating successful delineation of the vocal tract pathway, but also expansion into the cavities left by the teeth. The region segmented constitutes the oral vocal pathway superior to the larynx. Other than for the explicit case of

nasal phones, there was no visible coupling of the nasal tract. Efforts were also made to segment the nasal tract and paranasal sinuses (as explored in section 6.4) and the trachea. It is not anticipated that the sub-glottal structure will be included in simulation of voiced vowels until a suitable dynamic boundary mechanism is devised for the vocal folds.

The geometry of the larynx is obviously important to the aerodynamic and consequently acoustic nature of the source. While a reasonably detailed segmentation of the larynx is possible with the current imaging, it is anticipated that with advances in laryngeal imaging technologies a more detailed model may be possible in future [76, 74, 75]. For the purposes of the current study it is considered that the level of detail attained here is adequate.

As the segmentation region expands during segmentation, it is permitted to exceed the lips and nares. This allows for inclusion of their shape in the segmented model (as shown for the lips in Fig. 6.6), consequently incorporating their radiative characteristics in simulation.

Since the teeth (and other bony structures in the facial anatomy) do not appear on MR scans, the segmentation algorithm frequently expands into regions that are not part of the pathway. This is demonstrated by Fig. 6.6, where the segmented region has grown to include the lower and upper teeth (to the right hand side of the lips in the image). It also automatically moves into the trachea since the vocal folds demonstrate an averaged position (as shown in Fig. 6.7). These errors are corrected by hand using an adaptive paintbrush after automatic segmentation (which can be particularly time-consuming).

After segmentation, the region is saved as a surface mesh (in VTK PolyData format) and the segmentation itself is saved as an Analyze format image. Saving the segmentation separately allows for later modifications such as the inclusion of a standard model of the nasal tract and paranasal sinuses, or addition of detailed dental models. After segmentation a complete three-dimensional model of the vocal (and possibly nasal) tract is available, suitable for inspection and development of various types of numerical model. Such a model is shown in Fig. 6.8, which also shows the same model with the lip radiation dome cut away to reveal the lips. The next step is to develop a

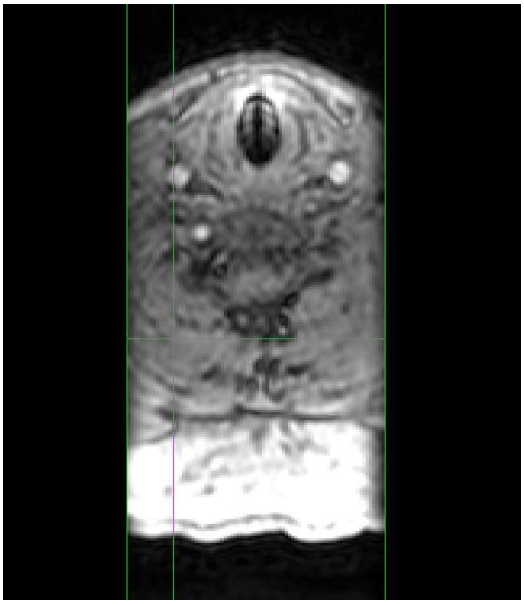


Figure 6.7: Coronal section of adult male during sustained phonation demonstrating average vocal fold positions with significant motion artefacts

isotropic three-dimensional grid suitable for DWM-style computation.

6.3.2 Development of Sampling Grids

The development of sampling grids was performed using image stencilling functions provided as part of the VTK toolkit. The polydata representation of the vocal tract is first transformed into an image stencil. This stencil is applied to a regular three-dimensional grid (in this case a three-dimensional vtkImageData) whose isotropic voxel dimensions correspond to the spatial sampling interval presented by the desired simulation sampling rate. Grid elements that fall inside the stencil geometry are assigned the scalar value 1.0, whilst elements outside the geometry are assigned 0.0. The shape of the geometry has hence been *stencilled* into the regular grid. This grid is then traversed, generating a graphical model of lines and points describing the construction and connectivity of the sampling grid, as shown in Fig. 6.9. This final graphical model is saved as a vtkUnstructuredGrid. Representing the sampling grid in this format is advantageous as it explicitly defines the

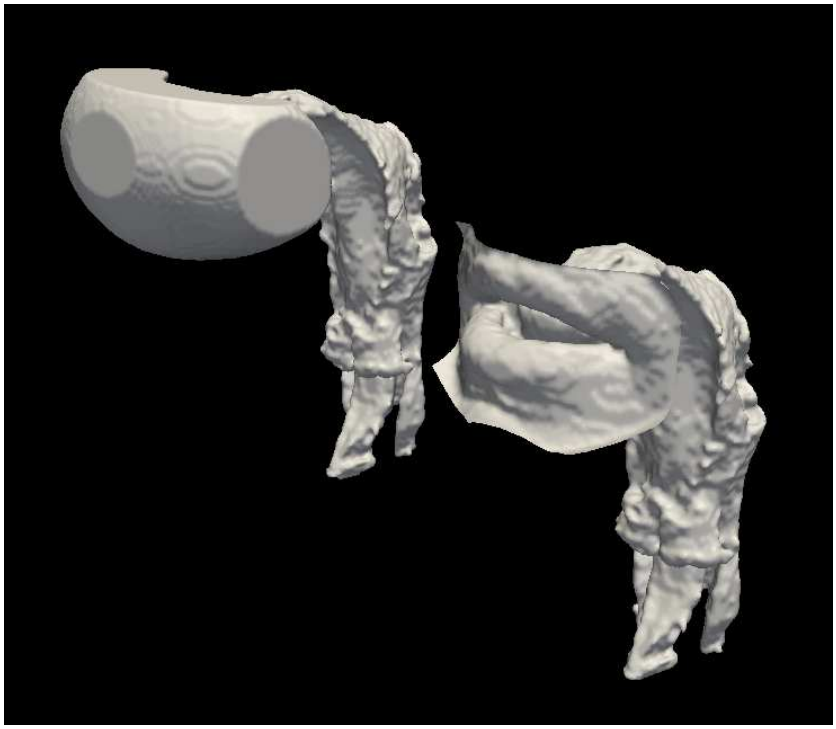


Figure 6.8: Three-dimensional view of graphical model of the segmented vocal tract for adult male phonation of /ɹː/ with and without lip radiation dome (left and right respectively)

cartesian geometric position of each sampling point in the network. The corresponding VTK class also features a number of member functions which allow neighbouring points to be determined. Since the sampling grids are intrinsically linked to a given spatial sampling interval, they are specific to a given system sampling rate and must be recomputed should this change. Fig. 6.9 shows the original PolyData model, together with grids of sampling rate 96kHz, 192kHz, 384kHz and 768kHz. Note that the 96kHz grid is not completely connected, and the 192kHz grid is very sparsely sampled.

Whilst it may be desirable to maintain the direct geometrical representation of the sampling grid during simulation, an efficient implementation suggests a more abstracted approach. The next step is hence the decomposition of this geometrically analogous model to a structure suitable for simulation.

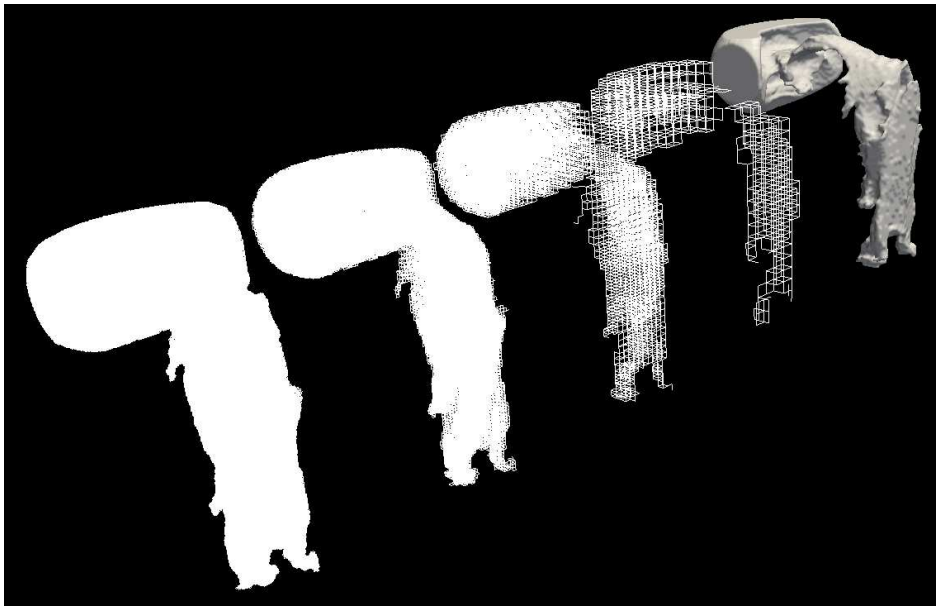


Figure 6.9: Original PolyData for adult male phonation of /ɪː/ (far right), with consequent sampling grids corresponding to system sampling rates of 768kHz (far left), 384kHz, 192kHz and 96kHz

6.3.3 Implementation

The major requirements of the implementation within the scope of this project are:

- Reasonable execution time for computation of an impulse response
- Flexibility - Reusability, adaptability, compilation on different architectures.

A reasonable execution time is a fundamental requirement. At this stage flexibility is more desirable than absolute optimisation. While the use of processor-specific intrinsics would accelerate execution it would hamper portability and within the scope of the project, accessible and highly manipulable programming is preferred. Flexibility also entails the ability to change methodologies as desired; changing boundary update equations for example should not require a rewrite. To this end a strongly object oriented approach is pursued.

For simulation of a wave-variable based scattering network, a node-centric approach is preferred (as explored briefly in section 3.11). In this case, each scattering point is an object, aware of its neighbours and responsible for its own update and shuffle operations. While this places a large demand on memory (as explored in section 3.11), it should not be computationally infeasible, and the approach is well suited to an object oriented methodology.

The low level arrangement of the data structure is shown in Fig. 6.10, and the complete collaboration diagram in Appendix B.

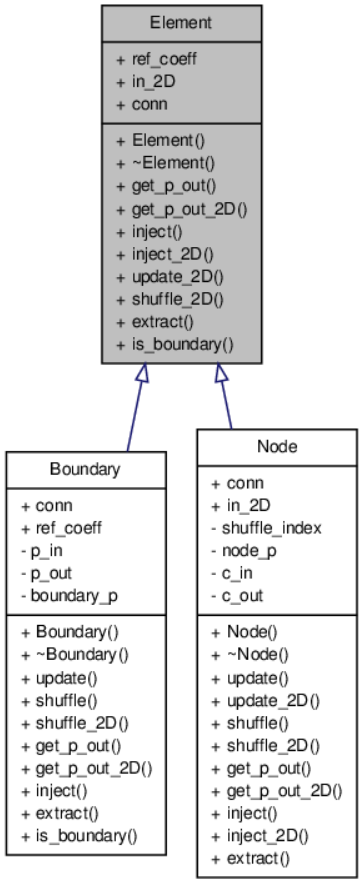


Figure 6.10: UML diagram of the data structure

The primary element is a virtual base class - the Element. This base class represents each point in a given sampling network, establishing the functions that any subclass must be expected to perform and its member

variables (particularly an array of pointers to neighbouring elements and variables to contain wave quantities). This base class is inherited by Node and Boundary, the former representing scattering nodes in a homogeneous medium (in this case air) and the latter representing boundary nodes. Each adopts functions suitable to its type. These elements are contained within the class Model, which holds arrays of pointers to all nodes, boundaries, source nodes and receiver nodes. The Model class is responsible for generating the data structure from any ingested sampling network (the model built in section 6.3.2), driving update of all contained elements and also for injection / extraction from the data structure.

There are three steps to generating the data structure from the graphical model of the sampling grid. The first is to generate Element objects for all points in the structure. The second is to traverse all points in the ingested sampling grid, establishing pointers to neighbours according to a transform table maintained in the Model class and casting all appropriate Elements to Nodes. A pointer to each Node is added to the Node array in Model. The final step is to traverse the Node array, generating Elements casted to the Boundary class and assigning them to any 'loose' connections in the data structure. A pointer to each Boundary is added to the Boundary array maintained in the Model class. Once the data structure is constructed, a complete traversal is performed to check for leaks (null connections) and physically impossible connections.

C++ is used to implement the system, for ease of integration with VTK and efficiency. Compilation is by gcc on a Linux platform, however there are no constraints to compilation on Windows/Macintosh systems other than VTK installation. The C++ standard library is avoided for the main data structure for efficiency, although vectors are occasionally used in the higher level interface to maintain lists of less significant objects. Compiler intrinsics are not yet implemented for transparency, although it's considered these will considerably improve performance. Double precision (64 bit) floating point variables are used to contain wave variables to provide a large dynamic range.

6.3.4 Visualisation

Visualisation is an extremely useful tool to ensure behaviour during simulation is acceptable. It can also be extremely informative as to the nature of the radiating behaviour, as time-domain simulation lends itself directly to wavefront visualisation. Since all data types used in the development of the simulation (with the exception of the final data structure) are based on VTK primitives, it is also remarkably straightforward.

Since visualisation in three-dimensions is complicated (due to the lack of a suitable viewing position), a view of user-selectable two-dimensional planes is preferred. To achieve this, the user selects any number of points from the original graphical sampling network and saves them as a separate vtkUnstructuredGrid. The higher-level user interface builds a new graphical grid structure based on these points and constructs a lookup table to map its constituent points onto the corresponding entries in the simulation data structure. This grid is displayed onscreen during simulation. The colour of each point on the grid is determined by the current pressure of its corresponding element in the data structure, mapped through a blue-red colour map.

An example of the resulting visualisation is shown in Fig. 6.11. Any number of different planes can be added to a single visualisation to inspect different aspects of the simulation. It should be remembered that visualisation significantly slows down simulation, so simulation is kept entirely decoupled from visualisation in the program structure and can be easily disabled.

6.3.5 Boundary Formulations

The boundary formulation currently used in this project is the simple one-connection wave variable reflection boundary derived in section 3.6. Reflective coefficients are set and stored in the Boundary class. While constituting a significant approximation with regard to anticipated reflective behaviours in the system, these boundaries demonstrate an acceptable performance for more simple geometries used in validation (in Chapter 7). In the vocal tract, the boundaries are expected to perform poorly in two major ways. The

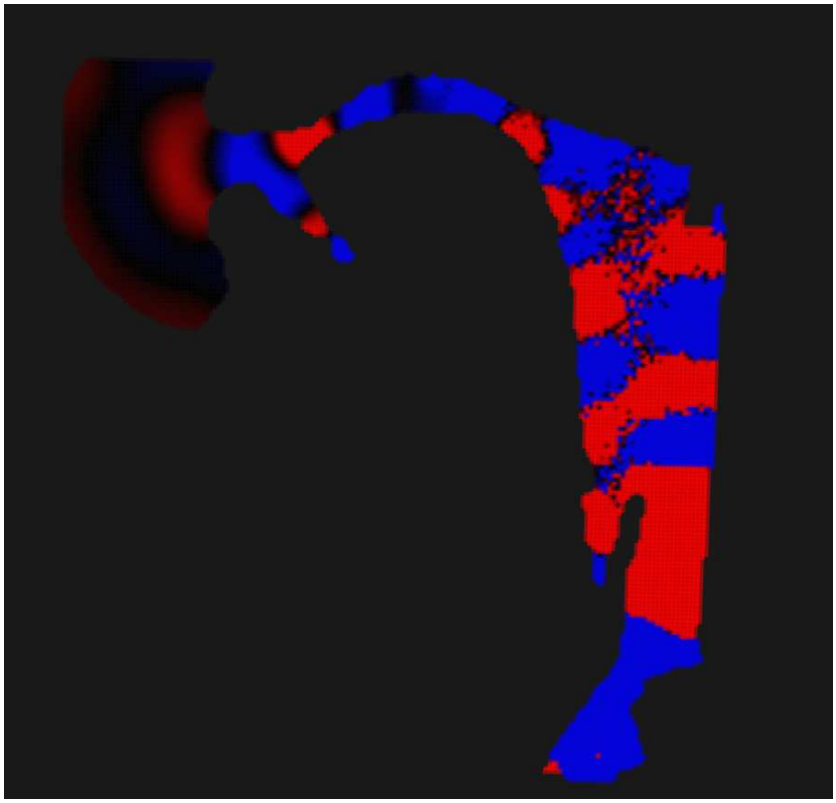


Figure 6.11: Visualisation of 768kHz simulation of adult male phonation on /ɹː/, using single mid-sagittal slice inspection

first is the frequency independence of the one-connection boundary, based on an assumed pure-real discontinuity in acoustic impedance. The reflective behaviour of the soft tissues of the vocal tract (and indeed nasal tract) are likely to exhibit frequency dependent behaviour which will not be reproduced. In the simulations performed here all vocal tract limits (with the exception of the exterior of the lip dome) will also be assigned a uniform reflective coefficient. This is unlikely to be the case in the real vocal tract, as the hard palate for example would be expected to exhibit a different reflective characteristic to that of the tongue. If desired, more accurate anatomically-specific reflective behaviours (and hence boundary formulations) can be introduced at at later date by inheritance of the Boundary class. This might allow for introduction of frequency-dependent wave-digital filters for example.

The second major shortcoming of the formulation is in its absolute rigidity.

Much of the vocal tract consists of soft tissues which might be expected to vibrate in mechanical sympathy with airflow or acoustic resonance, serving to modify the effective damping. While the acoustic manifestation of this effect could be represented by the modulation of reflective coefficients, a means of directly simulating the effect is not apparent.

6.3.6 Injection

Injection to the mesh will be by means of a point-oriented soft source mechanism, adding a wave variable signal to the existing field at a single node. The point at which injection is made is determined by selection of a single point in the graphical model of the sampling network, loaded into the program as an additional vtkUnstructuredGrid by the higher level user interface. This finds the point in the simulation data structure and adds it to a vector of injection nodes. A PCM .wav or data file containing the signal for injection is simultaneously loaded, whose contents are linked to a single entry in the injection vector. By this means it is possible to inject an unlimited number of signals to any point in the simulation simultaneously.

Initially, injection will be on a point-source basis. While the voice source is not a point source (perhaps better matched to a radiating line of sources), development of such a mechanism is beyond the immediate scope of the thesis. Validation tests demonstrate that this basic mechanism is suitable for determination of acoustic impulse responses (Chapter 7). The developed system could readily be used as a testbed to investigate such an approach. The point source in this case is positioned at the approximate centre point of the vocal folds.

The injection signal here is a bilateral sinc function, whose zero crossings correspond to a cut-off frequency of $20kHz$. This is used to prevent system excitation at excessively high frequencies, removing the potential for aliasing error. The support of the sinc function will be determined as a function of the simulation run-time. The signal will be recomputed for different system sampling rates.

6.3.7 Extraction

Extraction is executed similarly to the injection mechanism. A point is selected, saved as a vtkUnstructuredData type and ingested through the higher level interface. Any number of points can be added in a single file, pointers to which are added to a vector of extraction points. During simulation, wave variables are extracted at each relevant node and saved to individual vectors for each output point. At the conclusion of each simulation these are saved to individual data files, with names corresponding to the extraction point.

For the case of vocal simulations, an array of receivers is positioned at the centre of the lips, slightly outside the maximum lip protrusion. Extracting from a field of points rather than a single point allows for a wider inspection of a pressure field, especially useful where the field is likely to be spatially sensitive (such as in the case of lip radiation).

6.4 The Nasal Tract

Compared to segmentation of the vocal tract, segmentation of the nasal tract constitutes a significant challenge. As explored in section 4.6, the nasal tract is fundamentally different in nature to the vocal tract. Whereas the latter consists of a generally open, empty pathway the former is a labryinthine structure of more narrow pathways. Fig. 6.12 shows the nasal tract in axial and coronal projections. The conchae, particularly visible on Fig. 6.12b, are curved shelves of bone [43] projecting into the body of the nasal tract. Tracing the spaces around the conchae is reasonably straightforward, however it is not always completely clear where the pathways end and the surrounding bone and cartilage (which are not clear on scans) begins. The nature of this bone also complicates connection of the paranasal sinuses. These are considered to impart a significant effect on the vocal process [57, 59, 60], even if it is not yet entirely understood. At their most simple the paranasal sinuses can be seen to function as Helmholtz resonators, the effect of which is impacted by their point of coupling to the nasal tract - the ostia. The positions of these ostia are extremely hard to determine from scans, since

the surrounding wall is particularly thin (as in Fig. 6.12a).

(a) Axial (b) Coronal

Figure 6.12: Axial and coronal plane views of structural nasal tract scan (no phonation)

Since the nasal tract is a predominantly static structure, a different scan protocol can be used to collect higher resolution imaging with a significantly longer capture time. The head coil can also be used, allowing a more targetted image than is the case for the head and neck coil used for vocal tract capture. In this case the preferred structural scan results in a stack of 116x512x512 images of non-isotropic 1.00x0.75x0.75mm voxel spacing.

Segmentation of the nasal tract is approached using a combination of the automatic snake evolution algorithm used for vocal tract segmentation, and manual segmentation (using ITK-Snap's adaptive paintbrush). This can be extremely time-consuming, but need only be carried out once per subject. After segmentation the model can be reloaded in ITK-Snap and combined with a segmented vocal tract model.

6.5 2D Derivative Simulation

The same function that can be used to extract slices for visualisation (section 6.3.4) can be used to develop two-dimensional simulations from a three-dimensional sampling grid. Each class used in simulation has two-dimensional equivalents of each of its member functions. Where the Model class would normally be used to construct a data structure based on three-dimensional simulation, it uses a separate method to build an equivalent structure for two-dimensional simulation of a given slice. This is extremely useful to check the comparative performance of two- and three-dimensional simulations. It is only important that the sampling grid developed uses a spatial sampling interval that is consistent with two dimensional simulation (according to equation 3.22 of section 3.2).

It should however be remembered that two-dimensional decomposition of a three-dimensional graphical model is not expected to be correct. Existing two-dimensional models make a number of compromises to ensure correct reproduction of axial or tangential resonant behaviour. In this case were a midsagittal slice to be chosen (as in Fig. 6.13), the cross-tract distances are not likely to be indicative of changing cross-sectional area, hence the models represent a significant (and potentially physically inaccurate) abstraction.

Summary

In this chapter techniques for development of a three-dimensional DWM model of the voice have been described. The complications inherent to acquiring appropriate data are explored, along with a description of the resulting experimental protocol. While the development of geometrically accurate sampling grids suggests an appropriate analogue for numerical simulation, this is by no means an assurance of consequent acoustic correctness. For the results of numerical simulation to be useful, a rigourous validation of the techniques used must first be performed. This validation is approached in Chapter 7, for models of progressively increasing complexity.

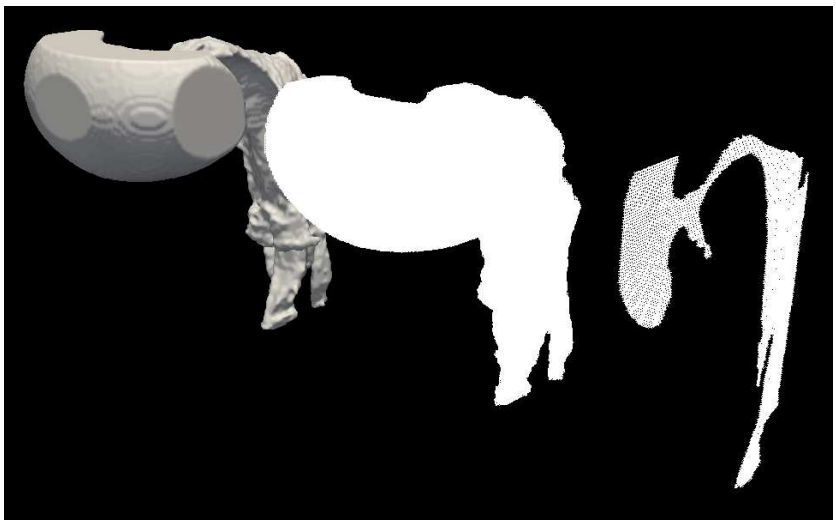


Figure 6.13: Extraction of a two-dimensional derivative mid-sagittal sampling grid (right) from a complete three-dimensional model

Chapter 7

Benchmarking and Validation

Introduction

Chapter 6 introduced the techniques employed in this project for numerical simulation of vocal-tract-like structures. While mathematically rigourous, no numerical model can realistically reproduce the true complexity of acoustic behaviour. To this end, a number of approximations and simplifications are made throughout the simulation process. Validation of the numerical simulation technique is hence fundamental to the value of the results. It must be possible to demonstrate the trustworthiness, accuracy and potential error inherent to such an approach. This chapter encompasses efforts made towards validation and benchmarking of the simulation technique. A series of structures are introduced, all analogous to the vocal tract, but of increasing geometrical complexity. A combination of direct mathematical determination and acoustic measurement are then used to assess the validity of corresponding simulations. Section 7.1 begins by considering simulation of a simple enclosed cuboid, using direct mathematical determination of anticipated resonant behaviours. Section 7.2 then continues to introduce an acoustic measurement technique used to validate more complex geometries. This measurement technique is itself assessed in section 7.3, along with the results of simulation of simple cylindrical structures. It continues to address validation of concatenated cylindrical arrangements. In section 7.4, the sim-

ulation of complex cylindrical analogues to the vocal tract are considered, using acoustic measurement alone as a benchmark. Finally, section 7.5 considers how recorded audio can be used to benchmark the results of simulation of full vocal tract models.

7.1 Cuboidal

The most fundamental validation exercise performed is simulation of the acoustic response of a simple cuboid of dimensions 3cm x 4cm x 17cm. Such a cuboid represents the most straightforward means of validation as its resonant frequencies can be analytically determined. The simulation implementation is kept entirely as intended for vocal tract simulatiovn, with the same source, excitation and extraction methodologies as described in chapter 6. The system sampling rate used is 960kHz and a sinc function designed to provide a cutoff frequency of 20kHz is injected. The cuboid graphical model is shown in Fig. 7.1.

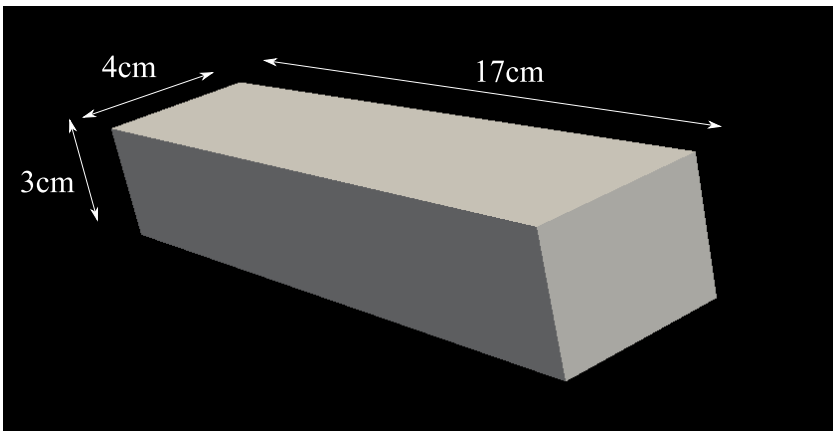


Figure 7.1: Cuboid Geometry

7.1.1 Lumped Measurement

As explored in sections 2.3 and 2.4, the frequencies of acoustic resonance in a closed one-dimensional system are easily determinable and represented by

(7.1) where n is mode number, c the speed of sound ($340.5ms^{-1}$ here) and L is the distance between boundaries.

$$f_R = \frac{cN}{2L} \tag{7.1}$$

A cuboid can be approximated as an interacting amalgamation of three such systems. The consequent resonant frequencies can hence be predicted by the universal modal equation (7.2), taking each dimension into account where $N_{x,y,z}$ is the Nth mode in each axis and l,w,h the corresponding distances.

$$f_{xyz} = \frac{c}{2}\sqrt{\left(\frac{N_x}{l}\right)^2 + \left(\frac{N_y}{w}\right)^2 + \left(\frac{N_z}{h}\right)^2} \tag{7.2}$$

Using (7.2) for the cuboid of Fig. 7.1 it is possible to predict resonant mode frequencies as given in Table 7.1, for a speed of sound $c = 340.5$ where N_z describes modes along the length (0.17m), N_y modes across the width (0.04m) and N_x modes across the height (0.03m).

In calculating these frequencies the system acoustics are considered to be linear, as would predominantly be the case for such a closed inert structure. The presence of the resonant modes is largely dependent on receiver position during simulation (positioning a receiver on a pressure node would result in a nullified output at that frequency). For maximum visibility of the resonant modes a chain of receivers is selected running diagonally across the sampling network, as shown in Fig. 7.2. For excitation of the maximal frequency range of resonant modes (but reduced energy at lower frequencies) injection is made in a corner of the cuboid, as can also be seen on Fig. 7.2. Multiple-source injection is avoided to reduce the risk of possible interference artefacting.

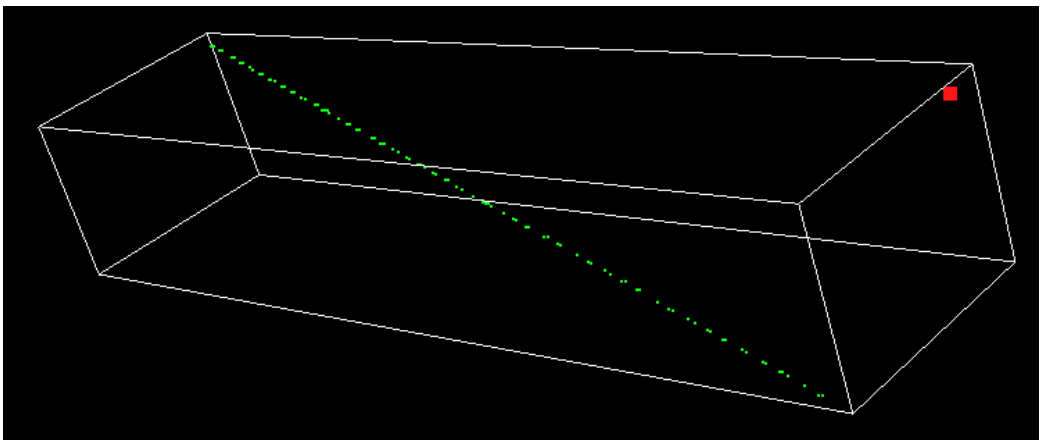


Figure 7.2: Source point (red) and receiver array (green) positions shown inside the cuboid wireframe

N_x	N_y	N_z	f_{xyz} (Hz)	f_{min} (Hz)	e (Hz)	M (Hz)	e_M (Hz)
0	0	1	1001	964	37	999	2
0	0	2	2003	1959	44	1989	14
0	0	3	3004	2953	52	2981	24
0	0	4	4006	3947	59	3974	32
0	0	5	5007	4941	66	4992	15
0	0	6	6009	5936	73	5963	46
0	0	7	7010	6930	80	6959	51
0	0	8	8012	7924	87	7999	13
0	0	9	9013	8919	95	8975	38
0	1	0	4256	4099	157	4219	37
0	2	0	8513	8229	284	-	-
1	0	0	5675	5422	253	-	-
2	0	0	11350	10873	477	-	-
0	1	1	4372	4217	155	4358	14
1	1	1	7164	6881	283	-	-
1	2	1	10280	9916	364	-	-
0	1	2	4704	4553	151	4661	43
0	1	3	5210	5064	146	5165	45
0	1	4	5845	5703	142	5770	75
0	1	5	6572	6433	139	6410	162
0	1	6	7364	7225	138	7283	81
0	1	7	8201	8063	138	-	-
0	1	8	9072	8932	140	-	-
0	1	9	9968	9825	142	-	-
1	0	1	5763	5512	251	-	-
0	2	1	8571	8288	283	-	-
1	0	2	6018	5773	245	5770	248
1	0	3	6421	6184	237	6366	55
1	0	4	6946	6718	228	-	-
1	0	5	7568	7348	220	7513	55
1	0	6	8265	8052	214	-	-
1	0	7	9019	8811	208	-	-
1	0	8	9818	9613	205	-	-

Table 7.1: Resonant modes for the cuboid of Fig. 7.1 where N_x, N_y, N_z are orders of modes in each axis, f_{xyz} is the mathematically approximated modal frequency, f_{min} is the modal frequency approximation under maximum error conditions, e is the magnitude of this error, M the absolute measured modal frequency and e_M the difference between measurement and analytical approximation of the mode.

7.1.2 Simulation

Simulation of the cuboid at 960kHz using a 3D rectilinear DWM results in 882245 air elements, 69526 one-connection boundaries (r=0.96), 92 receivers and a single source point. The spatial sampling interview of 0.614mm provides an arrangement of of approximately 48 x 65 x 276 nodes. Simulation ran for 8000 steps with the additive sinc injective supported to 801 samples, yielding a run time of approximately 30 minutes. The consequent impulse responses were analysed using a 16384-tap FFT after flat-top windowing. This yields a bin-width of approximately 59Hz, giving a potential error of ±30Hz with negligible scalloping loss [108]. The magnitude response for all receivers is plotted in Fig. 7.3.

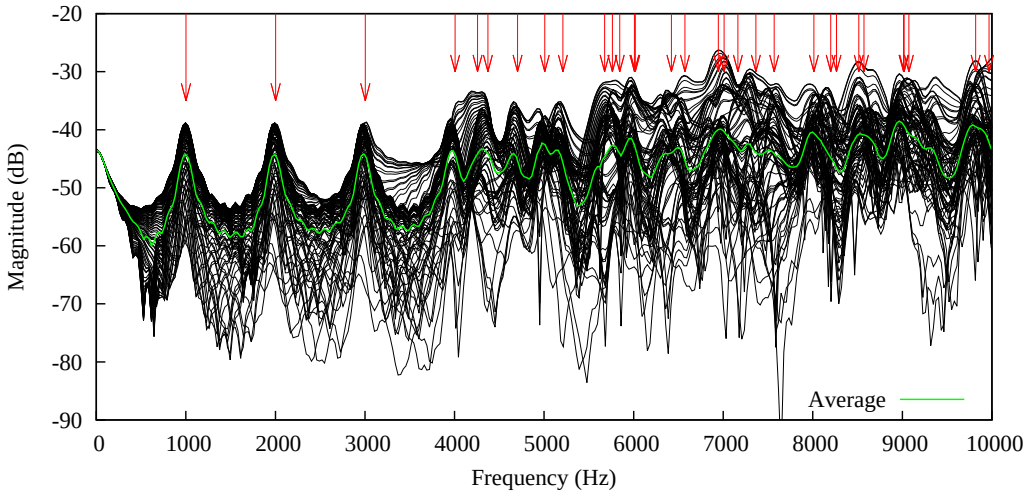


Figure 7.3: Magnitude responses of system at 92 diagonal receiver points shown in Fig. 7.2, with averaged frequency response (green) and calculated mode positions as per Table 7.1

Note that the variation in responses above 4-5kHz is significant, resulting in a dense layering of lines that is particularly difficult to interpret. To address this, the average magnitude response is calculated and plotted in green. This plot has no grounding in the acoustic nature of the system, but provides a suitable approximation of the system response. The resonant modes of the system were measured as the centre frequencies of the bins

representing regional maxima in the average magnitude response of Fig. 7.3 and are given in Table 7.1 in column M (in Hz).

Since the spatial sampling interval is fixed, a degree of spatial sampling error will occur, where the grid does not fit perfectly inside the containing geometry. In a fixed rectilinear geometry such as the cuboid the effect of this error is easy to predict. The nature of the grid construction means that spatial sampling error always extends the effective distance between boundaries. This puts the error in the region $0 \leq e_{spat} \leq 2d$, where d is the spatial sampling interval and e_{spat} an acoustic path difference. In the case of this simulation the maximum possible spatial sampling error between two boundaries is equivalent to a path extension of 1.23mm and the maximum frequency percentage error is given by (7.3) for each dimension, where L describes the actual pathlength.

$$\frac{e_{spat}}{L + e_{spat}} \tag{7.3}$$

For the spatial sampling interval in this system the maximum possible spatial sampling error corresponds to a frequency percentage error of 0.72% for the 0.17m dimension, 2.98% for the 0.04m dimension and 3.93% for the 0.03m dimension. In the case of a perfect fit, spatial sampling error will not be introduced. The maximum possible error limit is then extended by the FFT half bin-width of ±30Hz. These error conditions are included in a maximal error calculation for each resonant mode in Table 7.1, in column f_{min} (in Hz). The maximum possible difference between the calculated value and the simulated value where spatial sampling error and the FFT bin width is taken into account is quoted in column e (in Hz), and the actual difference after measurement from the average magnitude response is given in column e_M (in Hz). Where the resonant mode frequency was not clear (in the case of degenerate modes for example) a measurement was not made.

A further source of error in the simulation is numerical dispersion error, as introduced in section 3.10. The dispersion factor (k) for a 3D rectilinear grid is given by (7.4) as a function of spatial frequencies ξ_x, ξ_y and ξ_z. The topology-specific function b_{3r} is given in (7.5), where ξ represents the spatial

frequency normal and D describes the optimum spatial sampling interval (the distance covered by a wave travelling at ideal speed in an idealised medium during a single temporal sampling interval) [37].

$$k(\xi_x, \xi_y, \xi_z) = \frac{\sqrt{3}}{2\pi\xi} arctan \frac{\sqrt{4 - b_{3r}^2}}{b_{3r}} \tag{7.4}$$

$$b_{3r}(\xi_x, \xi_y, \xi_z) = \frac{2}{3} \left(cos(2\pi D\xi_x) + cos(2\pi D\xi_y) + cos(2\pi D\xi_z) \right) \tag{7.5}$$

It can be observed (particularly from Fig. 3.15 in section 3.10) that lower spatial frequencies exhibit lower numerical dispersion error. In this $960kHz$ system frequencies no higher than $10kHz$ are examined, hence spatial frequencies of $|\xi| > 0.01042$ are not exceeded. Within this range the normalised dispersion factor is never more than 0.99988, causing a maximum disparity in phase velocity across all directions and frequencies of interest of $\pm 0.00012c$. For the 9th lengthwise resonant mode given in Table 7.1 (9013 Hz) this would correspond to a maximum error of approximately 1Hz, which is insignificant in comparison to the spatial sampling error.

Table 7.1 demonstrates several important points. Most importantly, the three-dimensional digital waveguide mesh can be seen to reliably reproduce the resonant behaviour of a closed cuboid up to at least 10kHz. This is despite using minimal boundary formulations. It has been demonstrated that the errors present are largely due to spatial sampling error, the scale of which can be significant. Considering the resonant modes of the cuboid in Table 7.1 spatial sampling errors of over 200Hz could be experienced as low as 5.5kHz (for the case of the first resonant mode across the cuboid's smallest dimension). The significance of spatial sampling error increases as cross-sectional distances shrink, the only means of combating which is an increase in sampling frequency throughout the model. Table 7.1 also demonstrates that spatial sampling error only leads to a reduction in realised frequencies of resonance, since it is caused by an increase of the size of the sampling grid beyond the original geometry. Finally, it is also clear that the absolute

frequency shifts induced by spatial sampling error increase with frequency since the error factor is applied to the effective spatial period.

7.2 Acoustic Measurement

Beyond simple structures like the cuboid, direct mathematical determination of the acoustic response of a geometry becomes non-trivial and particularly approximate. To allow proper validation of the numerical simulation of such structures a means of measuring the actual response must be developed. Inspiration for such an approach has come from a method for the harmonic-independent determination of vocal formants during singing [13, 109]. In this case a broadband noise source is presented at the lips, coupled by an impedance matching horn to a cowl and attached microphone as per Fig. 7.4. The noise source is then calibrated to provide a flat spectrum for a reference acoustic load (of the closed mouth). The effective input impedance of the vocal tract was then approximated by considering the results to represent the vocal tract coupled in parallel with the reference impedance.

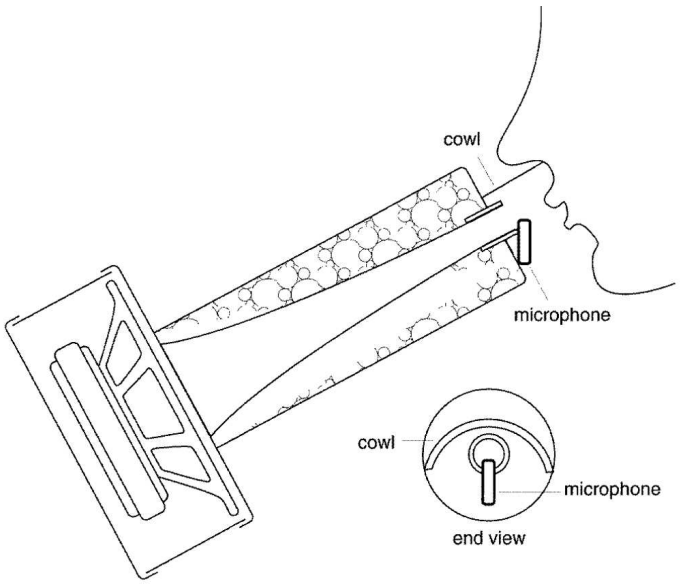


Figure 7.4: System diagram for noise-based determination of vocal tract input impedance - From [13]

Since measurement of vocal tract analogues does not demand in-vivo measurement in the manner of voice assessment, it is possible to position the acoustic source at the effective voice source. In the case of each of the geometries under test here, this means the reference load is in series with the system under test and can hence be directly coupled, as demonstrated in Fig. 7.5. The reference load in this case is provided by an acoustically long (5m) tube of cross-sectional area matching the input to the test geometry. This provides an approximation of the loading characteristic acoustic impedance without imparting significant resonant characteristics of its own.

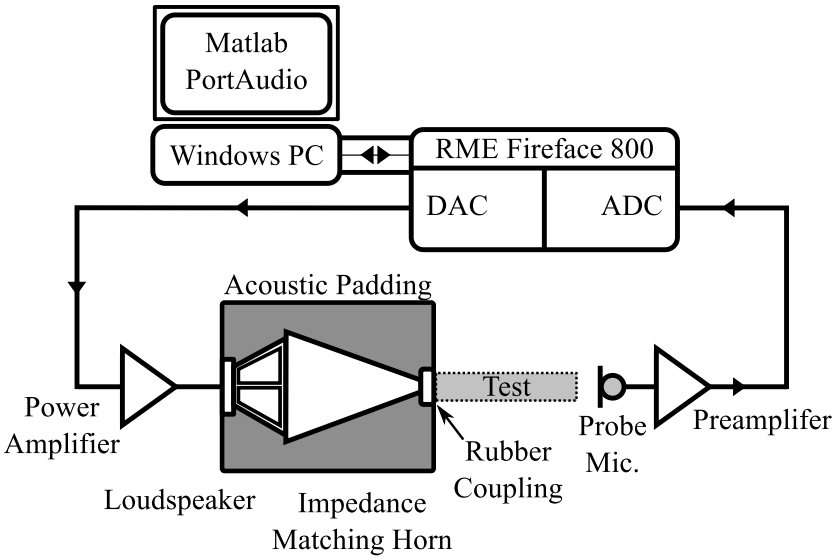


Figure 7.5: System diagram for exponential sine-sweep measurement of mechanical vocal tract analogues

7.2.1 Exponential Sine Sweep Measurement

Rather than a broadband noise source, excitation of the system is by means of exponential sine-sweep [110, 111]. This has been shown to be appropriate for determination of the acoustic impulse responses of room acoustics, offering a manipulable (and largely transducer independent) signal-noise ratio and direct separation of rising orders of harmonic distortion. The expression for a generic exponential sweep is given in (7.6), where ω_1 and ω_2 are angular

start and stop frequencies respectively and T the sweep duration.

$$x(t) = sin\left[\frac{\omega_1 T}{ln\left(\frac{\omega_2}{\omega_1}\right)}\left(e^{\frac{t}{T}ln\left(\frac{\omega_2}{\omega_1}\right)} - 1\right)\right] \tag{7.6}$$

By performing a full range exponential sine-sweep the frequency response of the combined transducer-plant can be obtained after convolution of the recorded output by the corrected inverse of the input sweep. Correction in this case is a -6dB/Oct envelope to compensate for decreasing energy at higher sweep frequencies. Usefully, the microphone response will be included in this measurement.

A frequency response obtained from a full range exponential sine sweep of the system shown in Fig. 7.5 is plotted in Fig. 7.6. The system consists of a Monacor SPH-60X 30W transducer unit driven by a Spirit Powerpad. The horn couples the 80mm speaker cone to a 2mm aperture, and is attached to the system under test via a tight rubber coupling. The acoustic response is recorded using a G.R.A.S. Sound and Vibration 40SA 80mm probe microphone with a 12AA power module of the same manufacturer. Both output and recorded signals are converted from/to 24-bit samples at 192kHz by an RME Fireface 800. Simultaneous signal playback and recording is performed through Matlab in double precision floating point using pa_wavplay [112], based on the PortAudio API [113]. Measurements were initally made in a five-sided semi-anechoic chamber with acoustic foam arranged on the sixth side (concrete floor) to approximate an anechoic condition. The same measurements were later repeated in a full six-sided anechoic chamber.

Fig. 7.6 demonstrates significant variation in the magnitude response of the transducer system across the sub-10kHz range, exceeding the dynamic range of the system and hence ruling out direct equalisation (without dropping frequencies into the numerical noise floor). For measurement purposes, this issue can be addressed by decomposition of the sweep into overlapping bands, for each of which the frequency response can be corrected within an appropriate dynamic range. The corrected gain level can also be shifted by direct manipulation of the sweep length.

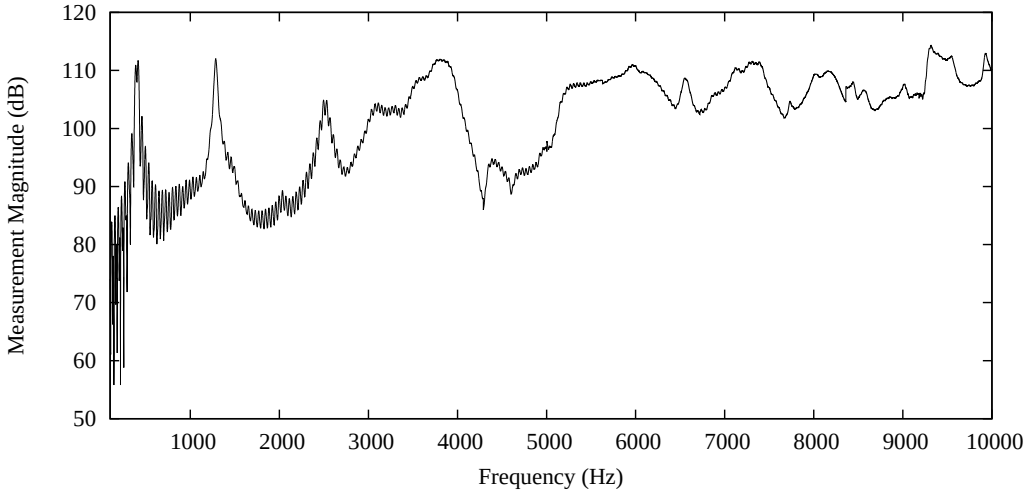


Figure 7.6: Combined transducer response obtained using full-range exponential sine sweep measurement

7.2.2 Transducer Equalisation

For complete inversion of the transducer colouration, a linear phase inversion of each band is required, together with band-specific manipulation of the sweep time to reach a given target gain. Inversion is approached in this case through a least-squared-error regression of the zero-order Volterra component of each band to an appropriate sinc-based bandpass kernel. The implementation is based on an all-pole autoregressive expression for the impulse response obtained by deconvolution of each chirp, as in (7.7), where $\hat{y}[n]$ is the system estimate, a_k the filter kernel and y the impulse response. The equation is resolved over a window of size p.

$$\hat{y}[n] = -\sum_{k=1}^{p} a_k y[n-k] \tag{7.7}$$

The estimation error e is defined by the difference between a sinc function $S_b[n]$ encapsulating each band b and the estimate as per (7.8):

$$e = sinc_b[n] - \hat{y}[n] \tag{7.8}$$

The consequent total squared estimation error E can then be defined as

(7.9), and since a minimised expression is sought, find the roots of the first derivative with respect to the filter coefficients, as in (7.10).

$$E = \sum_n e[n]^2 = \sum_n \left[S_b[n] - \hat{y}[n]\right]^2 \tag{7.9}$$

$$\frac{\partial}{\partial a_k}\left(\sum_n \left[S_b[n] - \hat{y}[n]\right]^2\right) = 0 \tag{7.10}$$

Rearranging (7.10) and substituting the derivative of (7.7) with respect to the filter coefficients, find (7.11), expressing the relationship between an autocorrelation of the system output (weighted by the filter coefficients) and a cross-correlation of the system output and desired sinc function.

$$\sum_n \sum_{i=1}^{p} a_k y[n-i]y[n-k] = -\sum_n S_b[n]\left(y[n-k]\right) \tag{7.11}$$

Where the number of filter coefficients exceeds the length of the impulse response, a delay is introduced to the target kernel for each band to centralise it within the window. The inversion of such a non-minimum phase system is by definition of infinite length, hence the approximation represented by the filter kernel increases in accuracy with its support. The system is finally resolved using the Levinson-Durbin recursion, exploiting the Toeplitz symmetry of the auto-correlation matrix [114].

The frequency response of the corrected source-system with reference load is plotted in Fig. 7.7, demonstrating a significant improvement in magnitude response consistency. Performance at higher frequencies is less impressive, but still within an acceptable range for deconvolution after measurement under the assumption of transducer linear time invariance. Vertical lines correspond to the limits of each band (with a 200Hz overlap either side of each) while the horizontal line indicates the target gain of 110dB. It should be noted that the system response is only linear-phase within each band, with a potentially inconsistent group delay between each.

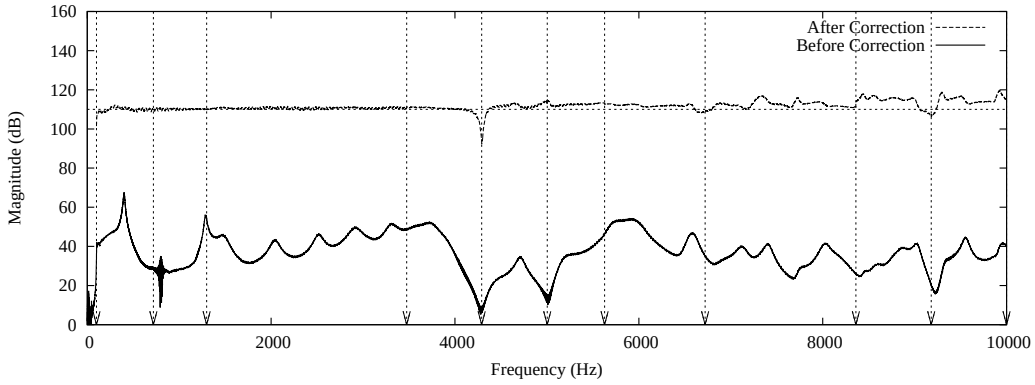


Figure 7.7: Frequency response of the band-adjusted source system before and after correction

7.3 Cylinders

As validation progresses, analogues of increasing geometry complexity are introduced. Simple cylindrical arrangements offer an increased complexity related to the simple cuboid yet their responses can still be approximated by direct mathematical determination. At this stage the most simple cylinders can also be used to validate the acoustic measurement process described in section 7.2.

Two types of cylinder-based analogues to the vocal tract are introduced. The first set are uniform quarter-wave cylinders, of lengths 160mm, 170mm and 180mm and diameters 16mm and 32mm. The changing lengths and cross-sectional areas allow for progressive cross-checking of the reproduction of axial and tangential resonant modes. The second set consist of concatenated pairs of cylinders with different cross-sectional areas, joined by a 0.013m linear connection and summing to 0.183m in length. These arrangements are shown in Figs. 7.9, 7.10 and 7.11. The concatenated cylinders are used to benchmark the representation of step changes in characteristic acoustic impedance, as introduced in section 4.3.2.

Mathematical approximation of each structure is performed in section 7.3.1, followed by presentation of the results of numerical simulation and corresponding acoustic measurements in section 7.3.2.

7.3.1 Lumped Calculations

In section 2.4 the acoustics of simple quarter-wave resonators was introduced, and an equation established for approximation of axial resonant frequencies. This described a simple one-dimensional treatment of the quarterwave formulation, augmented by approximation of the complex characteristic acoustic impedance of an unflanged open cylindrical end. The characteristic acoustic impedance Z_c [19, 20] is approximated in (7.12) where ρ_o is ambient air density, c the speed of sound, S the aperture cross-sectional area, k is angular wavenumber and r is the cross-sectional radius. The corresponding frequencies of resonance are expressed by (7.13).

$$Z_c = \rho_o c S \left[\frac{1}{4}(kr)^2 + j(0.6kr)\right] \tag{7.12}$$

$$f_R = \frac{(2m+1)c}{4(L+0.6r)} : m \in \mathbb{N}^0 \tag{7.13}$$

These frequencies are tabulated for all quarter-wave resonator configurations in Tables 7.2 and 7.3.

	Width 16mm		
n	160mm	170mm	180mm
0	517 (496)	487 (467)	461 (441)
1	1550 (1517)	1461 (1431)	1382 (1353)
2	2583 (2539)	2435 (2396)	2303 (2266)
3	3616 (3561)	3409 (3358)	3224 (3178)
4	4649 (4582)	4383 (4322)	4146 (4090)
5	5682 (5604)	5357 (5286)	5067 (5002)
6	6715 (6626)	6331 (6250)	5988 (5914)
7	7748 (7647)	7305 (7213)	6909 (6826)
8	8781 (8669)	8279 (8177)	7831 (7738)
9	9814 (9691)	9253 (9141)	8752 (8641)
10	10847 (10712)	10227 (10105)	9673 (9563)
11	11880 (11734)	11201 (11069)	10595 (10475)

Table 7.2: Calculated axial resonant mode frequencies (Hz) of 16mm-diameter uniform quarterwave resonators, where values in brackets correspond to maximal error conditions.

	Width 32mm		
n	160mm	170mm	180mm
0	502 (482)	474 (454)	449 (430)
1	1506 (1475)	1422 (1392)	1347 (1319)
2	2510 (2468)	2370 (2331)	2245 (2208)
3	3513 (3461)	3318 (3269)	3143 (3098)
4	4517 (4454)	4266 (4207)	4041 (3987)
5	5521 (5447)	5214 (5146)	4939 (4876)
6	6525 (6440)	6162 (6084)	5837 (5765)
7	7529 (7433)	7110 (7022)	6735 (6655)
8	8533 (8426)	8057 (7961)	7633 (7544)
9	9536 (9419)	9005 (8899)	8530 (8433)
10	10540 (10412)	9953 (9837)	9428 (9323)
11	11544 (11405)	10901 (10775)	10326 (10212)

Table 7.3: Calculated axial resonant mode frequencies (Hz) of 32mm-diameter uniform quarterwave resonators, where values in brackets correspond to maximal error conditions.

Uniform cylinders will also support a number of tangential resonant modes. The frequencies of these modes can be determined by the roots of the first derivative of a Bessel function of the first kind [5], the Maclaurin series for which is given in (7.14), where m represents the number of diametric nodal lines and n represents the number of circumferential nodal lines in the cylindrical cross-section.

$$J_m(x) = \left(\frac{x}{2}\right)^m \sum_{n=0}^{\infty} (-1)^n \frac{\frac{x}{2}^{2n}}{n!(n+m)!} \tag{7.14}$$

The roots of this Bessel function are provided in Table 7.4. The anticipated resonant frequencies are calculated using (7.15) after [115] and given for different m, n in Table 7.5.

$$f_{m,n} = J'_{m,n}\left[\frac{c}{2\pi r}\right] \tag{7.15}$$

Root α'_{mn} of $J'_m(x)$=0					
n	$m=0$	$m=1$	$m=2$	$m=3$	$m=4$
1	0.000	1.841	3.054	4.201	5.318
2	3.832	5.331	6.706	8.015	9.282
3	7.016	8.536	9.969	11.346	12.682
4	10.173	11.706	13.170	14.586	15.964
5	13.324	14.864	16.348	17.789	19.196

Table 7.4: Roots of the first derivative of spherical Bessel functions of the first kind, after [5], where n represents the number of diametric nodal lines and m the number of circumferential nodal lines.

From Table 7.5a it is clear that the narrow uniform cylinder is too small for the tangential resonant modes to appear in the sub-10kHz spectrum. Similarly, in the wider uniform cylinder there is only one tangential resonant mode firmly within the sub-10kHz range, although modes close to to 10kHz could still affect the visible magnitude response.

Mathematical determination of the resonant modes for the set of concatenated cylinders is not similarly straightforward. Section 4.3.2 introduces a one-dimensional approximation based on the assumption of a closed or open

n	$m = 0$	$m = 1$	$m = 2$
1	-	12471	20688
2	25958	36112	45427

(a) 160mm / 170mm / 180mm x 16mm

n	$m = 0$	$m = 1$	$m = 2$
1	-	6236 (5389)	10344
2	12979	18056	22713

(b) 160mm / 170mm / 180mm x 32mm

Table 7.5: Calculated tangential resonant mode frequencies (Hz) of uniform quarterwave resonators, where dimensions are given as length x diameter, n represents the number of diametric nodal lines, m the number of circumferential nodal lines and values in brackets describe the maximal error case.

end looking either way across the interface. The result is an equation of reactances whose cancellation leads to resonance at appropriate frequencies. This is repeated in (7.16), where A_1 and A_2 are the first and second cross-sectional areas and l_1 and l_2 represent the first and second cylinder lengths.

$$-\frac{\rho c}{A_1} cot(kl_1) + \frac{\rho c}{A_2} tan(kl_2) = 0 \tag{7.16}$$

The resulting approximate resonant frequencies for each concatenated cylindrical arrangement are given in Table 7.11.

R	f_R (Hz)
1	705 (671 ± 15)
2	1297 (1295 ± 15)
3	2708 (2636 ± 15)
4	3300 (3261 ± 15)
5	4711 (4603 ± 15)
6	5302 (5227 ± 15)
7	6714 (6568 ± 15)
8	7305 (7193 ± 15)
9	8717 (8534 ± 15)
10	9308 (9160 ± 15)

(a) 85mm x 8mm → 85mm x 16mm

R	f_R (Hz)
1	295 (312 ± 15)
2	1707 (1653 ± 15)
3	2298 (2279 ± 15)
4	3710 (3620 ± 15)
5	4301 (4245 ± 15)
6	5713 (5585 ± 15)
7	6304 (6210 ± 15)
8	7716 (7551 ± 15)
9	8307 (8176 ± 15)
10	9719 (9517 ± 15)

(b) 85mm x 16mm → 85mm x 8mm

R	f_R (Hz)
1	612 (595 ± 15)
2	1677 (1620 ± 15)
3	2327 (2300 ± 15)
4	3393 (3352 ± 15)
5	4618 (4540 ± 15)
6	5683 (5535 ± 15)
7	6333 (6222 ± 15)
8	7399 (7297 ± 15)
9	8624 (8571 ± 15)
10	9689 (9447 ± 15)

(c) 42.5mm x 8mm → 127.5mm x 16mm

R	f_R (Hz)
1	325 (340 ± 15)
2	1390 (1379 ± 15)
3	2615 (2566 ± 15)
4	3680 (3578 ± 15)
5	4330 (4260 ± 15)
6	5396 (5324 ± 15)
7	6621 (6512 ± 15)
8	7686 (7492 ± 15)
9	8336 (8184 ± 15)
10	9402 (9272 ± 15)

(d) 127.5mm x 16mm → 42.5mm x 8mm

R	f_R (Hz)
1	612 (595 ± 15)
2	1677 (1620 ± 15)
3	2327 (2300 ± 15)
4	3393 (3351 ± 15)
5	4618 (4540 ± 15)
6	5683 (5535 ± 15)
7	6333 (6222 ± 15)
8	7399 (7298 ± 15)
9	8624 (8484 ± 15)
10	9689 (9447 ± 15)

(e) 127.5mm x 8mm → 42.5mm x 16mm

R	f_R (Hz)
1	325 (340 ± 15)
2	1390 (1379 ± 15)
3	2615 (2566 ± 15)
4	3680 (3578 ± 15)
5	4330 (4260 ± 15)
6	5396 (5325 ± 15)
7	6621 (6512 ± 15)
8	7686 (7491 ± 15)
9	8336 (8183 ± 15)
10	9402 (9272 ± 15)

(f) 42.5mm x 16mm → 127.5mm x 8mm

Table 7.6: Approximate axial resonant mode frequencies (Hz) for concatenated cylinder arrangements as per Figs. 7.9, 7.10 and 7.11.

7.3.2 Simulation

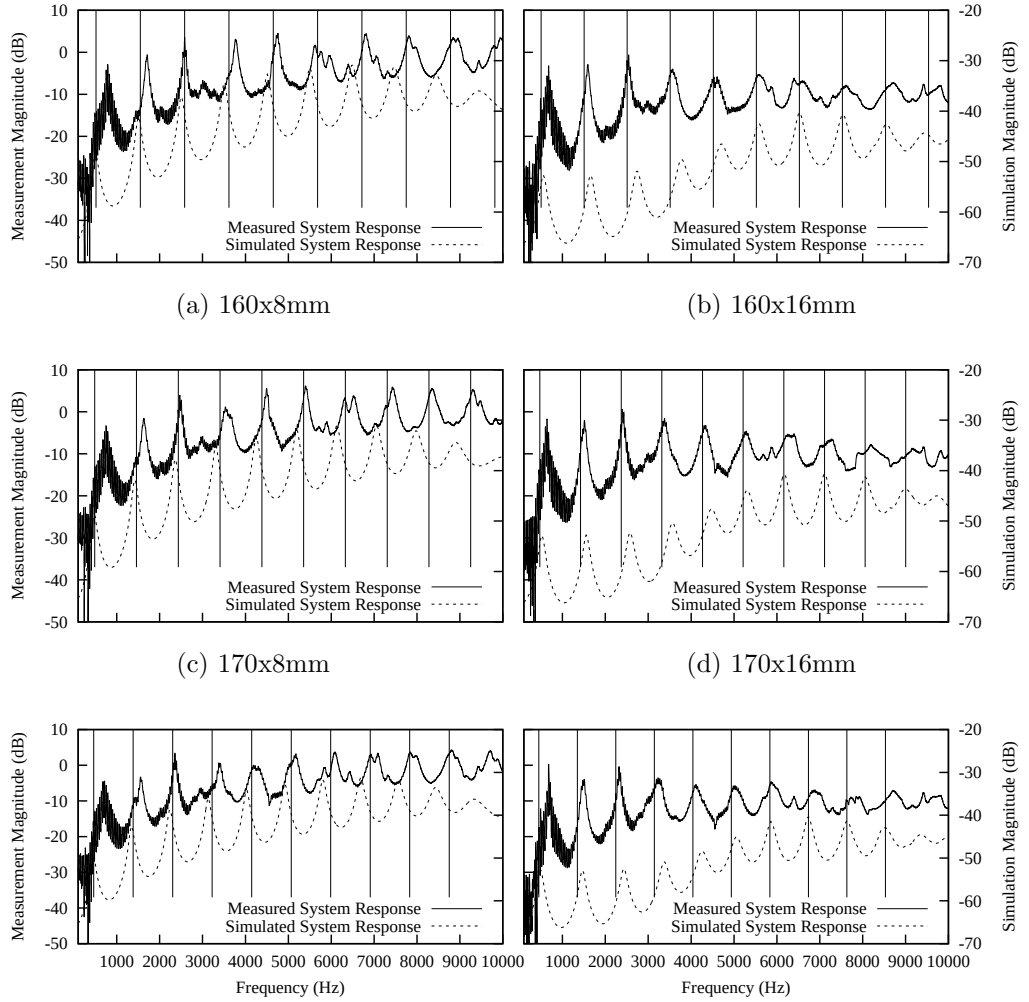
Simulation of the uniform cylinders was performed at 480kHz, and that for the concatenated cases at 720kHz to minimise spatial sampling error at the cylindrical interface. A sphere of diameter 0.06m was appended to the end of each graphical model, to allow direct simulation of the radiation impedance. A circular opening of radius 0.0026m was added to the closed end of each cylinder and coupled to a hemisphere, to represent radiation at the source aperture. Boundary elements constituting the two radiation domes have reflection coefficients set to 0.0 to approximate a free-field, the effectiveness of which is examined. Boundary elements in the body of the model have reflection coefficients of 0.99 to more closely approximate the mechanical analogues. A sinc function of 801-sample support is injected as in section 7.1.2, modified to provide a cutoff frequency of 20kHz at the lower sample rate. Injection is at the centre of the source opening and extraction at the centre of the main cylinder aperture. The simulation is run for 4000 time steps (equivalent to approximately 0.008s). The dispersion factor experienced in these simulations will be effectively double that of cuboid simulation due to the halved sample rate, yielding a maximum phase velocity error of $0.00024c$. For the example of the 10th resonant mode of the 170mm x 32mm quarterwave resonator (9005Hz) this corresponds to an error of approximately 3Hz, which can be disregarded.

The spatial sampling error in this case becomes more significant due to the reduced system sampling rate. For the quarterwave cylinders the maximal-error case can be approximated as $0.5d \leqslant e_{spat} \leqslant 1.5d$ for axial modes (since the cylinder is open ended each node can only be $0.5d$ away from the actual interface). For tangential modes the maximal-error case is $2d$ due to opposing boundaries, as with the cuboid simulation. The spatial sampling interval (d) in this case is 0.00123m. These error conditions are combined with the new reduced FFT half bin-width of $\pm$15Hz. The maximum expected error conditions for axial resonant modes in these quarterwave cylinders are given in brackets for each mode in Tables 7.2 and 7.3. For the 10th resonant mode of the 170mm x 32mm quarterwave resonator the maximum error condition

can yield to a movement in resonant frequency of approximately 100Hz, which is significant. The effect is more prevalent for the shorter cylinders.

An error analysis for the case of concatenated cylinders is slightly more complicated, but clearly relevant. In the case of spatial sampling error the grids representing either cylinder could be longer than the actual cylinders (hence the spatial error appears twice as often). It is possible to approximate the maximum shifts that might result by substitution of the potential length and radii extensions into (7.16) (where this extension is $2d$ for each constituent cylinder) and adjusting for the FFT half bin-width of $\pm$15Hz. These error conditions are given in brackets for each resonant mode in Table 7.11.

Simulated impulse responses are analysed as in section 7.1.2, and plotted in Fig. 7.8 for the case of the uniform cylinders and Figs. 7.9, 7.10 and 7.11 for concatenated cylinders.



(a) 160x8mm (b) 160x16mm

(c) 170x8mm (d) 170x16mm

(e) 180x8mm (f) 180x16mm

Figure 7.8: Simulated and measured magnitude responses of uniform quarterwave cylinders with calculated resonant modes as per Tables 7.2 and 7.3 shown as vertical lines. Dimensions given as length x radius

	Diameter 16mm								
	160mm			170mm			180mm		
n	Calc.	Sim.	Err.	Calc.	Sim.	Err.	Calc.	Sim.	Err.
0	517	502	15	487	467	20	461	441	20
1	1550	1505	45	1461	1431	30	1382	1353	29
2	2583	2505	78	2435	2396	39	2303	2266	37
3	3616	3505	111	3409	3358	51	3224	3178	46
4	4649	4508	141	4383	4322	61	4146	4090	56
5	5682	5504	178	5357	5286	71	5067	5002	65
6	6715	6489	226	6331	6250	81	5988	5914	74
7	7748	7471	277	7305	7213	92	6909	6826	83
8	8781	8445	336	8279	8177	102	7831	7738	93
9	9814	9452	362	9253	9141	112	8752	8641	111

Table 7.7: Calculated and simulated resonant mode frequencies (Hz) of 16mm-diameter uniform quarterwave resonators given with error figures.

	Diameter 32mm								
	160mm			170mm			180mm		
n	Calc.	Sim.	Err.	Calc.	Sim.	Err.	Calc.	Sim.	Err.
0	502	557	-55	474	520	-46	449	491	-42
1	1506	1655	-149	1422	1556	-134	1347	1469	-122
2	2510	2739	-229	2370	2578	-208	2245	2432	-187
3	3513	3779	-266	3318	3567	-249	3143	3373	-230
4	4517	4709	-192	4266	4479	-213	4041	4255	-214
5	5521	5581	-60	5214	5306	-92	4939	5061	-122
6	6525	6529	-4	6162	6170	-8	5837	5863	-26
7	7529	7540	-11	7110	7111	-1	6735	6735	0
8	8533	8555	-22	8057	8064	-7	7633	7639	-6
9	9536	9452	82	9005	9027	-22	8530	8540	-10

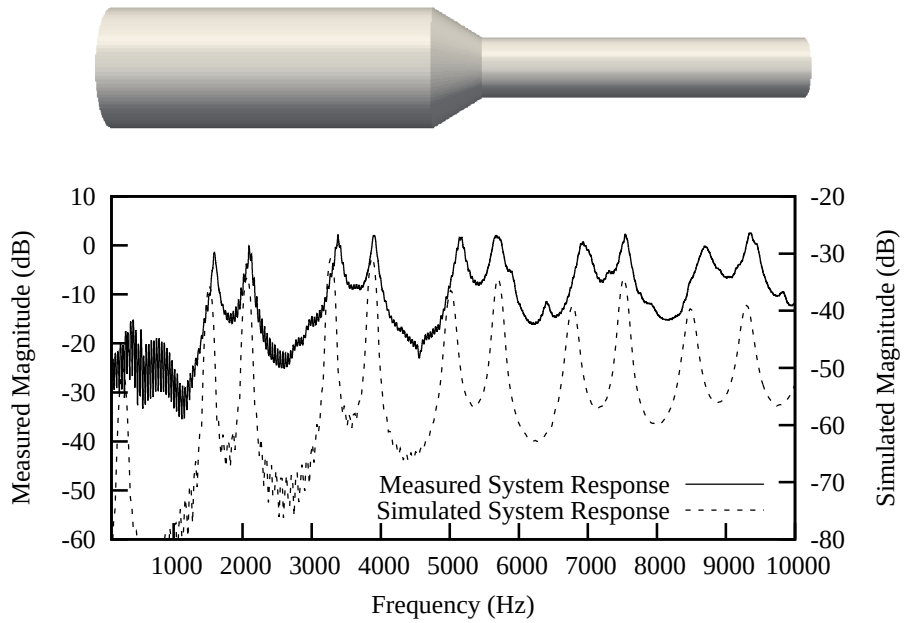
Table 7.8: Calculated and simulated resonant mode frequencies (Hz) of 32mm-diameter uniform quarterwave resonators given with error figures.

	Diameter 16mm								
	160mm			170mm			180mm		
n	Meas.	Sim.	Err.	Meas.	Sim.	Err.	Meas.	Sim.	Err.
0	791	502	289	753	487	266	709	461	248
1	1705	1505	200	1639	1461	178	1569	1382	187
2	2581	2505	76	2483	2435	48	2367	2303	64
3	3782	3505	277	3536	3409	127	3403	3224	179
4	4749	4508	241	4497	4383	114	4236	4146	90
5	5632	5504	128	5411	5357	54	5181	5067	114
6	6807	6489	318	6417	6331	86	6095	5988	107
7	7822	7471	351	7437	7305	132	7015	6909	106
8	8900	8445	455	8361	8279	82	7842	7831	11
9	9943	9452	491	9322	9253	69	8815	8752	63

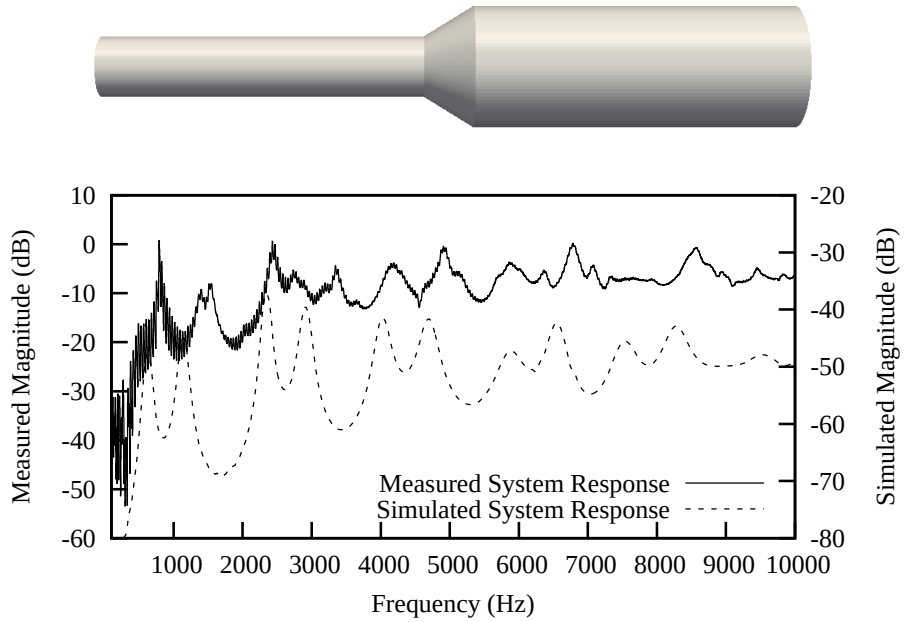
Table 7.9: Measured and simulated resonant mode frequencies (Hz) of 16mm-diameter uniform quarterwave resonators given with error figures.

	Diameter 32mm								
	160mm			170mm			180mm		
n	Meas.	Sim.	Err.	Meas.	Sim.	Err.	Meas.	Sim.	Err.
0	678	557	121	630	520	110	678	491	187
1	1592	1655	-63	1516	1556	-40	1500	1469	31
2	2537	2739	-202	2414	2578	-164	2323	2432	-109
3	3574	3779	-205	3388	3567	-179	3243	3373	-130
4	4573	4709	-136	4333	4479	-146	4113	4255	-142
5	5588	5581	7	5288	5306	-18	5027	5061	-34
6	6603	6529	74	6331	6170	161	5862	5863	-1
7	7614	7540	74	-	7111	-	6877	6735	142
8	8708	8555	153	-	8064	-	7800	7639	161
9	-	9452	-	-	9027	-	8714	8540	174

Table 7.10: Measured and simulated resonant mode frequencies (Hz) of 32mm-diameter uniform quarterwave resonators given with error figures.

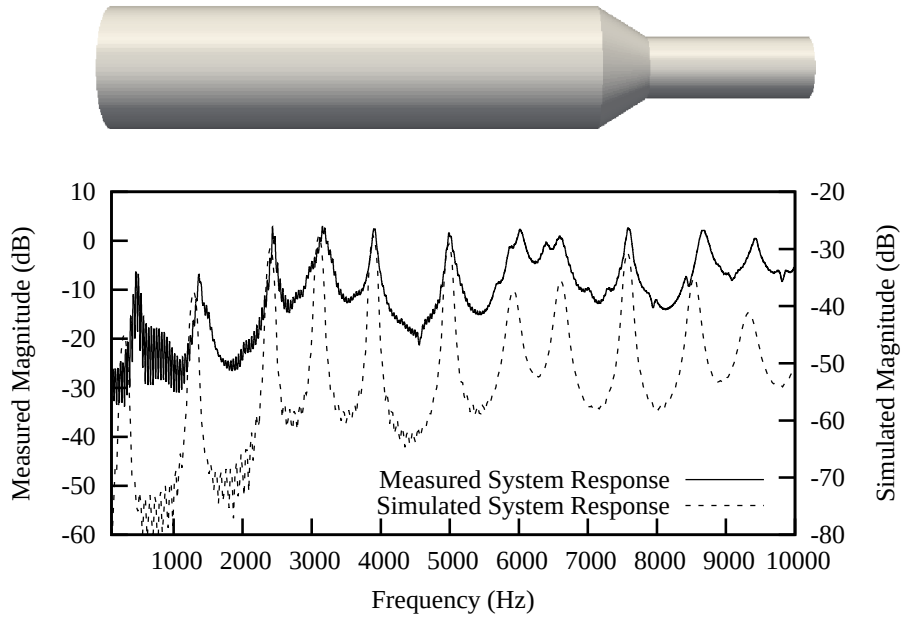


(a) 85x16mm → 85x8mm

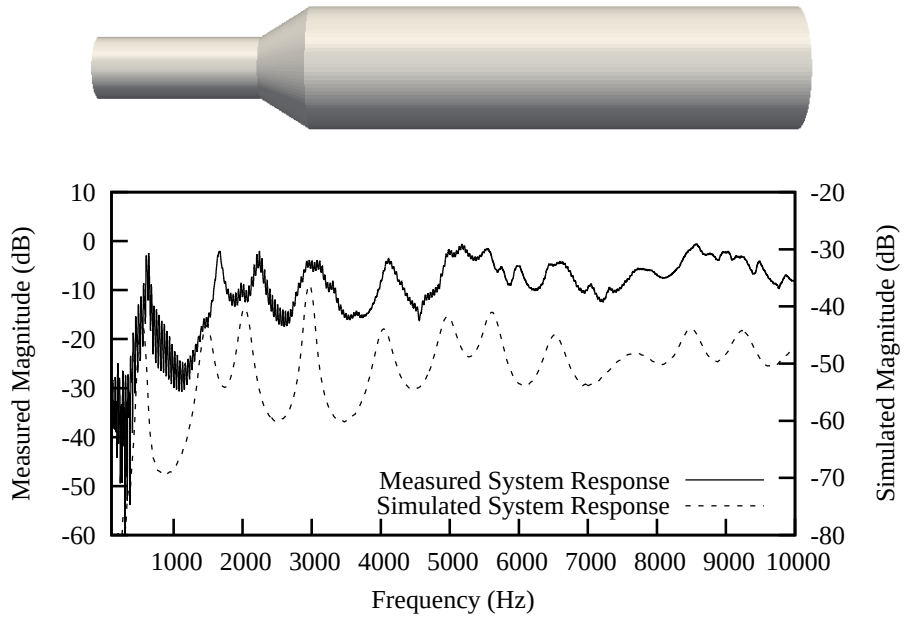


(b) 85x8mm → 85x16mm

Figure 7.9: Simulated and measured magnitude responses of concatenated cylindrical configurations. Dimensions are given as length × radius. Analogues displayed are closed on the left and open on the right.

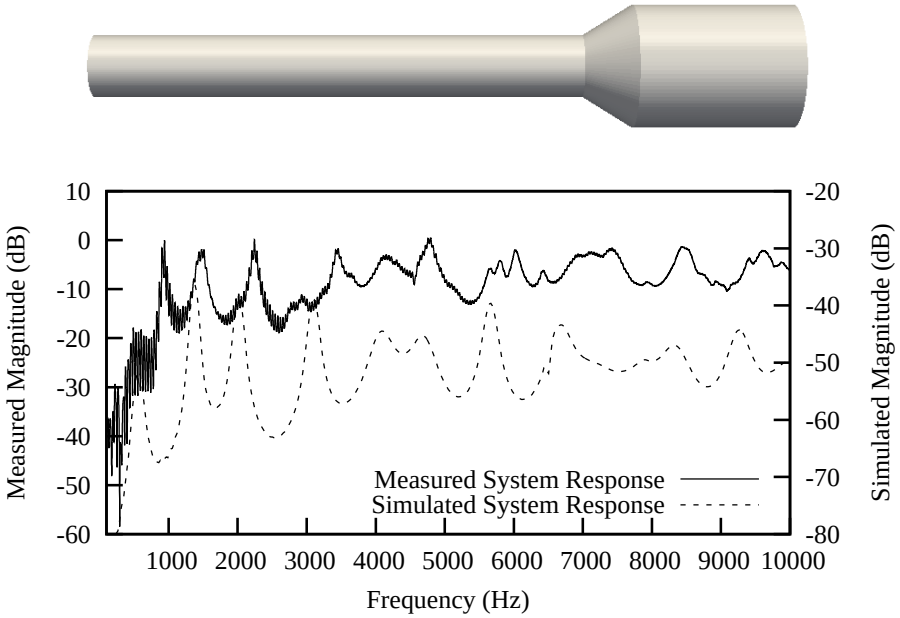


(a) 127.5x16mm → 42.5x8mm

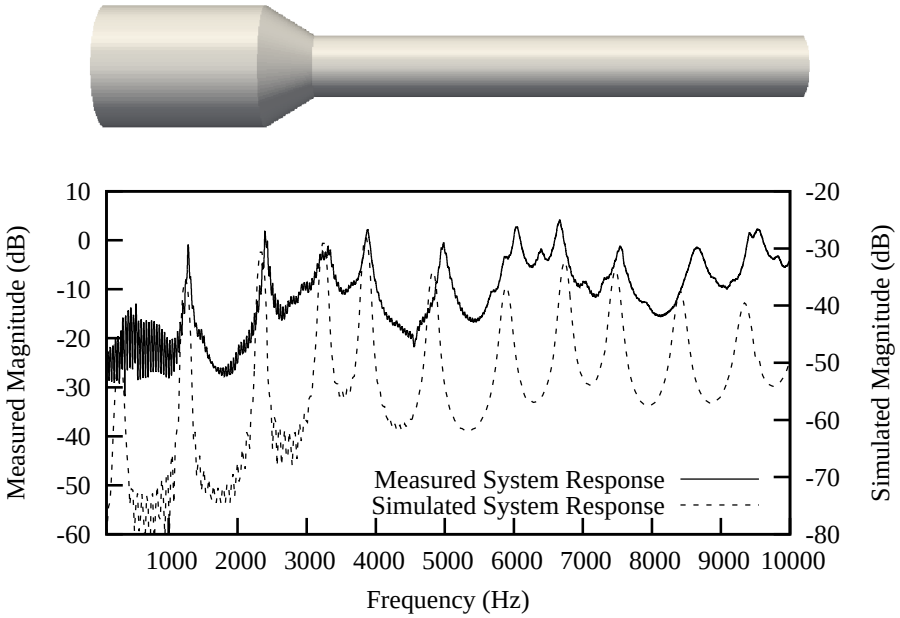


(b) 42.5x8mm → 127.5x16mm

Figure 7.10: Simulated and measured magnitude responses of concatenated cylindrical configurations. Dimensions are given as length × radius. Analogues displayed are closed on the left and open on the right.



(a) 127.5x8mm → 42.5x16mm



(b) 42.5x16mm → 127.5x8mm

Figure 7.11: Simulated and measured magnitude responses of concatenated cylindrical configurations. Dimensions are given as length × radius. Analogues displayed are closed on the left and open on the right.

R	C.	S.	e.
1	705	640	65
2	1297	1147	150
3	2708	2361	347
4	3300	2912	388
5	4711	4043	668
6	5302	5909	-607
7	6714	6546	168
8	7305	7561	-256
9	8717	8282	435
10	9308	9537	-229

(a) 85mm x 8mm → 85mm x 16mm

C.	S.	e.
295	271	24
1707	1522	185
2298	2067	231
3710	3281	429
4301	3880	421
5713	5701	12
6304	6785	-481
7716	7523	193
8307	8490	-183
9719	9307	412

(b) 85mm x 16mm → 85mm x 8mm

C.	S.	e.
612	545	67
1677	1491	186
2327	2030	297
3393	2966	427
4618	4980	-362
5683	5604	79
6333	6533	-200
7399	7709	-310
8624	8490	134
9689	9237	452

(c) 42.5mm x 8mm → 127.5mm x 16mm

R	C.	S.	e.
1	325	284	41
2	1390	1292	98
3	2615	2405	210
4	3680	3114	566
5	4330	3895	435
6	5396	4989	407
7	6621	6608	13
8	7686	7576	110
9	8336	8544	-208
10	9402	9329	73

(d) 127.5mm x 16mm → 42.5mm x 8mm

C.	S.	e.
612	573	39
1677	1384	293
2327	2014	313
3393	3104	289
4618	4119	499
5683	5663	20
6333	6700	-367
7399	7929	-530
8624	8308	316
9689	9275	414

(e) 127.5mm x 8mm → 42.5mm x 16mm

C.	S.	e.
325	299	26
1390	1245	145
2615	2345	270
3680	3243	437
4330	3858	472
5396	4825	571
6621	6738	-117
7686	7476	210
8336	8415	-79
9402	9351	51

(f) 42.5mm x 16mm → 127.5mm x 8mm

Table 7.11: Calculated and simulated axial resonant mode frequencies (Hz) for concatenated cylinder arrangements as per Figs. 7.9, 7.10 and 7.11, given with error figures.

R	M.	S.	e.
1	800	640	160
2	1484	1147	337
3	2446	2361	85
4	3360	2912	448
5	4182	4043	139
6	5878	5909	-31
7	6795	6546	249
8	-	7561	-
9	8575	8282	293
10	9461	9537	-76

(a) 85mm x 8mm → 85mm x 16mm

M.	S.	e.
416	271	145
1592	1522	70
2105	2067	38
3382	3281	101
3905	3880	25
5685	5701	-16
6940	6785	155
7545	7523	22
8708	8490	218
9360	9307	53

(b) 85mm x 16mm → 85mm x 8mm

M.	S.	e.
624	545	79
1667	1491	176
2253	2030	223
3035	2966	69
5172	4980	192
5550	5604	-54
6603	6533	70
7791	7709	82
8569	8490	79
-	9237	-

(c) 42.5mm x 8mm → 127.5mm x 16mm

R	M.	S.	e.
1	469	284	185
2	1368	1292	76
3	2446	2405	41
4	3174	3114	60
5	3911	3895	16
6	4995	4989	6
7	6603	6608	-5
8	7602	7576	26
9	8676	8544	132
10	9430	9329	101

(d) 127.5mm x 16mm → 42.5mm x 8mm

M.	S.	e.
507	573	-66
1491	1384	107
2244	2014	230
3442	3104	338
4157	4119	38
-	5663	-
-	6700	-
-	7929	-
8462	8308	154
9622	9275	347

(e) 127.5mm x 8mm → 42.5mm x 16mm

M.	S.	e.
475	299	176
1283	1245	38
2405	2345	60
3290	3243	47
3880	3858	22
4980	4825	155
6662	6738	-76
7545	7476	69
8661	8415	246
9527	9351	176

(f) 42.5mm x 16mm → 127.5mm x 8mm

Table 7.12: Measured and simulated axial resonant mode frequencies (Hz) for concatenated cylinder arrangements as per Figs. 7.9, 7.10 and 7.11, given with error figures.

The results for uniform cylinders are encouraging, both in terms of the accuracy of the acoustic measurement system and performance of the simulation. Fig. 7.8 demonstrates good matching between simulated and measured responses and close correlation with the mathematically approximated resonant modes. The centre frequencies of calculated, simulated and measured resonant modes are provided in Tables 7.7, 7.8, 7.9 and 7.10. There are clear differences between simulated and measured modes, more stark for the wider cylinders than the narrow equivalents, and constant in distance. This suggests an error related to cross-sectional area, most likely to be slight errors in the reproduction of the radiation field (and hence leading to poor representation of end-correction effects and consequent axial shifts). It is perhaps unsurprising that rectilinear representation of a circular cross-section leads to slight errors in representation of the characteristic acoustic impedance, especially for the case of the narrow cylinder for which the spatial sampling interval is a larger fraction of the diameter.

Differences between the mathematically approximated resonant mode frequencies and simulated results are smaller in most cases, although errors in the low frequency resonant modes of the larger diameter cylinders (Table 7.8) suggest a poor representation of low frequency behaviours. This is unsurprising considering the complex nature of the end-correction effect and its simple representation in the approximation. Inspection of Figure 7.8 confirms the strong performance of both measurement and simulation in these circumstances, by which each is clearly indicative of the correct behaviour, albeit shifted across the frequency range.

Simulation continues to perform well for the more complex geometrical configurations of the concatenated cylinders. Figs. 7.9, 7.10 and 7.11 give the simulated and measured responses for each of the concatenated cylindrical arrangements. Tables 7.12 and 7.11 list the simulated resonant modes against the analytically determined and measured equivalents. In all cases there is a one-to-one mapping between resonances in simulation, measurement and mathematical approximation. Figs. 7.9a, 7.10a, 7.10b and 7.11b exhibit clear, consistent behaviours right up to 10kHz (and conceivably beyond). Fig. 7.9b is interesting as the simulation provides results closer to the clear modal

separation expected from such a configuration. There are still strong trends between the two, although above 7kHz the simulation and measurement no longer match. The simulation of Fig. 7.11a is similarly interesting. While the measurement appears more trustworthy here than in Fig. 7.9b, unpredictable shifting in resonant mode frequencies is observed in the regions of 3kHz and 7kHz. The comparison of simulation and measurement at low frequency in this case (<1kHz) also demonstrates a shift of several hundred Hz in the first resonant mode. Despite these shifting errors, the overall performance of simulation is towards accurate reproduction of resonant modes across the entire 10kHz band.

Comparison of the results with the analytically determined cases is consistent at low frequencies, but quickly becomes erroneous for the cases of both simulation and measurement. This is evident in observation of the frequency errors of Table 7.11 and those observed in Figs. 7.9, 7.10 and 7.11. Considering likely error conditions, it is clear that the mathematical approximation represented by 7.16 is weak at low frequencies and without properly incorporating the effects of end correction (particularly at the interface). At this stage the value of the acoustic measurement as a means of establishing performance becomes evident. Table 7.12 provides a secondary measure of resonant behaviour, and while some frequency errors are clearly large, cross-referencing Figs. 7.9, 7.10 and 7.11 demonstrates consistent patterns with slight shifts.

7.4 Vocal Tract Analogues

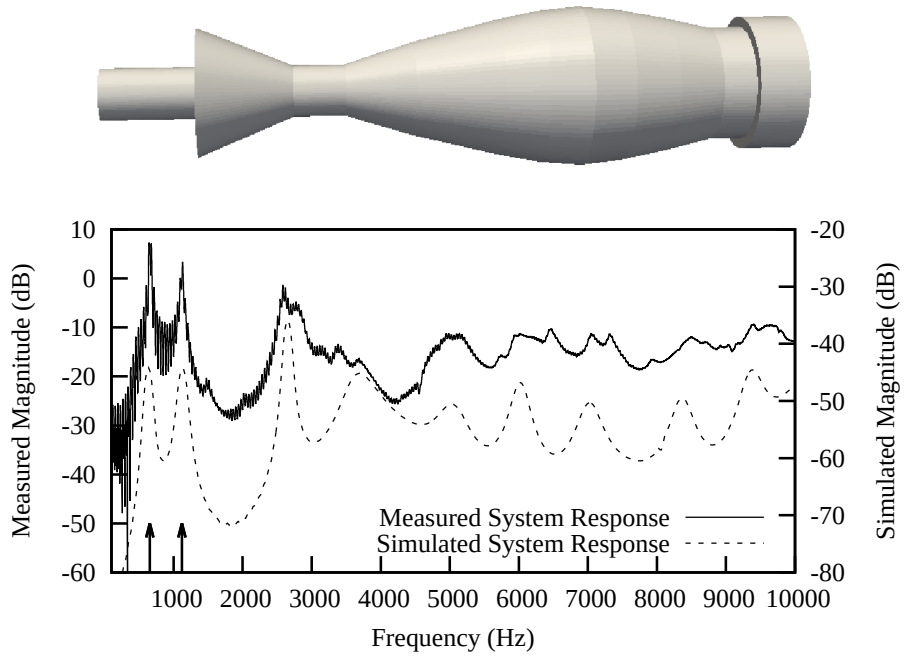
After concatenated cylinders, the next step of geometrical complexity is provided by real, mechanical analogues to the vocal tract. In this case, the VTM Vocal Tract Model kit, produced by Arai [9, 116] and based on geometries obtained from X-Ray imaging by Chiba and Kajiyama [53] are used. These are a series of clear acrylic cylinders, from which changing cross-sectional profiles are drilled, as in Figs. 7.12, 7.13 and 7.14. These analogues provide an appropriate approximation of the acoustic processes undergone in the vocal tract during phonation. The obvious omissions are the voice source and

vocal tract curvature (which has been shown to be significant [71]).

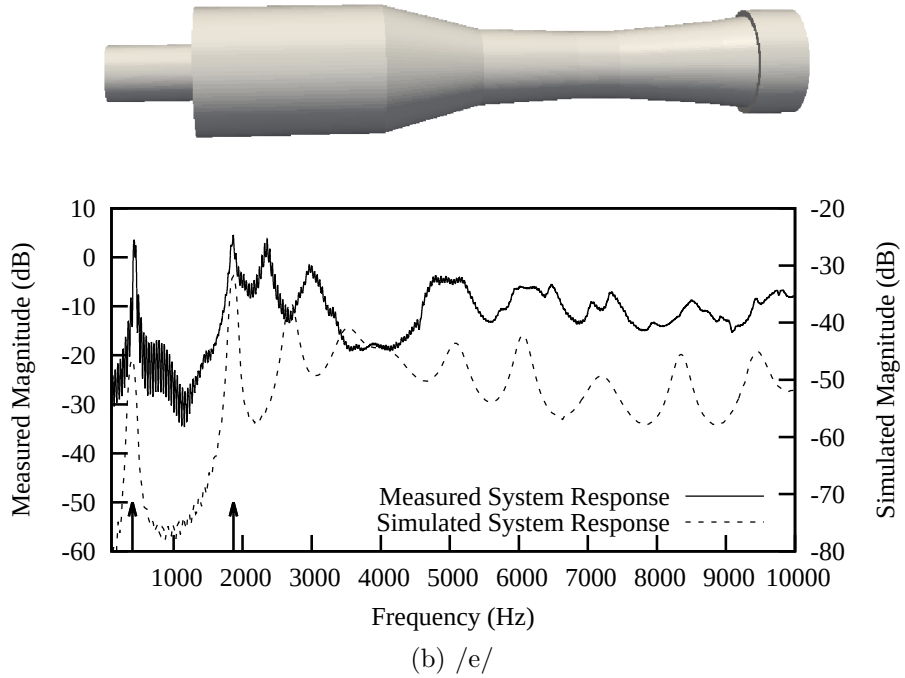
Mathematical determination of the acoustic response of these analogues is no longer considered to be reasonable, as even one-dimensional approximation is only valid for low frequencies (as demonstrated in section 7.3.2. Instead, validation depends primarily on the acoustic measurement technique introduced in section 7.2. The results of section 7.3.2 suggests that the measurement technique performs well across the 10kHz band.

7.4.1 Simulation

As for the concatenated cylinders simulations were run at a sampling rate of 720kHz, utilising an 801-sample support sinc function (with a bandstop frequency of 20kHz). Simulation is run for 10000 time steps, with extraction at the centroid of the mouth aperture. Injection is at the centre of the glottal face, parallel with the aperture. As with previous simulations numerical dispersion error is so small as to be insignificant at such a high sample rate. The simulations of section 7.3.2 suggest that spatial sampling error should not be an issue.

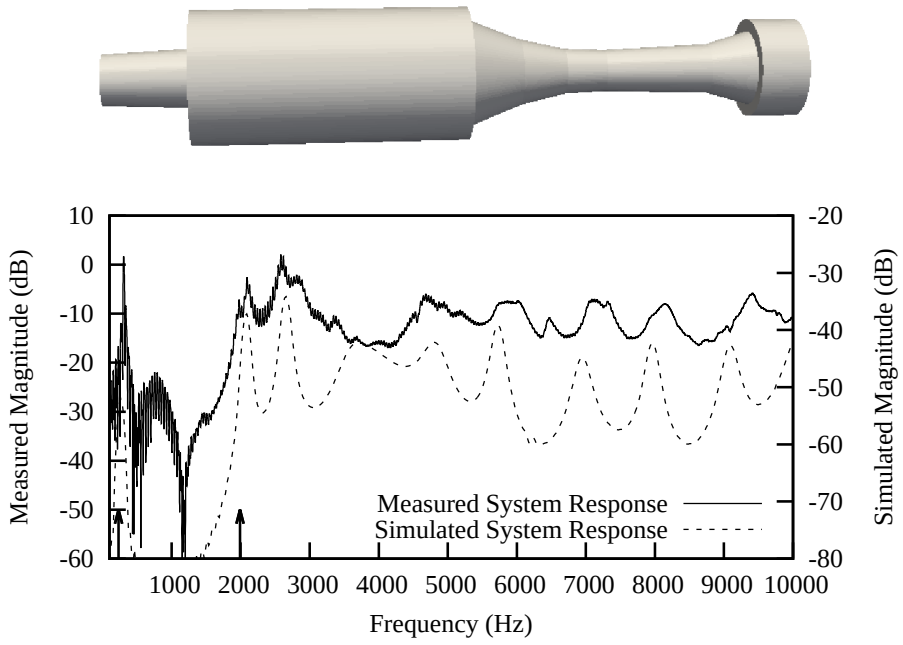


(a) /a/

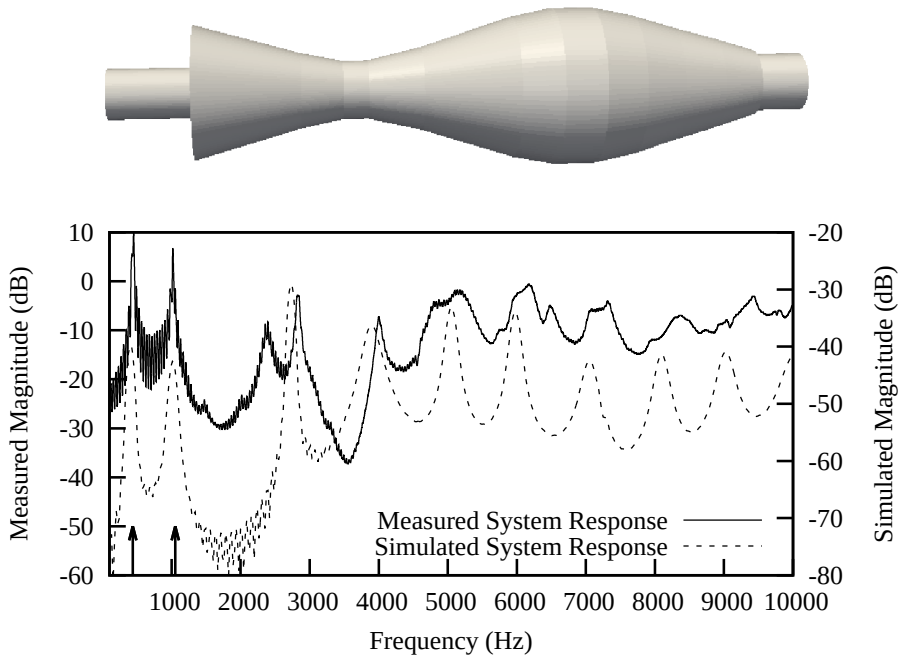


(b) /e/

Figure 7.12: Simulated and measured magnitude responses of vocal tract models with formant values as measured by Arai [9] indicated by arrows. Analogues displayed are closed on the left and open on the right.

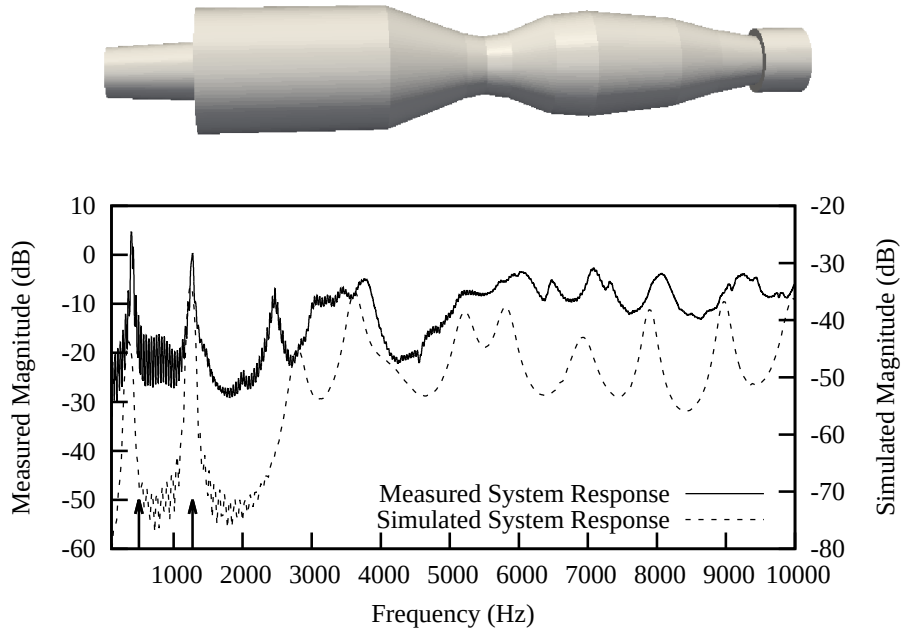


(a) /i/



(b) /o/

Figure 7.13: Simulated and measured magnitude responses of vocal tract models with formant values as measured by Arai [9] indicated by arrows. Analogues displayed are closed on the left and open on the right.



(a) /u/

Figure 7.14: Simulated and measured magnitude responses of vocal tract models with formant values as measured by Arai [9] indicated by arrows. Analogues displayed are closed on the left and open on the right.

The results of simulation and measurement are presented in Figs. 7.12, 7.13 and 7.14. The results are generally strong, exhibiting reasonable correlation between measured and simulated responses up to 10kHz although performance in the F3-F4 range is a concern for models 7.13b and 7.12b. Firstly, it is reassuring that the simulated and measured formants exactly agree, and both also agree with the formant measurements performed by Arai [9]. The best performance is perhaps observed in Fig. 7.12a for /a/, which provides an accurate representation of the measured response across the entire 10kHz band. Fig. 7.13a demonstrates a similarly strong performance for /i/, with the bandwidths of early formants the only mismatch between simulation and measurement. Simulation of /u/ exhibits generally consistent behaviours in Fig. 7.14a, although there is an upward frequency shift in F3 of ∼200Hz, which appears to continue in F4. Similar upward frequency shifts in F3 are seen in Figs. 7.13b and 7.12b, for /o/ and /e/. In the case of /o/, a small frequency shift effectively rolls F3 into F4 to create

a degenerate condition. In the case of /e/ the same shift in F3 is continued in F4.

Despite these issues, the overall trend in these simulations is positive. Slight shifts do occur, which are perhaps inevitable in any discrete representation of a continuously varying geometry. That the formant differences are inconsistent between geometries (and indeed certain concatenated cylinders) most likely rules out the possibility of errors in the measurements, instead suggesting certain-geometry dependent behaviours in simulation.

7.4.2 Listening Tests

For human listeners, vowel intelligibility is a perception-dependent measure of frequency-domain accuracy. Similarly, while frequency spectra provide a useful way of analysing each simulation in terms of linear system analysis, the most useful test is perhaps how they are perceived. To this end, a very simple listening test is performed, in which participants are played a series of audio clips. These consisted of acoustic recordings of the vocal tract analogues directly coupled to an electrolarynx, and convolutions of each simulated, downsampled impulse response with a recording of the electrolarynx alone. All acoustic recordings were made at 192kHz, using the same probe microphone as the main study and captured at 24-bit resolution using a RME Fireface 800 audio interface. No inversion of the microphone response was performed, and each audio clip was downsampled to 48kHz (after convolution in the case of the simulated version). The clips were played back in a random order over headphones, with the listener asked to make a choice from five buttons for each. These choices were labelled with the five vowel sounds targetted by the VTM analogues and given coarticulatory contexts. Each vowel must be assessed before the subject moves onto the next and each clip can be repeated as many times as desired. There were 20 participants, of different gender, language background and level of phonetic training. The results of this simple listening test were compiled into confusion matrices to demonstrate which vowel models are well identified and in which cases misidentification frequency occurs. These matrices are given in table 7.13 for

both cases.

	Interpretation				
	/a/	**/e/**	**/i/**	**/o/**	**/u/**
/a/	19	0	0	0	1
/e/	0	17	1	0	2
/i/	0	4	12	1	3
/o/	0	0	0	17	3
/u/	0	1	0	5	14

(a) Acoustic Recordings

	Interpretation				
	/a/	**/e/**	**/i/**	**/o/**	**/u/**
/a/	20	0	0	0	0
/e/	0	11	6	1	2
/i/	0	6	14	0	0
/o/	7	1	0	8	4
/u/	1	6	1	2	10

(b) 3D DWM Simulations

Table 7.13: Confusion Matrices for Recorded and Simulated Vowel Models

This is a particularly primitive listening test. Since the subjects are of such significantly differing backgrounds there are numerous uncontrolled variables, however with these shortcomings in mind it can still be informative. Firstly, it is clear that for each case the correct vowel is most frequently identified. Participants struggled to correctly identify /uː/ correctly, however this was consistent between the acoustic and simulated cases, suggesting a poor correlation between Japanese and English interpretations of the vowel. It is unsurprising to note that those simulations for which vowel identification was poor correlate to those in which a significant shifting of the third and fourth formants occurred. This was not the case for the acoustic recordings, suggesting that the error does indeed lie in simulation. It also suggests that all of the first four formants contribute to vowel intelligibility, in addition to naturalness and timbre. The increased energy regions at 4kHz do not appear to have affected intelligibility of the simulations, suggesting either that the error lies in measurement or simply that the importance of amplitude in this range is not significant.

7.5 Magnetic Resonance Imaging

In Chapter 6 the difficulties in validating MR imaging were introduced and discussed. Since the acoustic measurement technique introduced in section 7.2 is no longer an option and mathematical determination is unrealistic, direct recording of each phonation provides the only means of validation. As explained in section 6.1.1, each phonation was recorded before and after MRI scanning, in conditions matching that of the scanner as closely as possible.

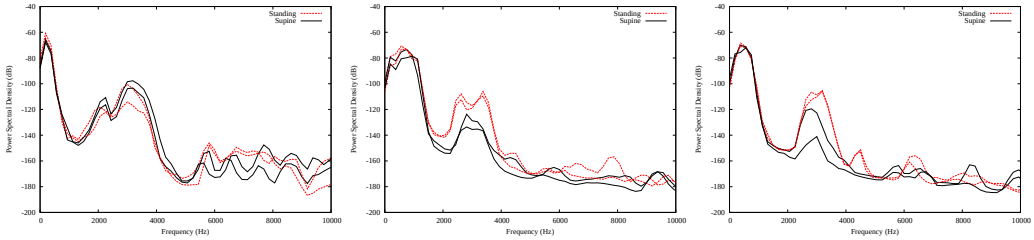
Before the results of such measurements can be trusted their validity must be ascertained. This encompasses three questions:

- Is supine phonation consistent with normal, standing phonation?
- Are recordings of successive phonations consistent?
- Are prolonged phonations adequately stable for the duration of a scan?

These questions are addressed in turn with respect to measurements of each of the participants, whose details are given in Table 6.2.

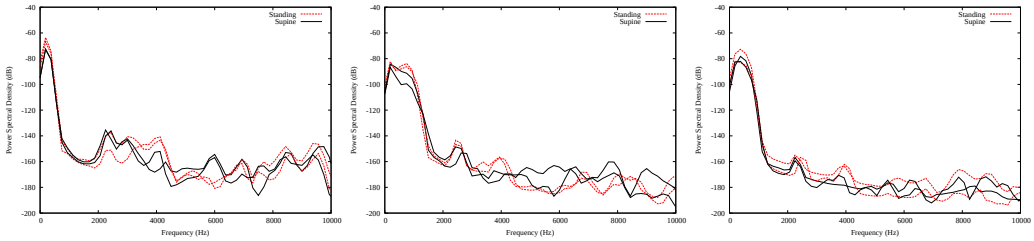
The Welch power spectral densities of each phonation (standing, supine, before, after) are calculated using a window length of 1024 samples and 50% window overlap and are plotted in Figs. 7.15 through 7.19 on a subject-by-subject basis. These suggest a strong consistency between standing and supine phonation up to approximately $4kHz$ in all subjects. Above this point consistency becomes both subject and vowel-dependent. There is an interesting trend towards a lowering of amplitude in the third and fourth formants in the supine case, although this cannot be isolated to fronted or back vowels. It is considered that supine phonation leads to inelective pharyngeal narrowing, hence a change in formant amplitudes might well suggest the use of compensatory articulation. It is observed that between supine and standing phonation the voice/vowel quality does not significantly change, even in those cases where the phonetic quality does. Jeff and Jasmine perhaps demonstrate the highest consistency between supine and standing phonation. Given their significant singing training this is perhaps understandable, since both are

used to maintaining vowel quality in potentially challenging feedback conditions. A further encouraging point is the absolute accuracy in matching of the first and second formants between standing and supine phonations. In some cases subjects are able to match all of the first four formants in both frequency and relative amplitude, as demonstrated by Jim in Figs. 7.18b and 7.18c and approached by Jasmine in Figs. 7.17a and 7.17b.



(a) /iː/ (b) /ɑː/ (c) /ɔː/

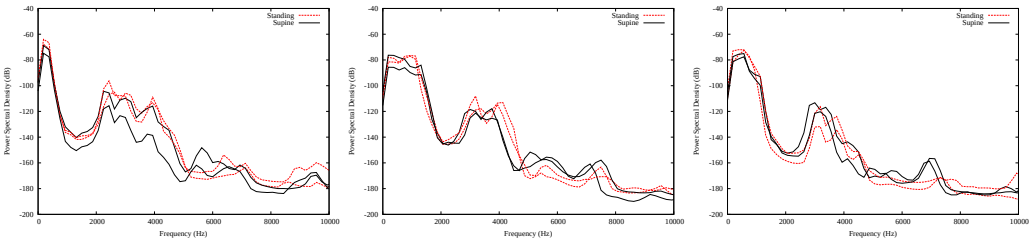
Figure 7.15: Welch power spectral densities for Standing/Supine phonations performed by Jack, before and after scanning



(a) /iː/ (b) /ɑː/ (c) /ɔː/

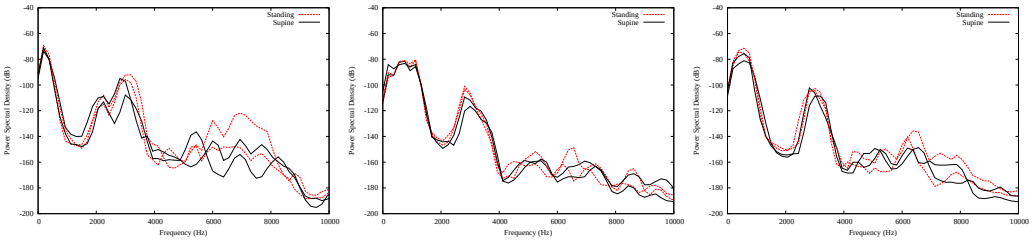
Figure 7.16: Welch power spectral densities for Standing/Supine phonations performed by Jill, before and after scanning

Successive supine phonations are shown to be very consistent, in most cases up to 8kHz and in some cases up to 10kHz. This is encouraging, especially considering that subsequent measurements were made over one hour apart and suggests that such audio recordings provide a valid benchmark for the models developed after MR imaging. The worst performing repetitions are perhaps Fig. 7.18a, 7.17a and 7.16b. It is reassuring to note that for other vowel arrangements successive phonations are not similarly



(a) /iː/ (b) /ɑː/ (c) /ɔː/

Figure 7.17: Welch power spectral densities for Standing/Supine phonations performed by Jasmine, before and after scanning



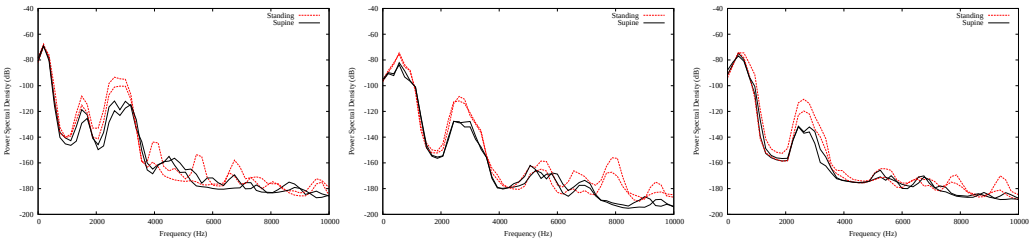
(a) /iː/ (b) /ɑː/ (c) /ɔː/

Figure 7.18: Welch power spectral densities for Standing/Supine phonations performed by Jim, before and after scanning

inconsistent. Additionally, subjects frequently produce more consistency in successive phonation while standing (see Figs. 7.17a and 7.19c).

To observe the effect of prolonged phonation, linear prediction of 1024-sample windowed segments (without overlap) of each phonation are performed. Since the recordings were made at 192kHz, 194 linear prediction coefficients were calculated to provide a suitable envelope to the formant structure. Treating the resulting coefficients as polynomials, the first four roots of each window were recorded and stored. This allows the visualisation of formant tracks, as shown in Figs. 7.20 to 7.24 for each of the subjects.

Fig. 7.20 shows the formant tracks for Jack. These measurements were recorded as part of the prototyping phase of the experiment, and featured repeated phonation (that is, the subject may stop phonating, inhale, then restart phonation during scanning). At approximately 10 seconds in Fig. 7.20a and 14 seconds in Fig. 7.20b the formants are clearly unsettled. In the case

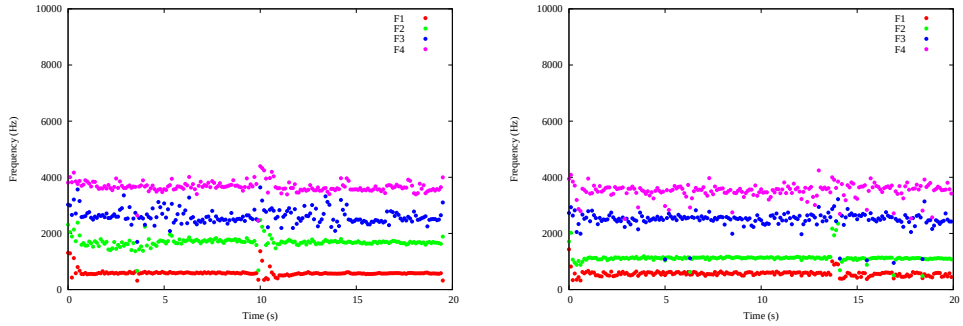


(a) /iː/ (b) /ɑː/ (c) /ɔː/

Figure 7.19: Welch power spectral densities for Standing/Supine phonations performed by Jeff, before and after scanning

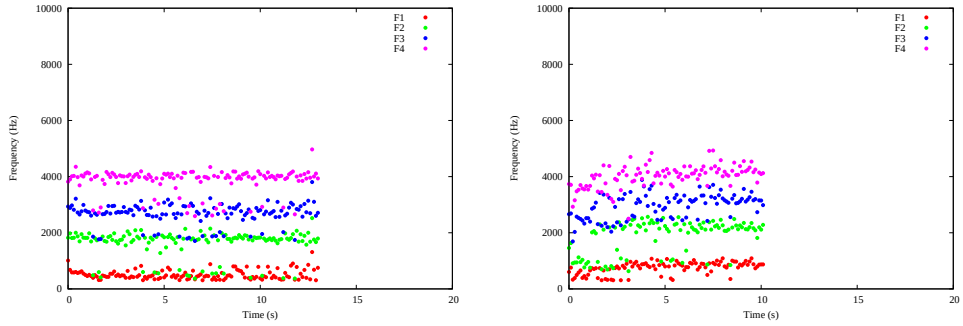
of Fig. 7.20a the formants settle down again to similar values to that before the break, whereas in Fig. 7.20b the formant structure is more unsettled. Onset in phonation leads to significant instability in the formant structure, underlining the value of using only a single phonation. Consequently, some of the following subjects display significantly shorter phonations.

In most cases, each subject demonstrates strong formant stability throughout phonation. Particularly strong cases are shown in Figs. 7.22a, 7.22b, 7.23a and 7.23b. It is also observed that the first and second formants demonstrate more stability than the third and fourth, as in Figs. 7.21b, 7.22b and 7.20a. This is perhaps unsurprising considering the greater sensitivity of higher frequency resonant behaviour to slight variation in articulation. The standard deviation of each of the Formant tracks for Jack's phonation on /ɛː/ is shown in Fig. 7.25, calculated at 10Hz. After onset this demonstrates stable phonation for the duration of the anticipated scan time. While there are some small (and acceptable) fluctuations in higher formants, little formant variation is exhibited until vowel release at approximately 20s.



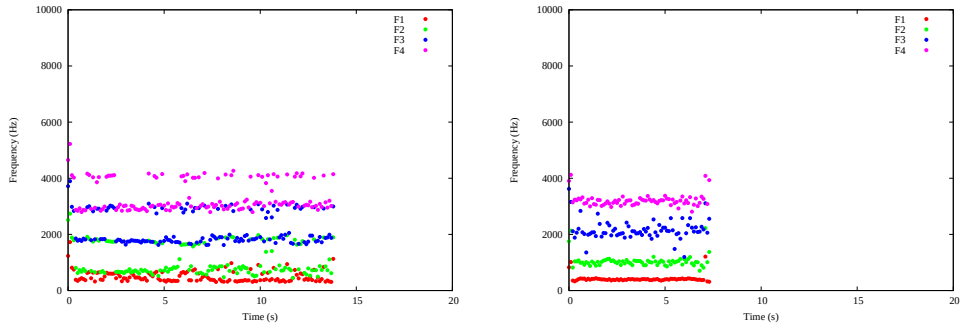
(a) /ɛː/ (b) /ʌː/

Figure 7.20: Formant tracks (from roots of 194 point linear prediction of 1024-sample windows) for supine phonations performed by Jack before scanning



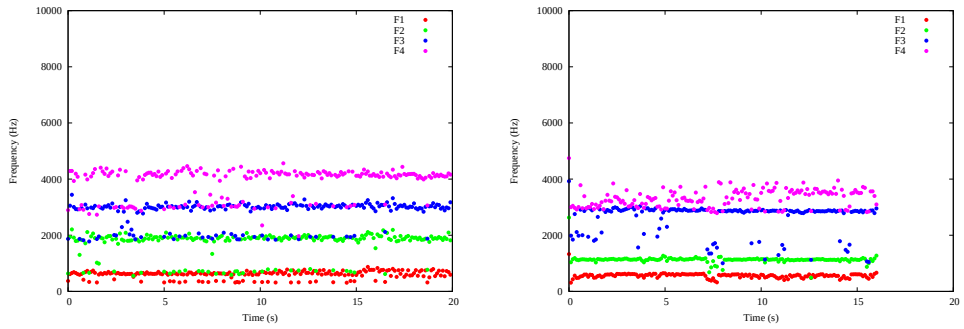
(a) /ɛː/ (b) /ʌː/

Figure 7.21: Formant tracks (from roots of 194 point linear prediction of 1024-sample windows) for supine phonations performed by Jill before scanning



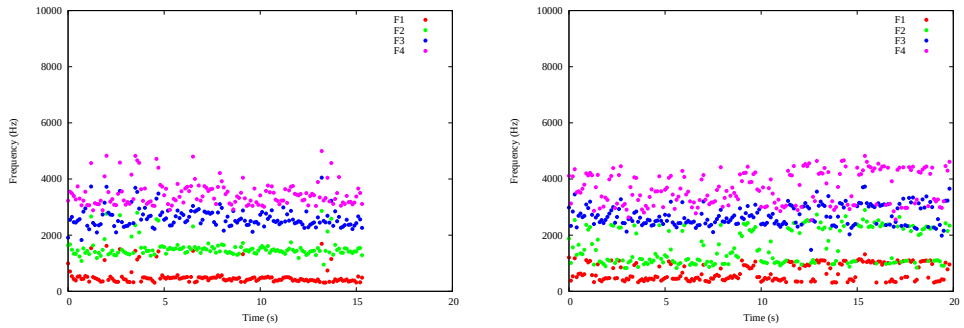
(a) /ɛː/ (b) /ʌː/

Figure 7.22: Formant tracks (from roots of 194 point linear prediction of 1024-sample windows) for supine phonations performed by Jasmine before scanning



(a) /ɛː/ (b) /ʌː/

Figure 7.23: Formant tracks (from roots of 194 point linear prediction of 1024-sample windows) for supine phonations performed by Jim before scanning



(a) /ɛː/ (b) /ʌː/

Figure 7.24: Formant tracks (from roots of 194 point linear prediction of 1024-sample windows) for supine phonations performed by Jeff before scanning

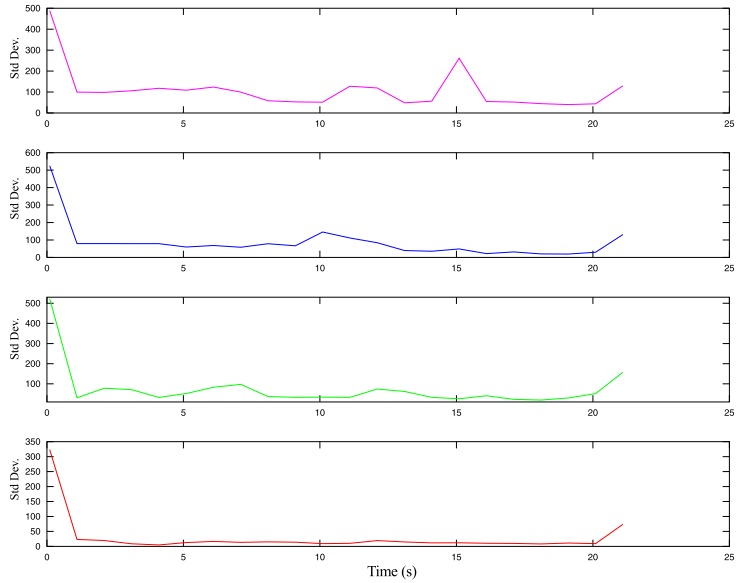


Figure 7.25: Phonation on /ɛː/ by Jack, displayed as 10Hz standard deviation of 1024-sample 194-coefficient LPC-based formant tracks using 0.0w overlap at 192kHz

Summary

This chapter has considered validation of the numerical simulation techniques developed in Chapter 6, using a range of vocal tract analogues of different geometrical complexity. In doing so, a number of important facts have been deduced. Spatial sampling error has been identified as being a more prevalent form of error than numerical dispersion error. The effect of spatial sampling error is seen to increase with frequency and inversely to geometry dimensions. It has been observed that the acoustic measurement technique is not necessarily perfect, particularly subject to dips in response at approximately 4kHz. It has however been demonstrated to be appropriate for validation through its accurate reproduction of more simple geometries. Simulation is seen to be acceptably accurate for both uniform cylinders, and simple concatenated cylindrical arrangements. It is also able to reproduce the acoustic behaviour of the mechanical vocal tract analogues, although inconsistent shifting of resonant modes is observed.

In section 7.5 it is noted that supine phonation can be considered consistent with standing phonation at least up to 4kHz, and above this frequency many behaviours are similar. It has been shown that successive supine phonations can be considered consistent at frequencies up to 8kHz in most cases. It is also observed that vocal tract stability is good for shorter, single phonations.

Having addressed the validity and likely errors observed in simulation of vocal-tract-like structures, Chapter 8 proceeds to present the results of simulation of vocal tract geometries extracted from MRI scanning.

Chapter 8

Results of Simulation

Introduction

In Chapter 6 the techniques used for numerical simulation using the three-dimensional digital waveguide mesh were developed. The reliability of its application to simulation of the acoustic field in vocal tract-like structures was consequently addressed in Chapter 7. In this Chapter the same techniques are applied to models of the vocal tract, developed as described in section 6.3. This results in a series of impulse responses, each representing a given vocal tract configuration as a linear time-invariant system. The results of such simulation are presented on a subject-by-subject basis, in section 8.1. In some cases certain vowel configurations are missing as severe motion artefacting in the imaging obstructs segmentation. Audio examples of each resynthesised vowel are presented and compared with those recorded during benchmarking for each subject. These examples are generated by convolution of each simulated impulse response with the corresponding electroglottographic signal. Section 8.3 continues to consider the contribution and applicability of the electroglottographic source signal to the resulting voice synthesis.

An important characteristic of the three-dimensional simulation in the context of the present hypothesis, is its performance relative to lower dimensionality derivative models. In section 8.2 corresponding models and the results of their simulation are presented and compared. Finally, section 8.4

describes the complete corpus of vocal tract imaging and benchmarking data collected during the course of this study.

8.1 Three-Dimensional Simulation

The first results presented here represent full three-dimensional simulation of the vocal tract models, as described in section 6.3.2. Each of these simulations is performed at 960kHz (corresponding to a waveguide length of 0.61mm). Single source and receiver points are used, injecting a sinc function tailored to a cutoff frequency of 20kHz. Boundary reflection coefficients on the edge of the radiating domes are assigned to $r = 0.0$ to provide a maximally absorptive condition. Those in the vocal tract itself are set to a uniform reflection coefficient of $r = 0.99$. This is experimentally deduced (not measured) to provide a reasonable formant bandwidth. Impulse responses are 8000-samples long, the short length of which causes occasionally ripple artefacting at lower frequencies. They are transformed using a zero-padded 16384 point FFT and the magnitude responses provided. For greater consistency with measured vowel power spectral densities, the impulse response is also convolved with the original electroglottographic source and the PSD plotted alongside the impulse magnitude response. This illuminates (but does not necessary match exactly) the harmonic contribution of the source.

8.1.1 Jack

Jack (male tenor - see Table 6.2) was the prototype subject for the project, meaning his scans were used to fine-tune the scanning protocol and procedures. He was advised to restart phonation after a pause if necessary during phonation. This was later seen to introduce slight instability in formant positions (due to onset) hence the advice was changed after his scans. Jack chose a target pitch of 220Hz for phonation, corresponding to the musical note A3.

Magnitude responses for simulation of each of the target vowels are plotted in Figs. 8.1 and 8.2, alongside power spectral densities for supine acoustic measurements of each vowel recorded before and after simulation (as de-

scribed in section 7.2). The power spectral density of the convolution of each impulse response with its corresponding electroglottogram signal is plotted on the same graph. The accuracy of the simulations (with respect to the benchmark audio recordings) varies between target phones. Low frequencies (particularly the first 2 formants) exhibit accurate behaviour with the first formant precisely located in all. For certain target phones a slight lowering of frequency in the third and fourth formants can be observed (also the second formant in the case of 8.1a). Some vowels exhibit an excess in high-frequency energy (see 8.1h), but in others the levels are well matched (as in 8.1c). In most cases the resynthesised vowels are suitable up to approximately 6kHz. above this point performance is erratic. Some phones, such as /ɛː/ (Fig. 8.1c) and /ʊː/ (Fig. 8.1h) exhibit behaviours that are characteristic of the target above 6kHz, while shifted in amplitude or frequency (consistent with the more complex vocal tract analogues of section 7.4). Others no longer bear any direct correspondence to the target phone at high frequencies (as is the case with /ɜː/ - Fig. 8.2c).

It is satisfying to observe that convolution of the source with the simulated impulse response results in power spectral densities of magnitude matching that of the recorded equivalents. This suggests that the very basic damping and absorptive behaviours implemented here are at least appropriate to re-synthesis of the VTTF.

Audio examples (simulated and recorded) can be found in Appendix D. The fixed graphical structure of the vocal tract models significantly affects the resulting resynthesised vowels. When compared to the recorded equivalents, none of the resynthesised examples exhibit an onset period other than those features attributable to the source (fundamental, amplitude changes). In the recorded audio, a period of formant settling is observable on spectrograms during vowel onset. Similarly, the absolute rigidity of the resonant structure during simulated phonation is unusual. With these considerations aside, the accuracy of a number of the resynthesised vowels is good (/ɪː/, /ɒː/ and /iː/ for example - see Appendix D). A number of the vowels produce the audible effect of constriction (/æː/ and /ɑː/ for example) which could be attributable to the supine condition were the audio recordings not made in the same

condition. It is possible the effect is caused by the frequency-independent damping behaviour imparted by the basic boundary formulation.

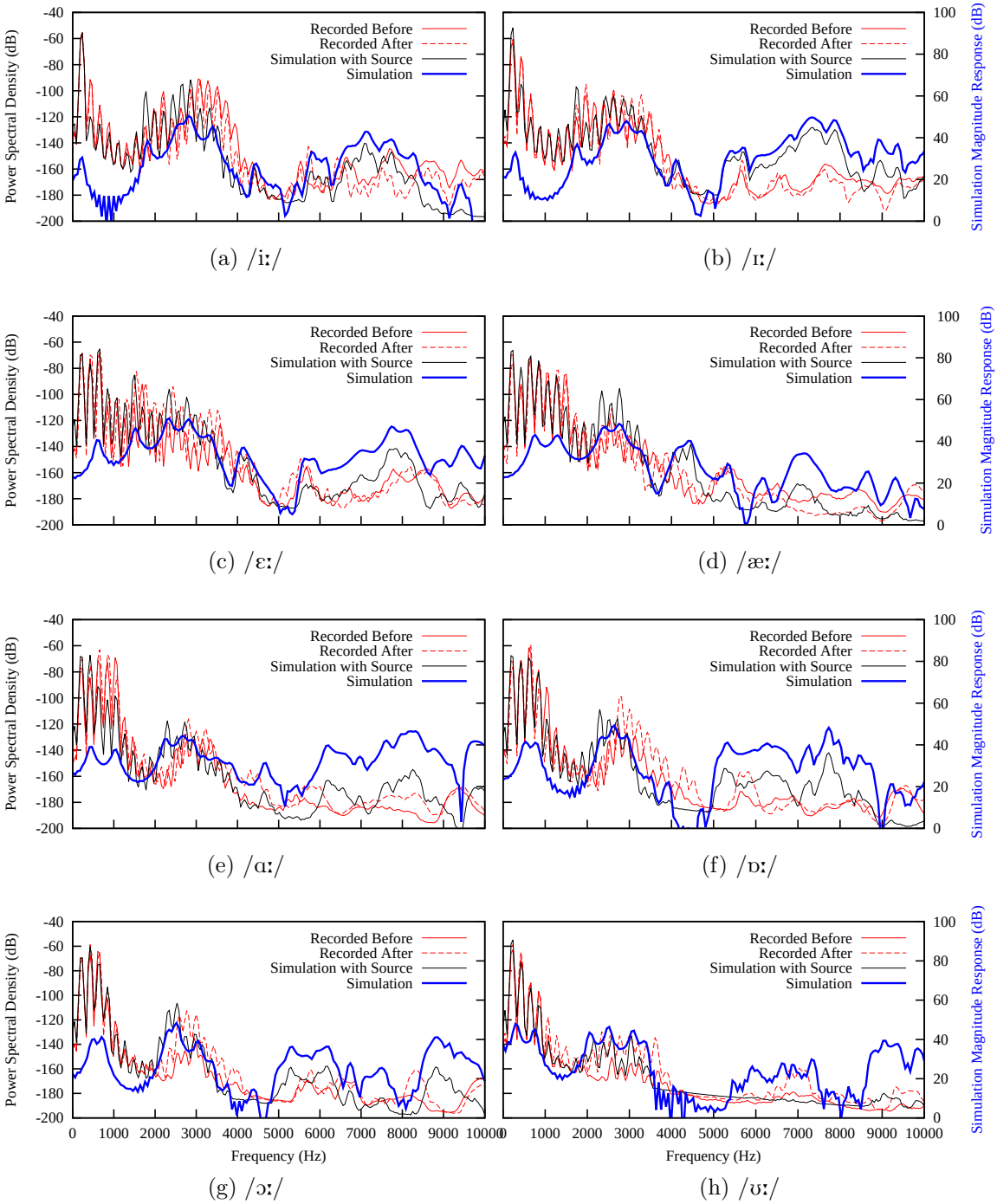
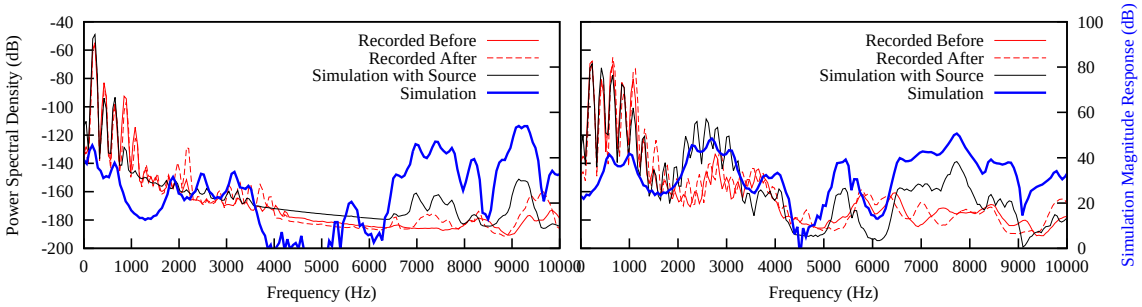
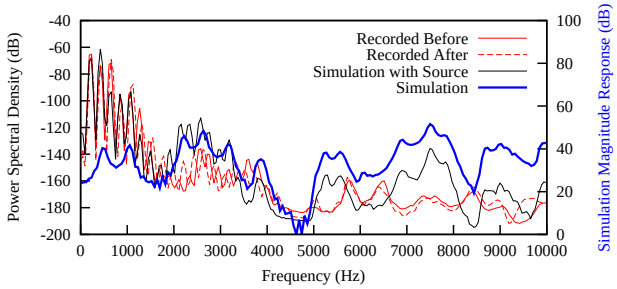


Figure 8.1: Jack - Male tenor - Comparison of three-dimensional simulations for each vocal tract model with recorded vowel power spectral densities



(a) /uː/ (b) /ʌː/



(c) /ɜː/

Figure 8.2: Jack - Male tenor - Comparison of three-dimensional simulations for each vocal tract model with recorded vowel power spectral densities

8.1.2 Jill

Jill is a trained female phonetician, with no singing background. Her vowel power spectral densities and simulation magnitude responses are plotted in Figs. 8.3 and 8.4. Jill chose to phonate at a pitch of 247Hz (corresponding to the note B3).

As with the previous subject, the simulation of certain phones demonstrates significant formant shifting and differences in formant amplitudes. The simulation of /iː/ for example (in Fig. 8.3a exhibits higher amplitudes in F3 and F4 than suggested by acoustic measurement. This trend continues to higher frequency behaviour and is particularly evident in /ɑː/, /ʊː/, /æː/ and /ɪː/. A number of the simulations hence exhibit excessive energy at high frequencies, which might be a manifestation of insufficient roll-off in the electroglottographic source waveform since it does not directly reflect the supra-glottal pressure waveform. It is also conceivable that this effect may be caused by the lacking representation of boundaries provided by frequency-independent damping (section 3.6). The overall trends of each simulation are however good.

A particularly interesting artefact can be seen in /ɜː/, where a trough in both recorded power spectral densities at 3kHz corresponds to a resonance in the simulated response of Fig. 8.4c. It is interesting to consider that this position lies in a conceivable range for a notch introduced by (even slight) nasal coupling in the recorded case.

The audio examples (Appendix D) reveal a number of features defining the nature of this three-dimensional simulation. Firstly, a number of the resynthesised vowels represent extremely strong matches to the real versions (such as /ɜː/, /ɔː/, /uː/, /ʊː/ and /iː/). Despite providing strong phonetic matches, the voice quality/timbre is often slightly different. This is consistent with the performance of the vowels in Figs. 8.3 and 8.4, as consistent formant behaviour is observed, coupled with an unpredictable high frequency response.

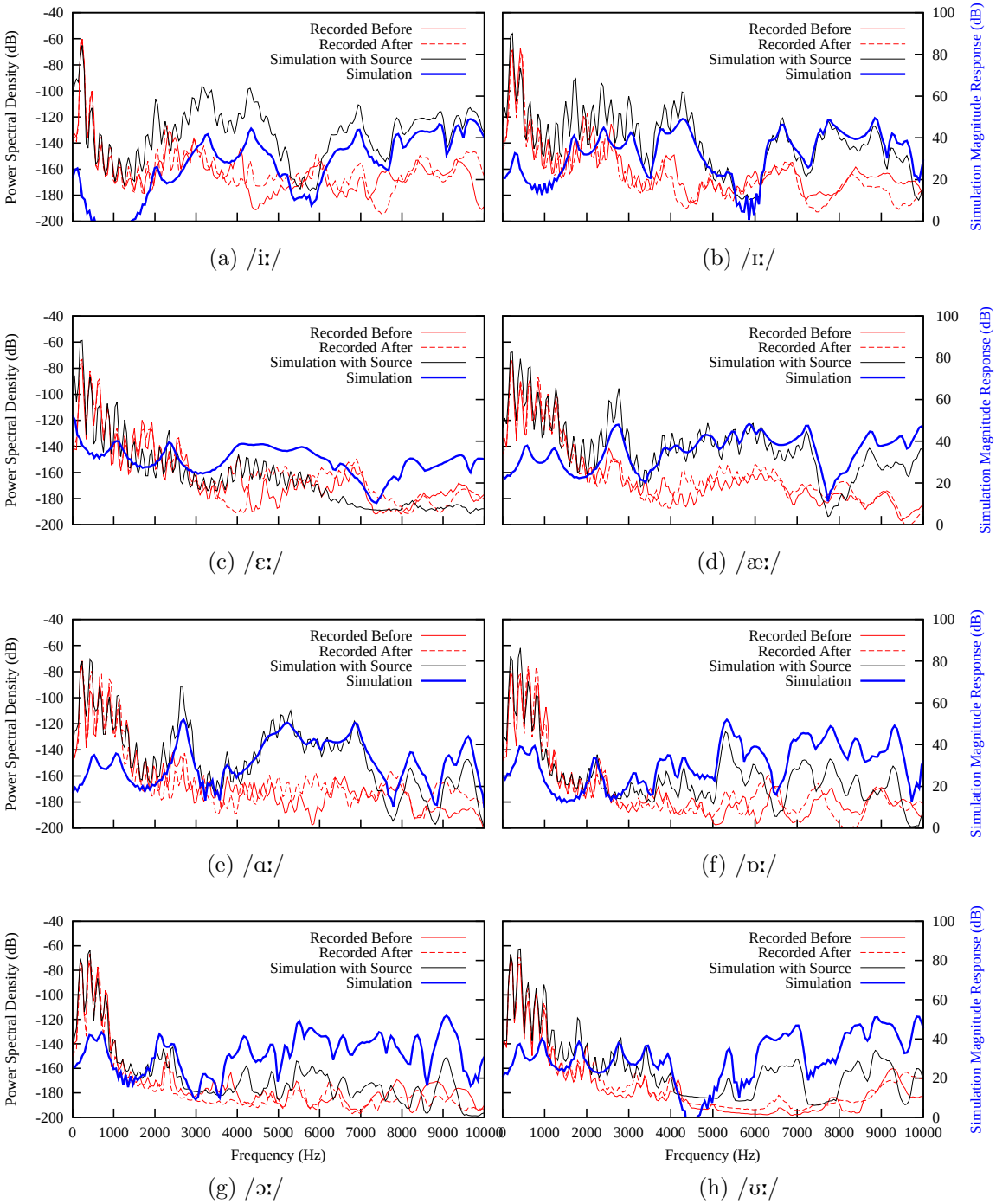
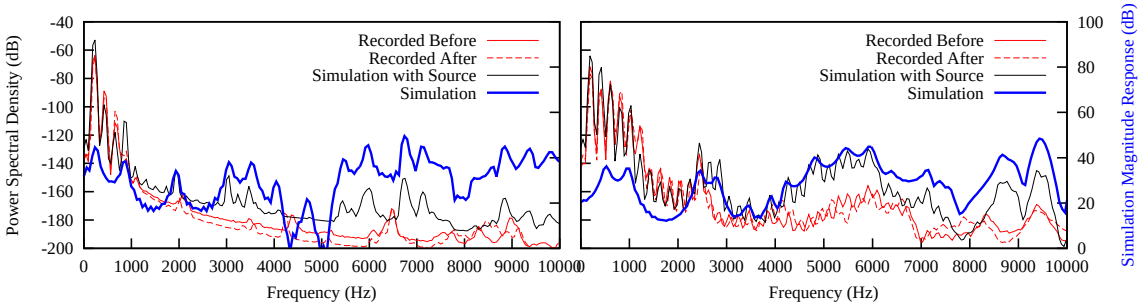
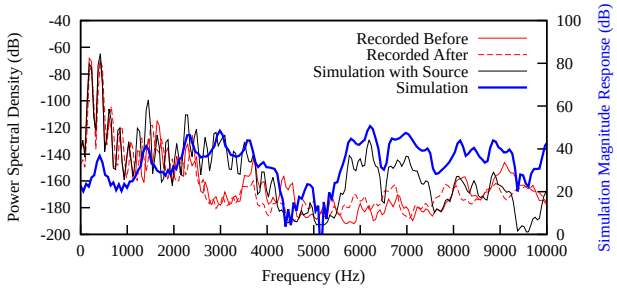


Figure 8.3: Jill - Female - Comparison of three-dimensional simulations for each vocal tract model with recorded vowel power spectral densities



(a) /uː/ (b) /ʌː/



(c) /ɜː/

Figure 8.4: Jill - Female - Comparison of three-dimensional simulations for each vocal tract model with recorded vowel power spectral densities

8.1.3 Jasmine

Jasmine is a professional, classically trained singer with a mezzo-soprano range. She chose to phonate at 262Hz, corresponding to the musical note C4.

A number of Jasmine's scans were not be included due to significant motion artefacting (inducing blurring in the imaging). The remaining phones demonstrate very strong performances, as shown in Fig. 8.5. All of Jasmine's recorded power spectral densities exhibit strong consistency between successive phonations, a fact encouraging further confidence in the simulations (as explored in section 7.5). Disregarding very slight formant shifts, all target phones can be considered accurate up to approximately 6kHz and some as high as 8kHz (/æː/ for example). The one slight exception is possibly /uː/, in which a resonance in simulation clashes with a trough in the benchmark audio. It is interesting that this error closely matches that of the /ɜː/ simulation of Jill, for which a frequency notch attributed to slight nasal coupling was considered. As with many of the subjects and phones, a number of the simulated responses exhibit excessive energy at high frequencies (particularly in /ɒː/ for example). It is still encouraging to observe strongly coherent trends between acoustic and simulated vowel qualities.

Audio examples for the recorded and resynthesised cases are provided in Appendix D. Similarly to Jill, the resynthesised audio examples produced for Jasmine suffer from the model's absolute rigidity. Both the lack of onset behaviours and complete formant stability contribute to a noticeable difference between synthesised and recorded equivalents. The overall phonetic content in each vowel is fairly consistent, resulting in resynthesised examples which typically match the target phone. As with Jill and Jack, whilst this phonetic content is satisfactory, the overall vowel timbre of the resynthesised examples is noticeably different.

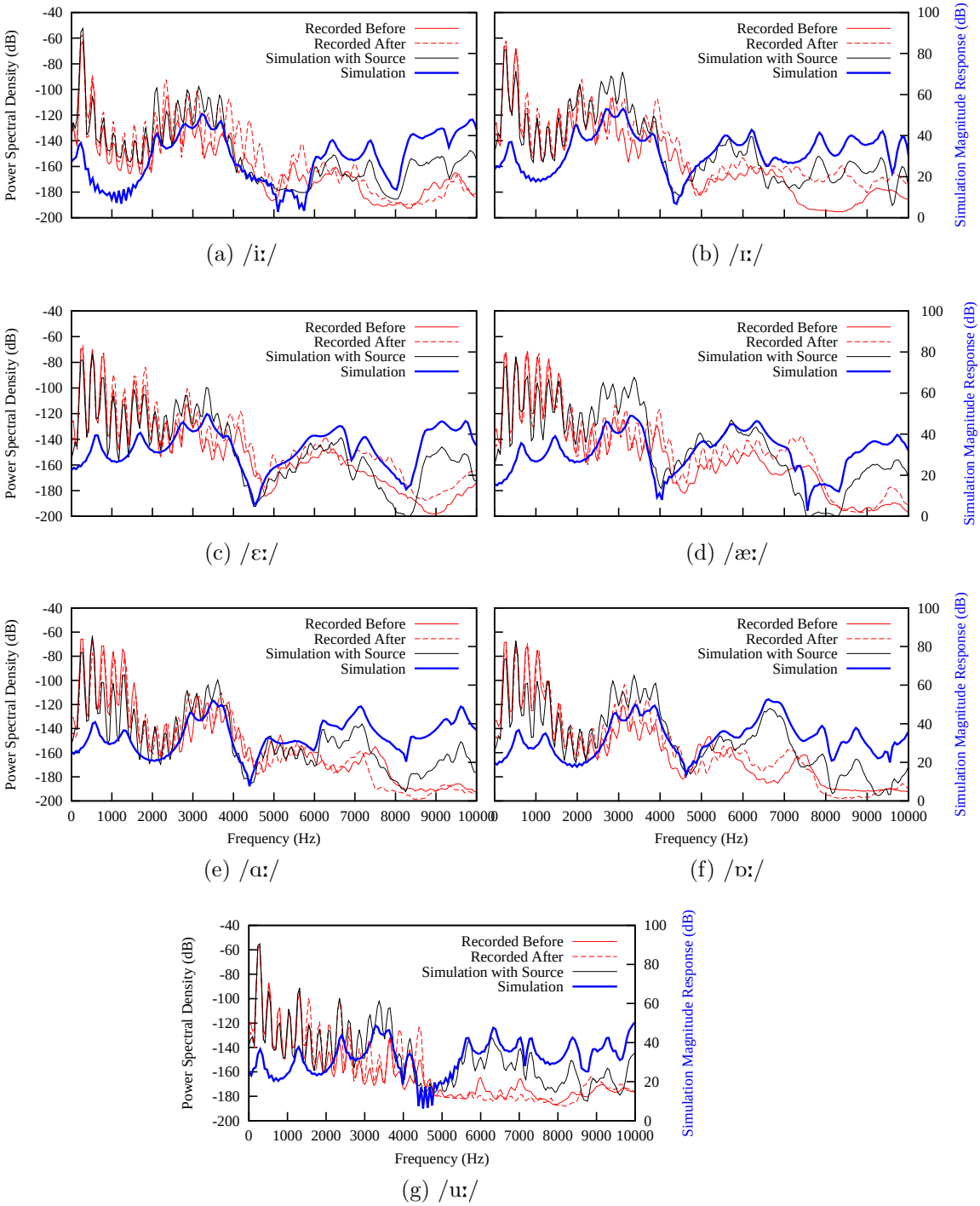


Figure 8.5: Jasmine - Female Mezzo-Soprano - Comparison of three-dimensional simulations for each vocal tract model with recorded vowel power spectral densities

8.1.4 Jim

Jim is a classically-trained (amateur) singer of tenor range. He chose a target pitch of 220Hz, corresponding to the musical note A3. The results for resynthesis of his target vowels are shown in Fig. 8.7, excluding those for /ɛː/ and /æː/ due to motion artefacting.

Jim demonstrates reasonable consistency between successive supine phonations, with some (expected) variations at higher frequencies (as in the case of /ɒː/). The simulated responses are generally accurate up to 5kHz, while exhibiting some slight shifts in F3 and F4. As with previous subjects, higher frequency behaviour follows the global trends of the recorded PSDs but with excessive high frequency energy in some cases (such as /ɒː/ and to a lesser extent /ɔː/ and /ʌː/).

Some of Jim's synthesised vowels are particularly accurate - /ɪː/, /ʊː/, /ɜː/ and /ʌː/ for example providing strong matches right up to 10kHz. Although each exhibits fair resonant shifts (typically lowering of F3 and F4), the global trends again closely match those of the recorded benchmarks. The strength of these vowels is further confirmed by audio comparison, as provided in Appendix D. The resynthesised vowels most closely matching the recorded equivalents are those identified as having particularly accurate frequency responses (/ɪː/, /ʌː/, /ɜː/ for example). It is worth noting that there is little distinction between /ʊː/ and /uː/ for both resynthesised and recorded cases, perhaps reflecting a lack of phonetic training.

Even the better examples of simulation for Jim exhibit the slight changes to vowel timbre identified for previous subjects. For Jim (and indeed Jasmine), these changes intuitively suggest constriction. This feature was also identified for the results produced for Jack, where they were attributed to basic boundary formulations. It is also conceivable that during segmentation of the MR imaging the actual cross-sectional area of the vocal tract is erroneously reduced. Post-processing of the MR imaging focusses on delimiting the vocal tract from the surrounding anatomy, including processing of the image to provide a steep contrast gradient. The scan protocol itself targets delineation of soft, wet tissue from air and since the surfaces of the vocal tract

are often covered with saliva it is quite conceivable that this layer contributes to effective narrowing in the resulting imaging.

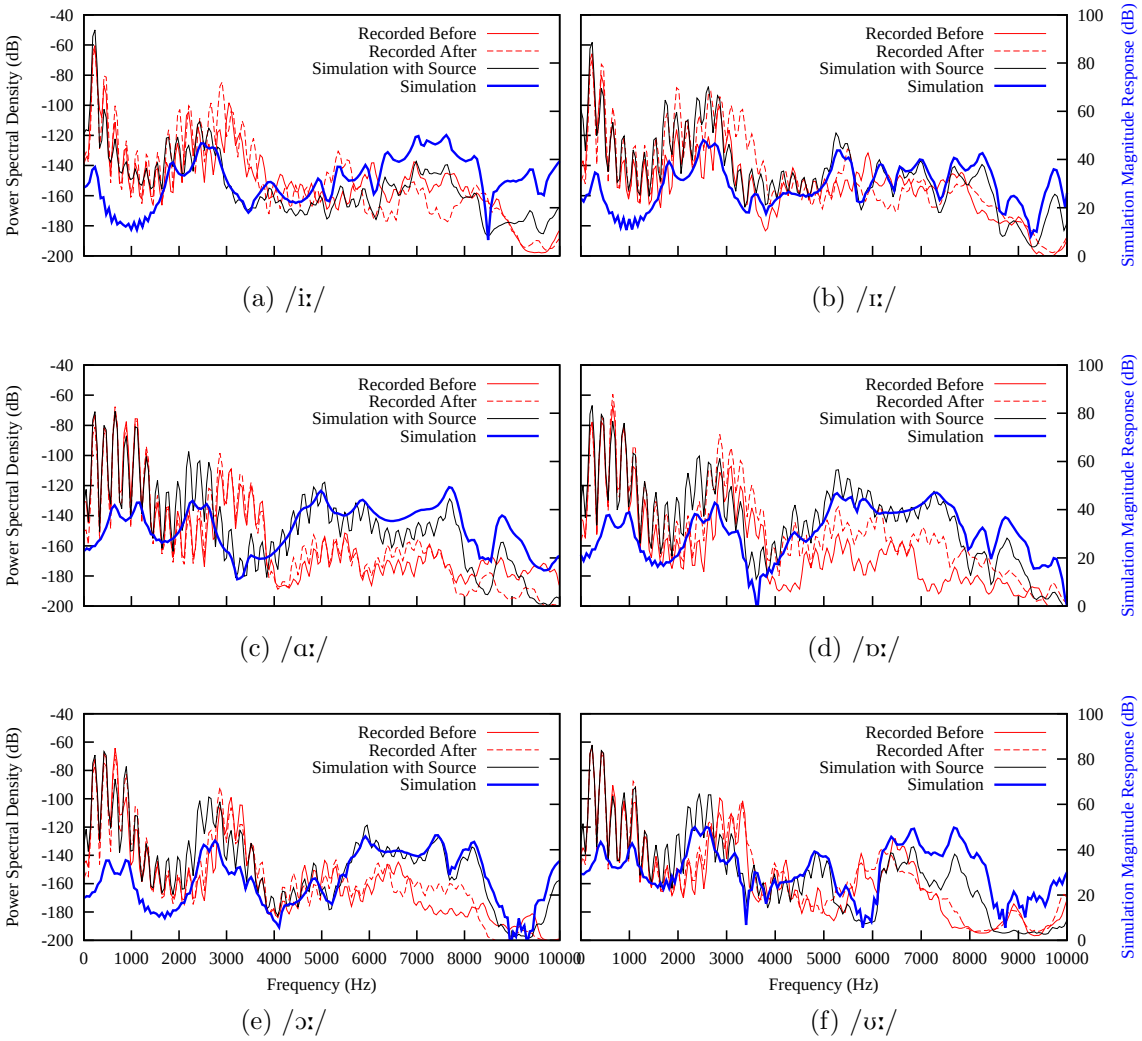
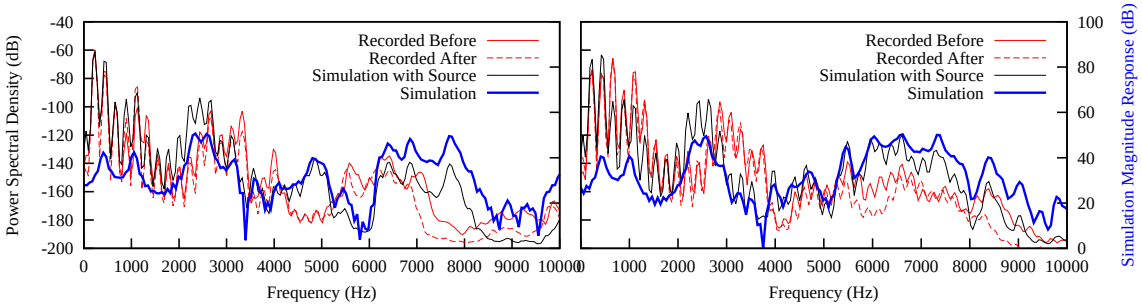
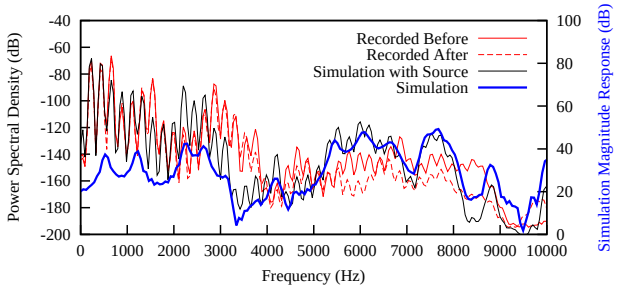


Figure 8.6: Jim - Male tenor - Comparison of three-dimensional simulations for each vocal tract model with recorded vowel power spectral densities



(a) /uː/ (b) /ʌː/



(c) /ɜː/

Figure 8.7: Jim - Male tenor - Comparison of three-dimensional simulations for each vocal tract model with recorded vowel power spectral densities

8.1.5 Jeff

Jeff is a professional classically trained opera singer with a bass/baritone range. He chose a target pitch of 110Hz, corresponding to the note A2.

While the simulation of Jeff's standard target vowels exhibits very slight lowering of F3 an F4 (as observed in all subjects), the simulations are reasonably accurate, as demonstrated in Figs. 8.8 and 8.9. The most impressive is perhaps /ɪː/, which, disregarding slight issues at 2.8kHz and 5kHz is accurate up to 10kHz. Other vowel configurations offer a similarly good performance, although many exhibit severe resonance shifts (such as /æː/). The simulations also demonstrate the increased energy at higher frequencies found in other subjects. With these resonance shifts and high frequency levels in mind, the responses of /æː/, /ɑː/ and /ɒː/ (for example) appear more accurate, although the impact of such frequency domain errors upon the vowel quality cannot be underestimated.

Audio comparison of the recorded vowels and simulated equivalents is provided in Appendix D. The overall correlation between these is remarkable. All provide acceptable matches in phonetic quality and (in contrast to other subjects) in most cases vowel timbre is very well matched. This is particularly interesting considering that a number of the simulated vowels do not match exactly the measured PSDs above 5kHz. It is worth mentioning that the MR imaging collected for Jeff featured extremely sharp boundaries and very little motion artefacting. This has inevitably contributed to the ease and accuracy with which the vocal tract is segmented.

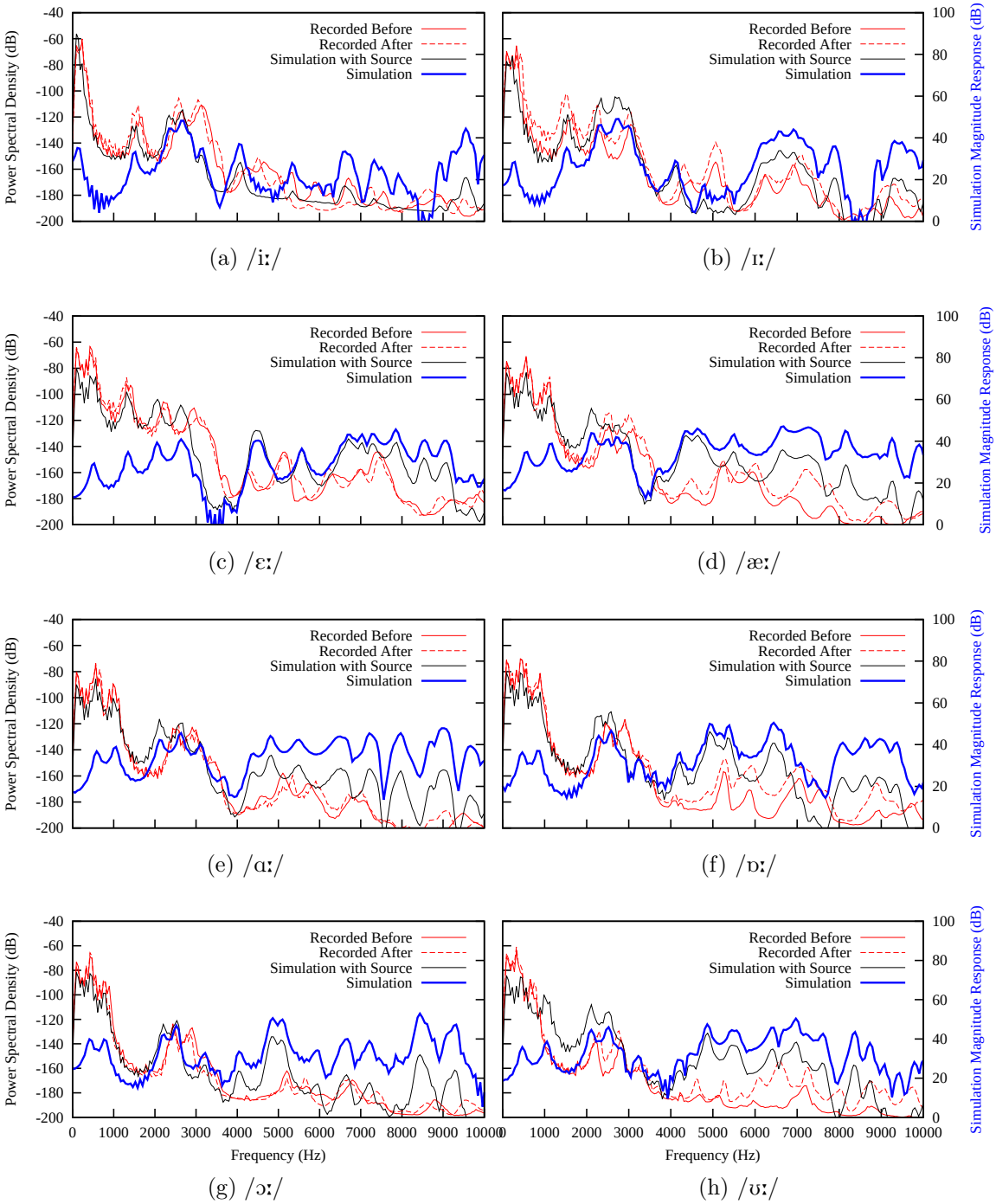
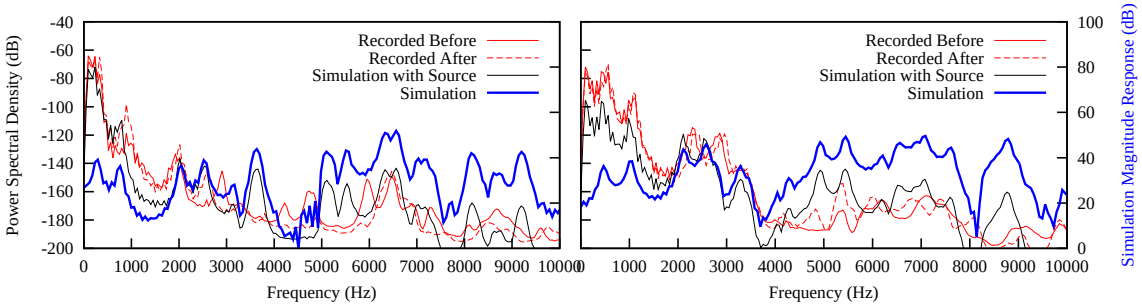
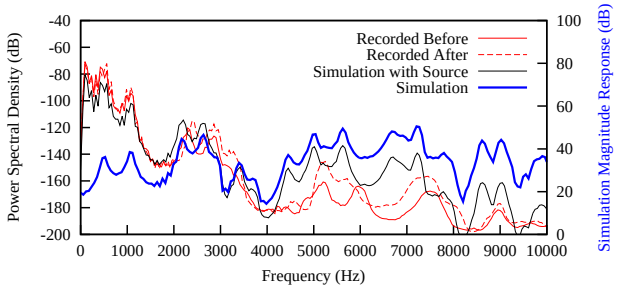


Figure 8.8: Jeff - Male tenor - Comparison of three-dimensional simulations for each vocal tract model with recorded vowel power spectral densities



(a) /uː/ (b) /ʌː/



(c) /ɜː/

Figure 8.9: Jeff - Male tenor - Comparison of three-dimensional simulations for each vocal tract model with recorded vowel power spectral densities

8.2 Derivative Simulations

While this project concerns the application of the three-dimensional digital waveguide mesh to vocal tract modelling, assessment of the hypothesis (see section 1.2) demands consideration of comparable lower dimensionality approaches. To this end a range of digital waveguide-based methodologies are now considered, drawing from the history of its application to vocal tract modelling.

Section 8.2.1 considers the development of cross-sectional area functions from the MR imaging and goes on to apply the traditional one-dimensional Kelly-Lochbaum methodology to the data. Section 8.2.2 continues to apply a two-dimensional digital waveguide mesh to the midsagittal plane of the vocal tract models. Finally, section 8.2.3 considers the application of Mullen's impedance-mapped vocal tract model to the developed cross-sectional area functions.

8.2.1 One-Dimensional Simulation

Traditionally, one-dimensional geometrical models of the vocal tract were developed by extrapolation of cross-sectional area functions from planar, midsaggittal X-Ray imaging. The acquisition of three-dimensional MR imaging allows for the development of more accurate cross-sectional area functions, based on measurement of the actual cross-sectional area. One-dimensional simulation of these area functions then provides an appropriate baseline for comparison of the full three-dimensional models.

Area functions are obtained by iterative bisection [117]. A line is drawn between the source and receiver coordinates used in three-dimensional simulation, and the normal plane computed about its centre-point, as demonstrated in Figs. 8.10 and 8.11.

The area and centroid of the intersection of this plane with the vocal tract model is found and the bisection point moved from the centre-point of the initial line to the calculated centroid. This procedure is performed recursively, resulting in a complete cross-sectional area function and a line drawing the area-wise centreline of the vocal tract. The final dissected model

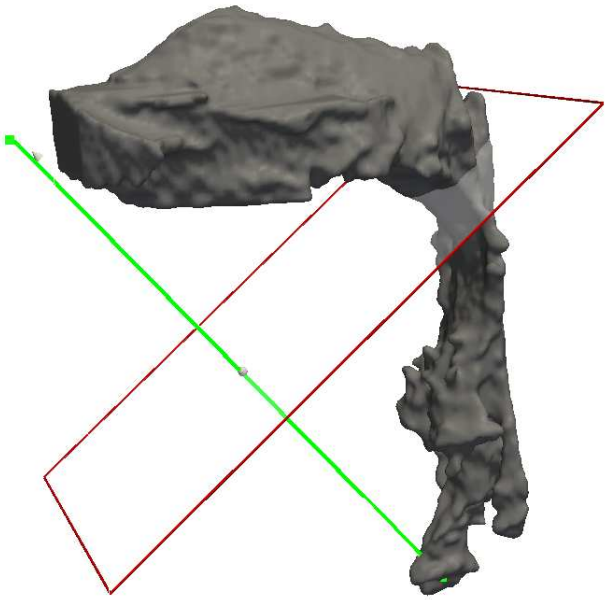


Figure 8.10: First stage of iterative-bisection of /ɑː/ model - Green line connects source and receiver points, red line delimits bisecting normal plane

and resulting centreline is shown in Fig. 8.12, with a cylindrical interpretation of the measured cross-sectional area function.

The implementation is based on the Kelly-Lochbaum model as described in section 5.6.2. Simple chains of digital waveguides are connected by one-dimensional scattering junctions. The reflective coefficients of these junctions are determined by the step in characteristic acoustic impedance, as a function of cross-sectional area changes between each pair of cylinders. In this case the radiating dome is replaced with a pressure phase-inverting reflection coefficient of $r = -0.9$ at the lips. The glottis is terminated in a phase preserving reflection coefficient of $r = 1.0$. A loss factor of 0.999 is introduced for each cylinder, found to provide appropriate formant bandwidths without a more accurate physical correlation. Simulations are run at 384kHz, corresponding to a waveguide length of 0.88mm. The model in Fig. 8.12 features 14 sections, corresponding to 4 iterative bisections. Each 5000-sample impulse is zero-padded and transformed by a 16384 point FFT. The magnitude responses of each simulation are plotted alongside the three-dimensional simulations and measured PSDs in Figs. 8.13 through 8.17.

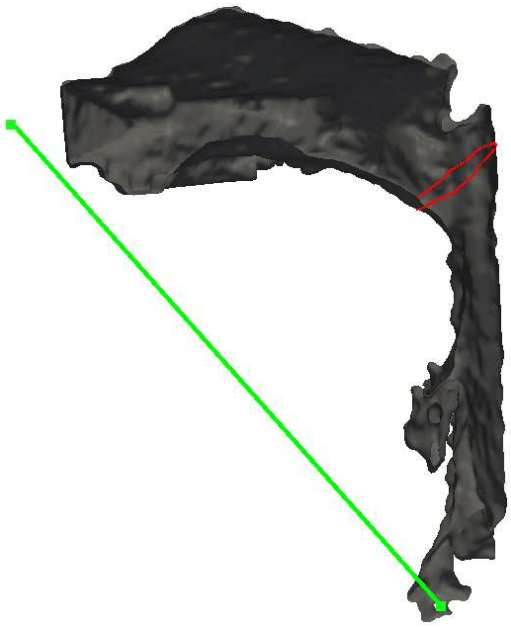


Figure 8.11: Halved /ɑː/ model after first bisection, showing extracted cross-section in red

The one-dimensional simulations perform well with regard to placement of the first 3 formants. Fig. 8.14a for example exhibits good accuracy in F1, F2 and F3 for the one-dimensional simulation. F1 could be considered slightly low in frequency, however this is consistent with the three-dimensional simulation and might simply illuminate spectral characteristics not exhibited by the fixed-fundamental frequency in the measured PSD. It is curious to observe variability in performance of the one-dimensional simulations between subjects. Fig. 8.13 demonstrates good matching of the first three formants for each of the three target vowels for Jack. Meanwhile, Jim demonstrates a very poor performance for one-dimensional simulation of /iː/ in Fig. 8.16a but a strong F1-F3 performance for /ɜː/ in Fig. 8.16b.

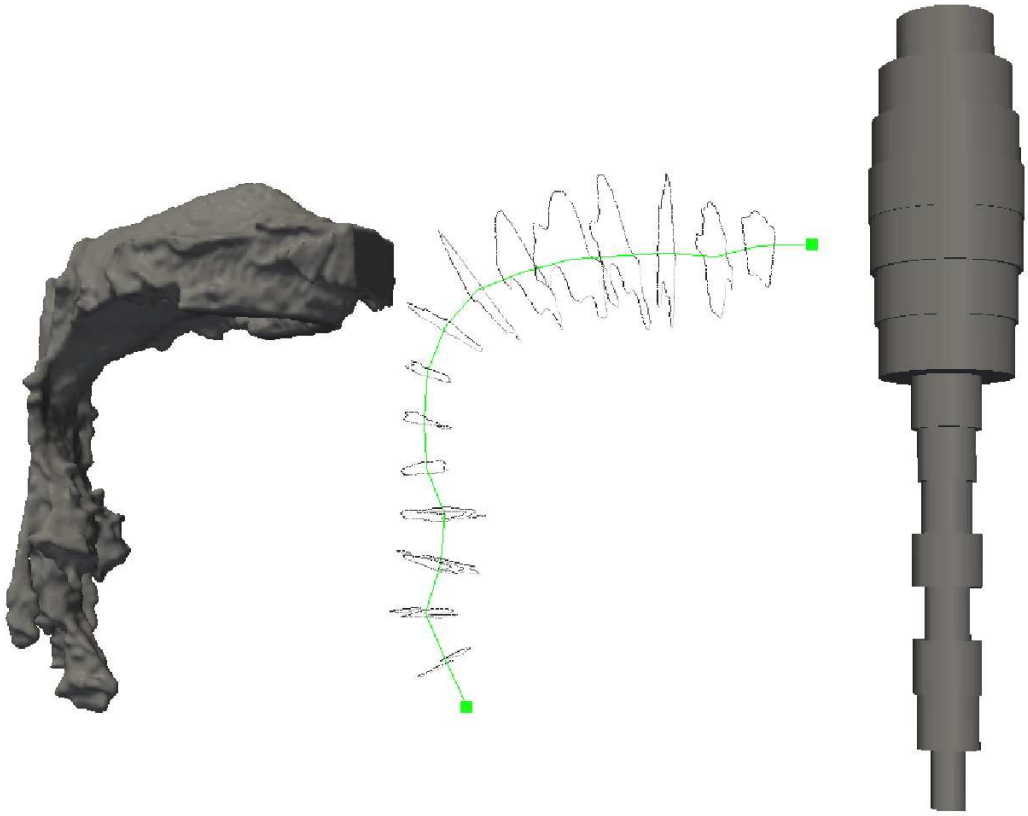
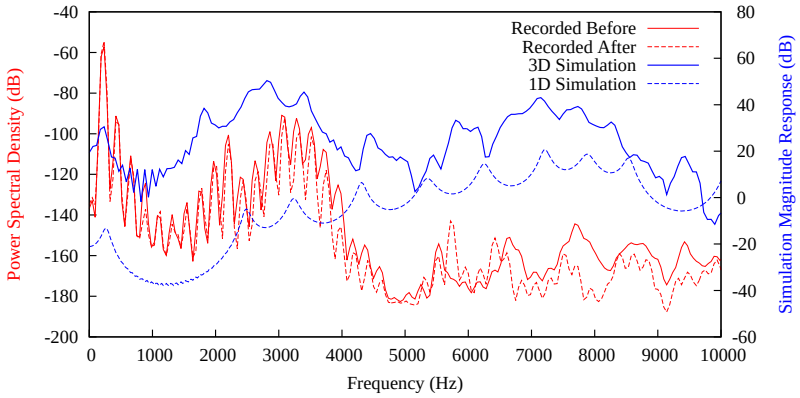
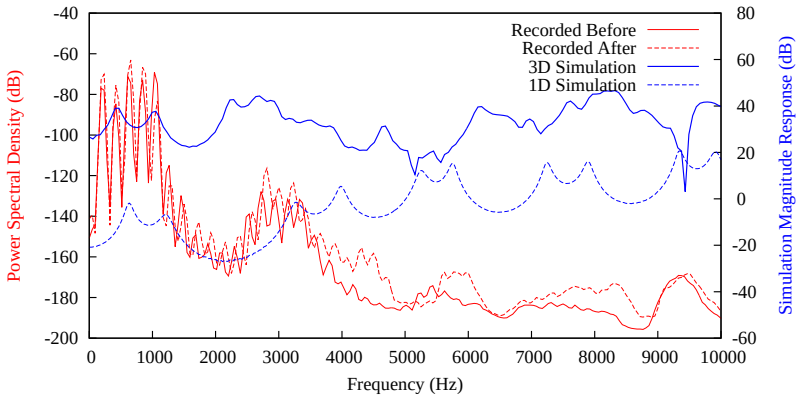


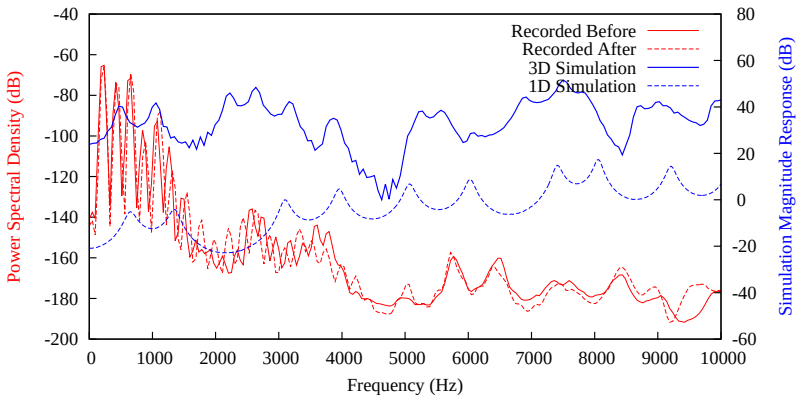
Figure 8.12: (Left) - Vocal tract model for /ɑː/ - (Centre) - 14 stage cross-sectional decomposition of the model with centreline in green - (Right) - Corresponding cylindrical analogue



(a) /iː/

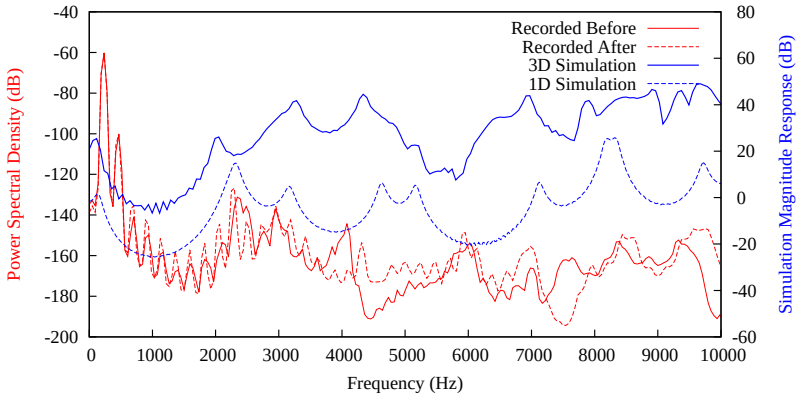


(b) /ɑː/

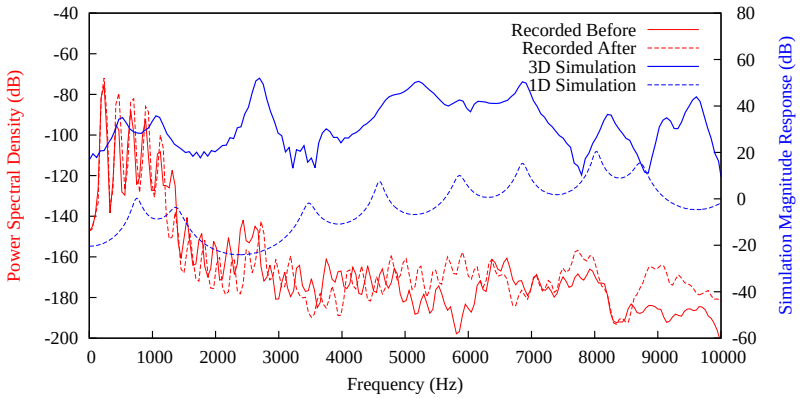


(c) /ɜː/

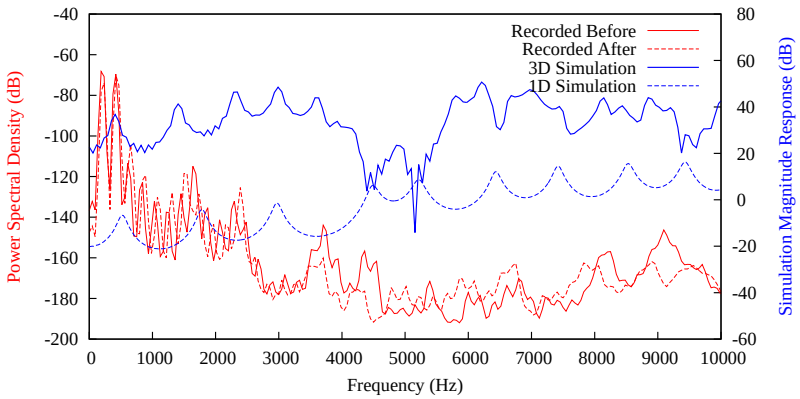
Figure 8.13: Jack - One-dimensional Kelly-Lochbaum model simulation using cross-sectional area functions derived from vocal tract models



(a) /iː/

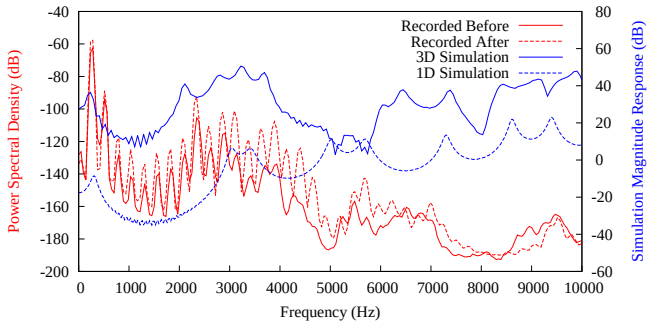


(b) /ɑː/

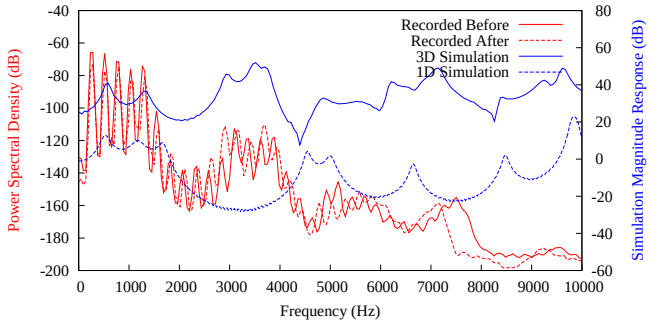


(c) /ɜː/

Figure 8.14: Jill - One-dimensional Kelly-Lochbaum model simulation using cross-sectional area functions derived from vocal tract models

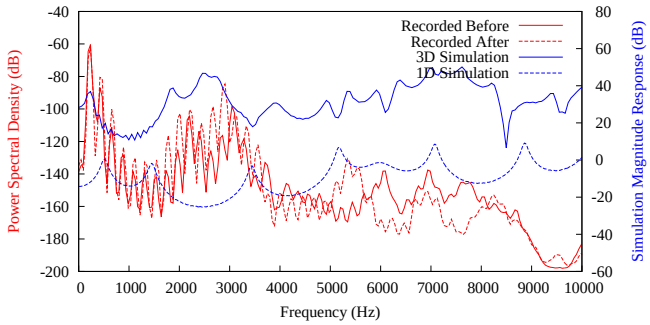


(a) /iː/

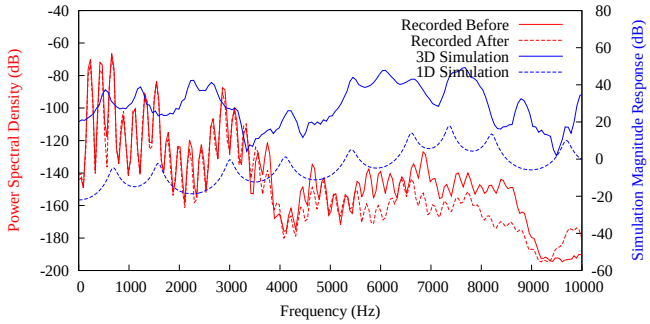


(b) /ɑː/

Figure 8.15: Jasmine - One-dimensional Kelly-Lochbaum model simulation using cross-sectional area functions derived from vocal tract models

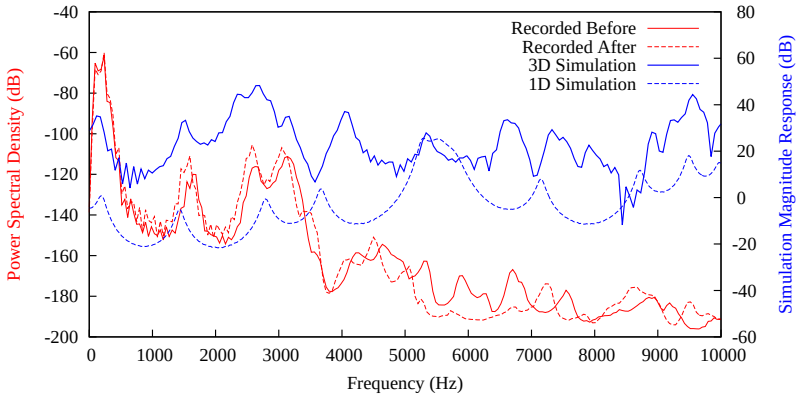


(a) /iː/

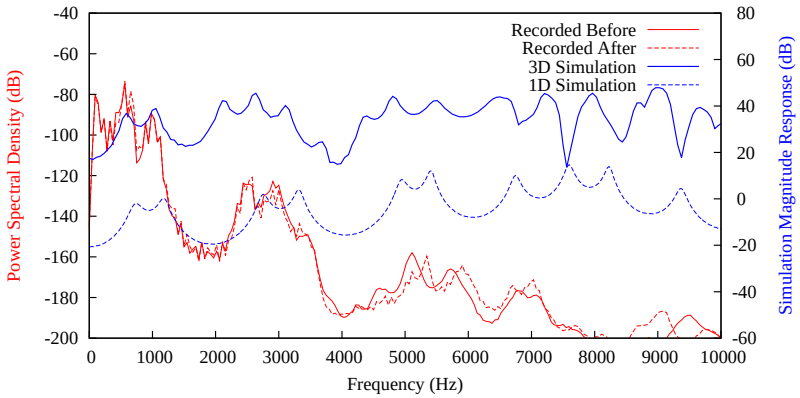


(b) /ɜː/

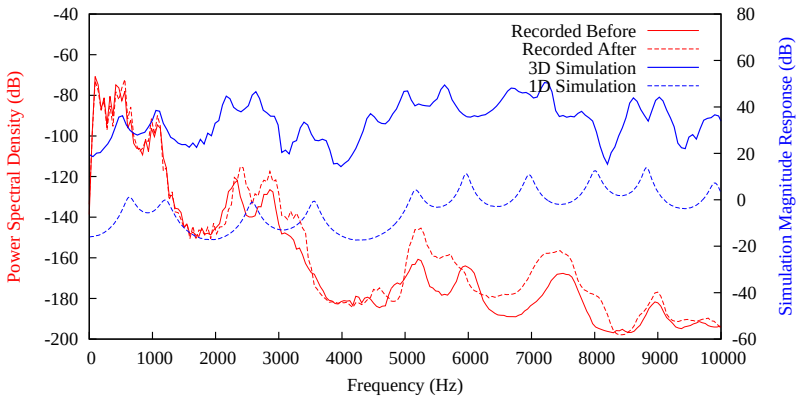
Figure 8.16: Jim - One-dimensional Kelly-Lochbaum model simulation using cross-sectional area functions derived from vocal tract models



(a) /iː/



(b) /ɑː/



(c) /ɜː/

Figure 8.17: Jeff - One-dimensional Kelly-Lochbaum model simulation using cross-sectional area functions derived from vocal tract models

8.2.2 Two-Dimensional Simulation

As explored in section 6.5, two-dimensional simulation introduces awkward ambiguities with regard to geometrical representation. Two-dimensional simulation can either imply impedance mapping of cross-sectional areas (as explored in section 5.8), or direct modelling of planes of a three-dimensional model. Simulation using impedance mapping is explored in section 8.2.3 and direct planar simulation is explored here.

Fig. 8.18 shows a midsagittal slice of a three-dimensional sampling grid developed for Jeff's phonation on /ɑː/. Just as the three-dimensional simulation is performed, a derivative simulation can be performed on this two-dimensional grid by triggering different update and shuffle functions in the implementation. The shortcomings of two-dimensional simulation are however exacerbated in this case. The resulting grid does not bear the cross-sectional relationships required for accurate reproduction of one-dimensional behaviours, hence formant patterns are not reliably reproduced. As with impedance-mapped grids it could be argued that acoustic behaviour is reproduced in a single plane, however without the interactions and overall behaviour of the complete vocal tract (and indeed an associated formant pattern) the value of these simulations is negligible.

The same sampling grids are used as for three-dimensional simulation, for consistency in the spatial sampling error experienced. This spatial sampling interval corresponds to a system sampling frequency of 784.840kHz in two-dimensions by (3.22) (in section 3.2). Source and receiver positions are snapped to the nearest nodes to their positions in three-dimensional simulation. The injective in this case is a sinc function, with zero crossings corresponding to a cutoff frequency of 20kHz and support of 8001 samples. The simulation is run for 30000 timesteps, and the resulting impulse response transformed by a 65536-sample FFT.

Magnitude responses for several phones produced by Jeff are shown in Fig. 8.19, plotted against the measured PSDs and three-dimensional simulation.

It is interesting (and perhaps unsurprising) to observe a lack of coherence

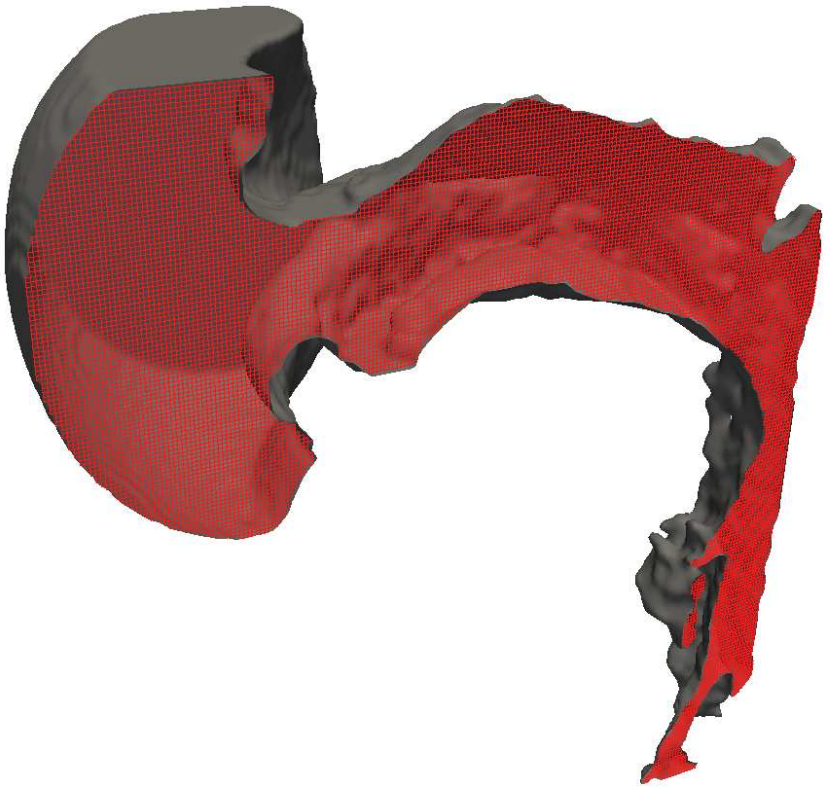


Figure 8.18: Mid-sagittal two-dimensional widthwise sampling grid for Jeff's Phonation on /ɑː/, at 783.840kHz, shown with its original segmented surface model

between the two-dimensional simulations and measured/three-dimensional simulated responses. Indeed, section 8.2.1 demonstrates a more accurate reproduction of behaviours by the one-dimensional model. Accurate matching of the first two formants can be observed in Figs. 8.19b and 8.19c for mid-sagittal simulation of /ɪː/ and /ʌː/, although simulation of /ɑː/ in Fig. 8.19a shows questionable representation of the first formant alone.

Beyond the first two resonant modes, single-plane two-dimensional simulation is seen to bear little resemblance to the actual analogue response. High frequency behaviour appears almost random and neglects to form any kind of reasonable characterisation of each vowel.

8.2.3 Impedance-Mapped Two-Dimensional Simulation

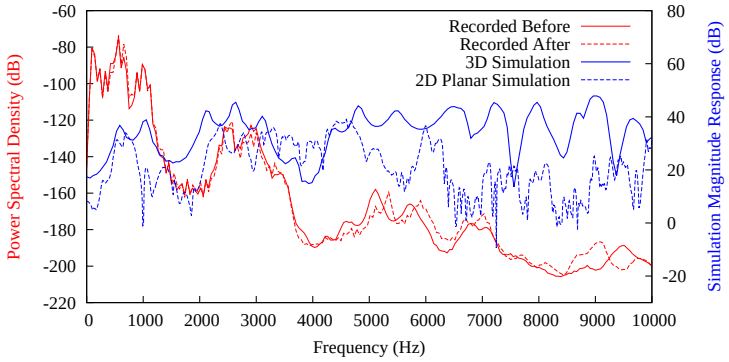
The dynamic impedance-mapped vocal tract model was introduced in section 5.8. This provides a novel means of real-time articulation of vocal tract modelling. It also presented a new paradigm of multi-dimensional representation of the vocal tract, whereby the simple two-dimensional plane (and its inherent shortcomings) are replaced by cosine-bell mapped impedance contours. These contours map to cylindrical radii or cross-sectional areas, neatly sidestepping the lost-volume issue in widthwise (planar) two-dimensional models.

VocalTract (shown in Fig. 5.10) provides Mullen's implementation of this system, based on C++ using an MFC user interface. This allows user manipulation of the vocal tract cross-sectional area function, enabling real-time changes to the resulting vocal tract transfer function. A number of base cross-sectional area functions are provided, allowing the user to switch between preset vowel models. These models are loaded from separate files at runtime. By replacing the contents of these files it is possible to substitute original area functions for the presets, in this case allowing area functions obtained from one-dimensional decomposition of MR imaging to be simulated using the two-dimensional impedance mapped methodology. This provides a further benchmark for comparison of the full three-dimensional simulations.

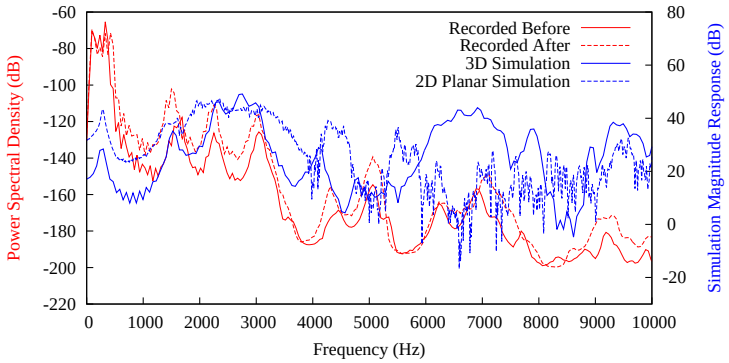
The precompiled VocalTract software (available online [118]) is run on a Linux system using the WINE compatibility layer [119]. Preset area functions are replaced in turn with area functions obtained by iterative bisection (see section 8.2.1). Simulations were run at a sampling rate of 176.4kHz (the highest rate offered by VocalTract) yielding an effective waveguide length of 2.75mm. Impedance contours are based on cylindrical radii (giving a geometrically correct analogue - see section 5.8). The LF-based source file is replaced by a simple dirac impulse, producing a periodic impulse response when the software is running. Each consequent impulse response is windowed to 5000 samples and transformed by means of a 16384 point FFT. The results are shown in Figs. 8.20 through 8.24.

At low frequencies, the similarities between two-dimensional impedance-

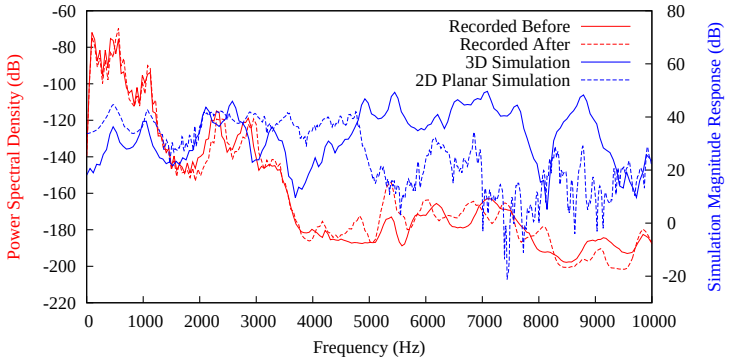
mapped simulation and three-dimensional simulation are often striking. The first two formants are correctly reproduced, occasionally providing slight frequency shifts in F2 (as is the case for 8.24b, 8.22a and 8.20b). For the simulations of /ɜː/ in Figs. 8.20b and 8.21b it is extremely interesting to observe a more accurate representation of the measured PSDs by the two-dimensional case than in the full three-dimensional case. It is unsurprising at least that the impedance-mapping approach works best for a nearly-neutral condition, since this condition least illuminates the ambiguities in contour mapping a three-dimensional volume to a single plane.



(a) /ɑː/

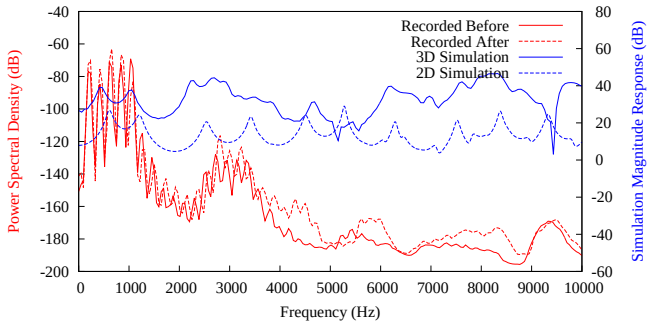


(b) /ɪː/

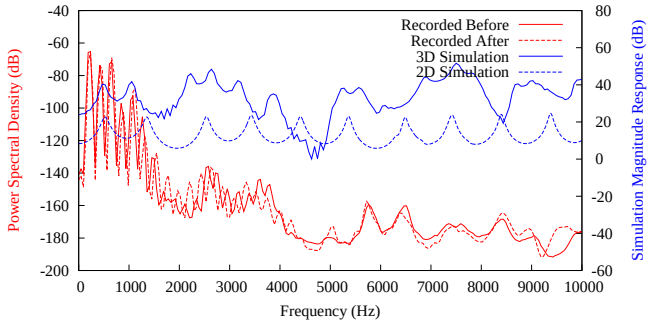


(c) /ʌː/

Figure 8.19: Jeff - Bass - Two-dimensional widthwise simulations compared with three-dimensional simulation and measured power spectral densities

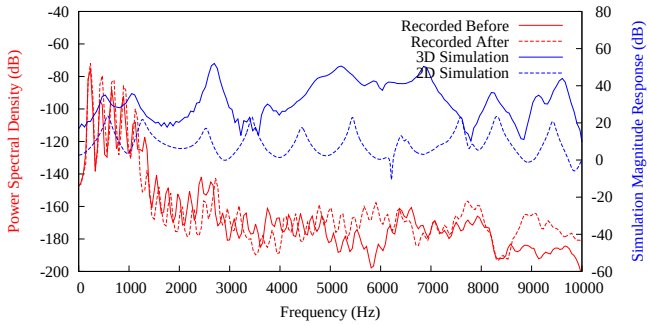


(a) /ɑː/

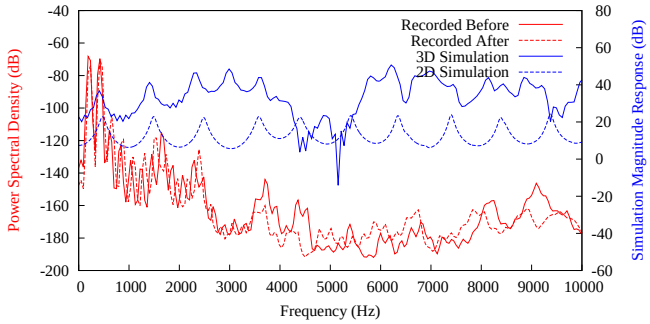


(b) /ɜː/

Figure 8.20: Jack - Two-dimensional impedance mapped simulation using cross-sectional area functions derived from vocal tract models

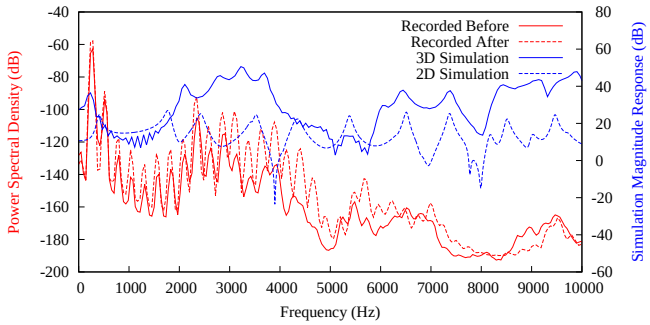


(a) /ɑː/

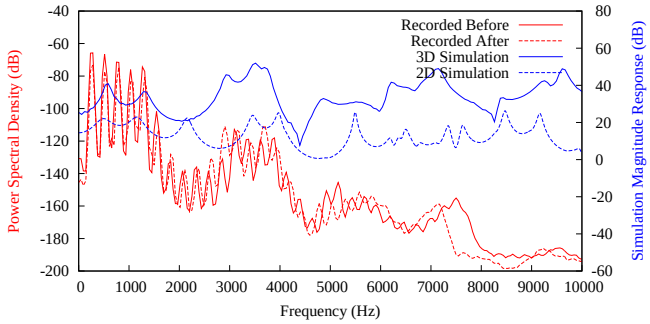


(b) /ɜː/

Figure 8.21: Jill - Two-dimensional impedance mapped simulation using cross-sectional area functions derived from vocal tract models

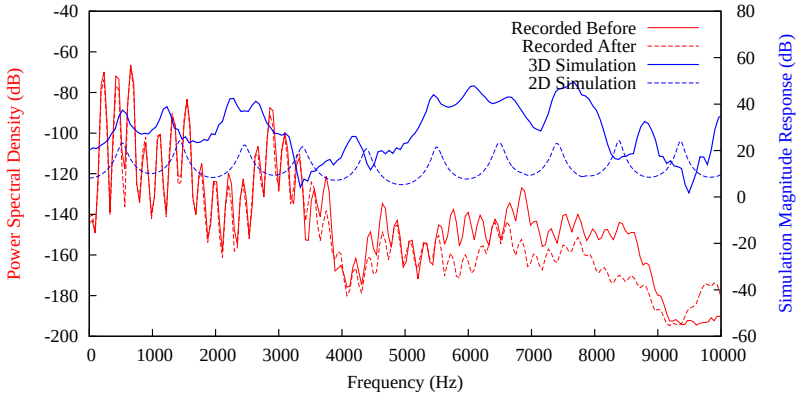


(a) /iː/



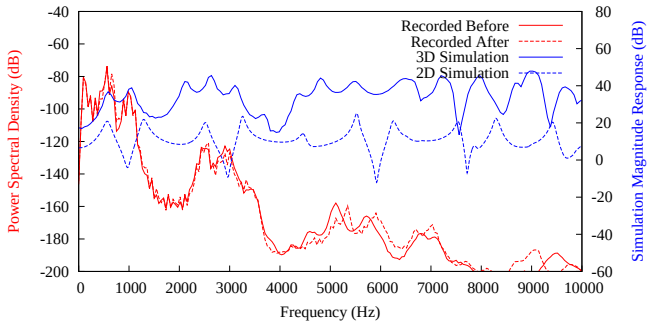
(b) /ɑː/

Figure 8.22: Jasmine - Two-dimensional impedance mapped simulation using cross-sectional area functions derived from vocal tract models

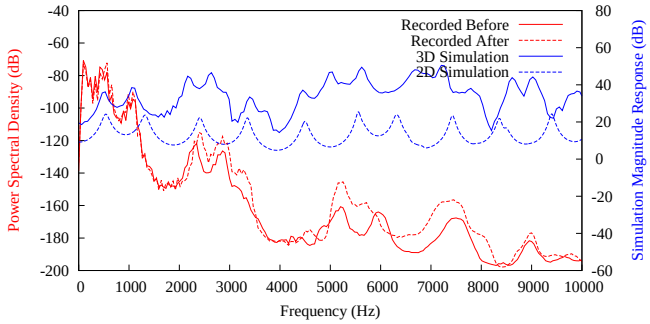


(a) /ɜː/

Figure 8.23: Jim - Two-dimensional impedance mapped simulation using cross-sectional area functions derived from vocal tract models



(a) /ɑː/



(b) /ɜː/

Figure 8.24: Jeff - Two-dimensional impedance mapped simulation using cross-sectional area functions derived from vocal tract models

8.3 The Source

The naturalness of the resynthesised examples here is inescapably linked to the source content. Regardless of the accuracy of the vocal tract transfer function, convolution by an unnatural source function will inevitably lead to unnatural voice synthesis. It is important to remember that the electroglottographic waveform used as a source in section 8.1 is not accurately representative of supra-glottal sound pressure variation (as explored in section EGG). It does however include desirable spectral characteristics and most importantly the stochastic variability in period, amplitude and spectral tilt intrinsic to real human phonation.

To decouple these characteristics of the source from the VTTF, each of the impulse responses is convolved with a simple cyclic LF-model waveform (see section 5.3 for definition of the Liljencrantz-Fant model). The source waveform is tailored to the target pitch for each subject, and given short linear ramps at onset and offset to provide a more typical source behaviour. The waveform is created at 192kHz and the impulse response downsampled to match this, before resampling of the convolved output to 48kHz for playback. The results are presented in Appendix D. On listening it is clear that when convolved with the strongly deterministic LF source waveform the resulting naturalness is diminished.

To explore further the nature of source characteristics imparted upon resulting voice synthesis, the electroglottographic waveforms recorded for each subject were swapped. For continuity in vocal range, the source waveforms for each of the tenors were swapped. Jack's and Jim's source recordings were hence swapped before convolution with the VTTFs (tenor ranges) and Jasmine and Jill's contributions similarly. Audio examples are provided in Appendix D. Here, appropriate outputs are still observed, although identifiability is reduced. This suggests that it is the variability of a real source waveform that is of greater importance to perceived naturalness than any physical correlation of the source and vocal tract. It should of course be remembered that during phonation vocal source coupling is considered to have an important role in phonation, although in this case the source and vocal

Static Protocol	
T_e	1.7ms
T_R	4.8ms
Flip Angle	5°
Bandwidth	±41.67Hz
FOV	$260mm^3$
Slice Width	2mm (50% separation)
Matrix	192 × 192

Table 8.1: MRI protocol developed for static vowel scanning

tract are considered linearly separable.

8.4 Corpus

In the course of this project a substantial body of data has been developed, both with the current research project and future studies in mind.

Data was collected for 6 subjects from strong singing or phonetic backgrounds. Of these, five are currently available. The MR imaging is of three types:

- Static Scans
- Structural Scans
- Dynamic Scans

Static scanning encompasses the standard procedure outlined in sections 6.1.2 and 6.2.2, the protocol for which is provided in Table 8.1. For subjects Jack, Jeff and Jim, structural scans were also carried out. These scans are recorded over a greater period of time and target delineation of the limits of the nasal tract. The protocol developed for these scans is given in Table 8.2.

Finally, dynamic scans were made for subjects Jack and John (yet to be released). These are recorded in a single midsagittal plane, demonstrating articulator movements for simple diphthongs. The protocol developed for this dynamic scanning is detailed in Table 8.3.

Structural Protocol	
T_e	3.1ms
T_R	8ms
Flip Angle	30°
Bandwidth	±31.25Hz
FOV	$260mm^3$
Slice Width	1.6mm (50% separation)
Matrix	256 × 256

Table 8.2: MRI protocol developed for structural scanning of the nasal tract

Dynamic Protocol	
T_e	1ms
T_R	3.2ms
Flip Angle	5°
Bandwidth	±41.67Hz
FOV	$220mm^3$
Slice Width	10mm (50% separation)
Matrix	160 × 96
Framerate	4.29Hz

Table 8.3: MRI protocol developed for dynamic midsagittal scanning of the vocal tract

All imaging is delivered in both DICOM and NifTI file formats. The resulting vocal tract models are delivered in multiple forms. Firstly, segmentation of each vowel model is delivered in Analyze file format, to allow it to be recombined with the original imaging and edited if desired. The segmentation surface mesh is saved separately as a VTK XML PolyData type. These models include the radiating lip dome (see section 6.3.1) for all processed models. Segmentation of vocal tract models is also provided for a number of models without this radiating dome. These are useful for the development of derivative models.

Cross-sectional area functions are calculated for each vocal tract model without the radiating dome. These are delivered as ASCII text files, with the radius for each cylinder provided in metres, tab-delimited from the length of the corresponding section. Each cylinder is carriage-return-delimited from the next. A three-dimensional cylindrical graphical model of each cylindrical arrangement is provided in VTK XML PolyData format. The vocal tract centreline developed in the course of iterative bisection is also provided and delivered in the VTK XML UnstructuredGrid file format.

Each static vowel model carries associated audio and electroglottographic data as outlined in section 6.1.1. This data can be used for benchmarking resulting synthesis strategies and for assessing the reliability of the assumed phonetic condition of the imaging. All recordings are made at 192kHz at 24-bit resolution. Audio and electroglottographic data are provided in both extracted PCM (.wav) and original formats (stereo tracks in Audacity project files).

It is hoped that the collected corpus will be of significant value to future voice research.

Summary

In this chapter, the results of three-dimensional simulation of the complete vocal tract models are presented and audio examples provided. Lower dimensionality derivative simulations are also produced for comparison. Direct two-dimensional midsagittal simulations are presented, followed by the closely

related dynamic impedance-mapped equivalents. Classical one-dimensional simulations are then provided as a benchmark. The influence of the source waveform on the audio examples has been considered. Finally, the collected corpus has been described.

Three-dimensional simulation is found to provide a strong match to the recorded vowels in terms of phonetic content. Overall voice quality is not equally well matched, although in most cases they sound similar. These perceptual findings support the presented magnitude responses in that lower frequency behaviours are reproduced whereas higher frequency behaviours undergo (in some cases significant) shifting or other modification.

Direct two-dimensional simulation is unpredictable, due to the limited geometrical analogue it provides. In some cases it is found to approximate early formant frequencies, but in other guises its behaviour is completely unreliable. Impedance-mapped two-dimensional simulation in contrast provides reliable results where the vocal tract is maximally neutral. Where significant discontinuities are experienced in the cross-sectional area function, the accuracy of the response declines with the further abstraction of the geometrical analogue.

One-dimensional simulation is interesting in the accuracy with which it can reproduce early formant frequencies. Its shortcomings are however well understood. By focussing on the dominant one-dimensional acoustic behaviour of the vocal tract, strong representation of early formant behaviour is nearly assured. A one-dimensional representation does not however allow for accurate (or even appropriate) reproduction of higher frequency behaviours. Consequently, the response of the one-dimensional model above the first two formants is typically inaccurate and simulated vowel quality suffers as a result.

Finally in section 8.3 it is noted that an accurate vocal tract model will not result in accurate voice resynthesis when convolved with an unrealistic source waveform. The source is observed to carry important cues towards the perception of naturalness. By contrast, an inaccurate vocal tract model will result in inaccurate, but realistic voice synthesis when combined with a real source waveform.

Chapter 9

Summary and Conclusions

In this thesis, the application of the three-dimensional digital waveguide mesh to the vocal tract has been explored. The fundamentals of acoustics and the human voice have been introduced, followed by exploration of techniques for time-domain numerical acoustics modelling and their application to the voice. Experimental protocols and processes have been developed for the output of MR imaging of the vocal tract in different articulatory settings and for the capture of appropriate benchmark audio recordings. The three-dimensional simulation techniques have undergone extensive validation, before their application to models of the vocal tract extracted from the MRI data. The results of simulation are compared to those of lower dimensionality derivative models. The outcomes are then compared by means of spectral plots and original audio recordings of each vowel phonation.

9.1 Summary of Results

Acoustic Measurement of Vocal Tract Analogues Using Banded Exponential Sine Sweeps

A technique has been developed for broadband acoustic measurement of cylindrical configurations and comparable vocal tract analogues. This is based on exponential sine sweep measurement, allowing an arbitrarily high signal-to-noise ratio and the removal of harmonic distortion. The experimental setup

is low cost, performing equalisation of a transducer mechanism by decomposition of the audible spectrum into bands, for which inverse-filtering can be performed. The technique allows for acoustic measurement of mechanical constructs which can be used in validation of numerical acoustics models. In the course of this thesis the measurement technique itself is validated against analytically determined and simulated frequency responses for cylindrical resonators of increasing geometrical complexity. The technique is seen to perform well across the sub-10kHz range. The system has two shortcomings of note. Firstly, it is unable to correct for zeros occurring in the transducer response. In this case, the technique would seek to expand the sweep length around the zero frequency to increase the injected energy. While this approach works for correction of troughs in the frequency response, increasing transfer magnitude at a zero by simple increase of the input is theoretically impossible. The second shortcoming is computational demand, although this is flexible. In the course of the thesis the process is run at maximum possible accuracy, using high sampling rates, iterative methods for determination of optimum band widths, long impulse responses and consequently long FFT windows. It is quite possible to run the system at lower specifications, reducing computational demand accordingly, although precision will likely suffer as a result.

Collection and Benchmarking MR Imaging of the Vocal Tract

A data set comprising MR imaging of a range of subjects from trained singing or phonetic backgrounds is presented. Scanning protocols are developed specifically targeting delineation of the vocal (and nasal) tract boundaries, whilst keeping capture times as short as possible. An experimental process has been developed incorporating the collection of benchmark audio recordings and electroglottographic waveforms both before and after the scanning procedures. The vocal tract has been segmented from a subset of the collected articulatory configurations and provided for each subject together with cross-sectional area functions and the associated vocal tract centre-line. A range of additional scans are recorded for further research, including vowels sung at

pitches spanning the vocal range, in different registers (head/chest/falsetto), high detail scans of the nasal tract and dynamic midsagittal scans of common diphthongs. The imaging augments existing data targetting the vocal tract in that it utilises the very latest MR imaging technologies to achieve a very high resolution capture in a short time. This is particularly crucial in imaging of the geometrically intricate vocal tract, which requires a significant level of skill to hold stable for any length of time.

Validating Application of the 3D DWM to Vocal Tract-like Structures

Application of the developed acoustic measurement technique to cylindrical structures and vocal tract analogues of increasing geometrical complexity has provided progressive validation of the applied numerical simulation schemes. It is demonstrated that for the most simple cuboid and cylindrical structures the three-dimensional digital waveguide mesh provides a close match to the measured acoustic behaviours across the sub-10kHz range. Overall behaviours from the analytically determined solutions are closely matched, although significant errors appear at higher frequencies. These are attributed to the limited mathematical representation of the analytical solution. The acoustically determined responses are found to exhibit consistent, shifted matches to their simulated equivalents, the reliability of which is more convincingly apparent in plots (eg. Fig. 7.9) than through the absolute frequency errors identified. For stepped concatenated arrangements a similarly strong performance is observed although for some particular geometrical arrangements the performance of simulation is poor. This is particularly apparent where glottis-end, longer and narrower cylinders are connected to lip-end cylinders of greater diameter. A further issue throughout simulation of the concatenated cylinders is slight frequency shifting of the third and to a lesser extent fourth resonant modes in the 2.5-4kHz range. This shifting simply induces slight errors between measured and analytically predicted resonant mode frequencies. Both these issues are potentially related to the representation of the complex radiating interface at the point of concatenation, which

determines the extent of formant separation.

Consequent application of the technique to complex vocal tract analogues reveals an often accurate performance. The first two resonant modes (formants in this context) are always accurately reproduced. From 4kHz up there is a very strong match between simulated and measured responses. As with some concatenated cylindrical arrangements the 2.5-4kHz range is a problem for certain simulations, leading to occasional shifts in F3 and F4. In the worst case, these shifts can cause the formation of a degenerate mode.

The methods employed here constitute a particularly basic implementation of current modelling technologies. While the mesh itself is created at a very high resolution (to provide as close a geometrical match as possible to the structure) the boundary formulations used are minimal, to reduce computational load. Despite this, the simulation produces results of remarkable accuracy. This suggests that even in such a basic implementation the methodology is suitable for representation of vocal tract geometries. It must of course be remembered that while the boundary formulations used have been shown to be effective for modelling reflections in a uniform acrylic structure, this may not also be the case for the varying complex surfaces of the vocal tract itself.

3D DWM Representation of Vocal Tract Acoustics

After rigorous validation, the three-dimensional DWM simulation technique is applied to the models of the vocal tract developed from MR imaging. In some cases, motion artefacting was so severe as to prevent segmentation of phones. Simulations were performed at 960kHz, for which computation of 10000-sample impulse responses was possible in approximately 30 minutes. Throughout this thesis, the simplicity of the implementation has been repeatedly underlined. Despite this, simulation appears to offer a good representation of the vocal tract transfer function. While the simulated impulse responses do not result in transfer functions exactly matching those recorded, it is emphasised that the recordings represent the complete voice production system, instead of the vocal tract transfer function alone. In all cases, the first

two formants are exactly reproduced. Following this, F3 and F4 placement is commonly accurate, although in a number of cases very slight frequency shifting is observed, consistent with that observed for simulation of the complex vocal tract analogues. At higher frequencies, behaviour is inconsistent. In some cases simulation provides a strong match to the recorded equivalents, while in others there is only a weak correlation.

Audio comparison of simulated and recorded equivalents is telling. While the trend is towards correct reproduction of phonetic quality, the overall vowel quality of simulation is not seen to exactly reproduce that of the recordings. This is perhaps unsurprising considering again that the vocal tract alone is represented in simulation.

An interesting audio artefact in resynthesis is introduced by the absolute rigidity of the model. A human phonation can be expected to produce small changes in resonant behaviour since, as observed elsewhere, holding the vocal tract completely stationary constitutes a significant challenge, even with training. In resynthesis these cues are not produced, with the simulated voice artificially producing perfectly stable formants for around 16 seconds. Similar cues towards variability in human phonation are observed in the source waveform. Vowel resynthesis using a simple, periodic source model results in a cyclic behaviour which the listener can quickly identify as unlikely, if not impossible in a human phonation. Use of an electroglottographic waveform circumvents this issue, as much of the inherent variation expected in human voice production is reproduced. This underlines the importance of the source waveform to accurate reproduction of human phonation in any physical model.

There are numerous potential sources of error in the vocal tract models, most of which are described in areas for further work. Beyond missing geometrical features, the most obvious shortcoming with respect to the simulation methodology alone is perhaps the limited representation of acoustic behaviours at boundaries. The boundary formulations used here are extremely limited, a particular concern being their frequency independence. This reflects a lack of physiologically accurate damping and also the assumption of uniform behaviours across all boundaries in the vocal tract tract model.

Acoustic behaviours in constrictions are very heavily damped due to flow effects in the boundary layer.

Derivative Lower Dimensionality Equivalent Digital Waveguide-based Models of the Vocal Tract

After development of the full three-dimensional models, derivative equivalents have been constructed using cross-sectional area functions obtained by iterative bisection. These models were one-dimensional Kelly-Lochbaum models, direct 2D midsagittal meshes and impedance-mapped 2D meshes. The relative performance of these models is assessed. Kelly-Lochbaum models are seen to provide accurate representation of early formant behaviours, with well-known shortcomings in the high frequency response due to representation of one-dimensional behaviours alone. Two-dimensional midsagittal simulation typically results in very poor frequency-domain representation due to the limited geometrical analogue. In the best cases the first two formants are accurately reproduced. Impedance mapping of a 2D DWM can result in accurate formant reproduction, although its performance is found to decrease in accuracy as discontinuities in characteristic acoustic impedance become more significant.

9.2 Hypothesis Revisited

Restatement of Hypothesis

> Time-domain acoustical simulation of the vocal tract using a three-dimensional digital waveguide mesh can result in more accurate reproduction of the voice than lower dimensionality equivalents.

Discussion

Treatment of this hypothesis necessarily begins with validation of the implementation. Development of a system able to simulate such complex geomet-

rical arrangements is not straightforward and to inspire trust in the results a level of confidence must be achieved. The system was hence applied to a range of models by way of validation. The earliest offered directly analytically determinable resonant behaviours (cuboid) for which 3D simulation offered an entirely accurate representation. The same was true for simple quarterwave resonators, for which the frequencies of simulated resonant modes matched those of calculation (to within known error margins). These results inspired confidence in the implementation, demonstrating that the complex stages in decomposition of a graphical model into an efficient data structure are effective.

Assured that the software developed constitutes a solid implementation of the theory, attention was turned towards assessing the applicability of the methodology to reproduction of the intricate geometrical configurations of the vocal tract. This stage in the validation was approached through development of an acoustic measurement system, against which the results of simulation can be checked when the structures under test become too complex for analytical derivation. The measured technique was found to be effective and was consequently used for comparison with the simulation of vocal tract analogues composed of concatenated cylinders. Simulation of these concatenated cylindrical resonators produced results that matched expectation in most cases. Slight errors (with few exceptions) in the simulated, measured and analytically determined frequencies of resonance were offset by closely matched overall resonant patterns. This inspires confidence in the simulation technique within known bounds, and demonstrates its capacity for reproducing complex acoustic behaviours. For instance, shifts in characteristic acoustic impedance caused by transitions in cross-sectional area. The complexity of the vocal tract models was consequently increased to that of vocal tract analogues derived from cross-sectional area functions obtained from X-Ray imagine (for Japanese vowels).

In the case of these vocal tract models, correlation between simulation and measurement remained strong, although frequency shifting in F3 and F4 was observed in some cases. Although the shifts are quite small they demonstrated the capacity to form degenerate modes between F3 and F4.

Despite this, reliable reproduction of resonant behaviours at high frequencies suggests the methodology is still appropriate. With the last stage of validation complete, the next step entails application of the technique to models of the vocal tract itself.

Development of the vocal tract models is covered in detail in this thesis. As with most data acquired here, this too must be validated, in this case to investigate whether the imaging is an appropriate representation of the vocal tract configuration for each phone. To approach this, a protocol is devised whereby each subject repeats a similar procedure to that executed in the MRI scanner, only in an anechoic chamber, before and after scanning. Recordings of acoustic outputs are then used to assess the consistency (are successive phonations the same?) and stability (does each phonation remain the same for the duration of a scan?) of each utterance. The results were encouraging, demonstrating that the subjects in most cases were able to repeat the same phonation with consistency in phonetic quality over a time frame of several hours. The phonations also typically remained stable for the duration of the scan (after onset). These findings suggest both that the MR imaging was likely to be a true representation of each vowel, and that the audio recorded could be used as a form of benchmark for resynthesis.

The developed simulation technique was consequently applied to the vocal tract models and compared with power spectral densities of the recorded vowels. These power spectral densities provide only a limited accuracy benchmark due to their characteristic averaging of frequency behaviours (which in the case of the vocal tract is not likely to remain perfectly static). The key resonant characteristics of each phone are however clear.

The results are not always consistent. While some simulations provided accurate matches across the entire range of interest, others were only able to produce the first 4 formants correctly. Even in the latter cases, resynthesised audio examples were convincing, and the best cases provide very strong matches to their recorded acoustic equivalents. It is interesting to observe such an accurate representation of voice production through simulation of the vocal tract in isolation of the rest of the voice production anatomy. This is particularly the case considering the basic boundary formulations used.

Decomposition of the vocal tract models into lower dimensionality equivalents demonstrated interesting trends. One-dimensional simulation appears to be reliable up to no higher than 2kHz, although it is able to reproduce the first two formants with confidence. Two-dimensional widthwise simulation meanwhile performs poorly and is able to correctly produce the first two formants in the best cases. At higher frequencies the response begins to appear effectively random. The two-dimensional impedance-mapped approach in contrast works particularly well. In some cases the system is able to provide a strong correlation up to 6kHz, although it is inconsistent at these higher frequencies.

Comparison reveals that of the approaches considered here, three-dimensional simulation provides the most accurate representation of the vocal tract transfer function. It is not perfectly consistent, nor perfectly accurate in even the best cases. Despite this it does the best job of recreating the full-range frequency responses obtained from audio recording.

It is therefore straightforward to conclude that the three-dimensional simulation is more accurate, but this does not satisfy the hypothesis.

Audio comparison of vowels simulated with artificial and real source waveforms illuminates the difficulty in decoupling the correctness of the vocal tract model from the nature of the source. It is heard that when the more accurate vocal tract model is used in combination with a synthetic source, the outcome is poor. Conversely, the use of a very basic vocal tract model with a recorded source waveform results in very convincing voice synthesis. To this end, it is possible to state progression to a higher dimensionality vocal tract model (as posed in the hypothesis) does not result in more natural voice synthesis. Even in a simple form - accurate only to early formant frequencies, the vocal tract transfer function appears to provide the necessary cues as to suggest a human utterance. Full three-dimensional simulation results in a vocal tract model that is similarly natural, yet more *accurate*, and it is here that the value of three-dimensional simulation lies.

Undoubtedly, three-dimensional simulation comes at a large computational cost, although run times and memory demands are still reasonable for a standard desktop computer. The remaining justification for less accu-

rate (and less computationally demanding) lower-dimensionality modelling lies in its potential for real-time operation, a potential that high resolution three-dimensional simulation is some way off achieving.

9.3 Novelty and Contribution

Development of a three-dimensional digital waveguide mesh-based numerical model of the vocal tract based on MRI data

Recent research has modelled the vocal tract using multi-dimensional transmission line matrix methods, finite-volume and finite-element techniques. This constitutes the first extensive and exhaustively-validated application of the three-dimensional digital waveguide mesh to the acoustics of the vocal tract.

Validation of the application of the three-dimensional digital waveguide mesh to mechanical vocal tract analogues

Through acoustic measurement, analytical determination and error analysis the validity of the application of the three-dimensional digital waveguide mesh using basic one-connection boundary formulations is established.

A novel technique for broadband measurement of the acoustic response of mechanical vocal tract analogues

While many similar approaches have been taken to transducer inversion in acoustic measurement, to the author's knowledge this constitutes the first approach to correction by which the region of interest is decomposed into sub-bands and the length of each exponential sine is modulated to provide gain adjustments before regressive inversion.

A set of MRI data specifically targeting phoneme reproduction for a range of trained subjects, including benchmark audio recordings

Several studies have collected MR imaging targetting the vocal tract, however the author is aware of no other dataset combining such a range of different image types, subjects from different backgrounds and a collection of benchmarking audio data. The dataset also represents the latest capabilities in MR scanning.

A comparison of three-dimensional, two-dimensional, impedance mapped and one-dimensional digital waveguide-based models of the vocal tract

This study provides the first comparison of equivalent three-dimensional, two-dimensional, two-dimensional impedance mapped and one-dimensional digital waveguide-based simulations of the vocal tract transfer function, each of which is based on the same data.

9.4 Further Work

This study constitutes the first step in the development of an exciting platform for the investigation and modelling of the human voice production system. Areas for further work focus on the shortcomings of the existing implementation, and the development of additional features for the same platform. This drives the research towards a more complete representation of the voice anatomy. Consideration is also paid to how the model might be applied in practice. The most pressing concerns for future work are perhaps development of a geometrically complete model, encompassing the nasal tract, teeth and tongue. The vocal tract alone of course constitutes only a single component of the vocal anatomy. Inclusion of the nasal tract especially is crucial and the acquired corpus provides the necessary base data. Development of a suitable technique for segmentation is required. Furthering the theme of geometrical completeness, an appropriate source model would then also be needed. This might demand a step towards development of a dynamic model

due to the naturally dynamic nature of the vocal source. Building a fully dynamic model would allow for reproduction of frication and plosive sounds, and might not be too demanding a task considering the ease with which the digital waveguide mesh can be transformed into a dynamic form using impedance mapping. Finally, the numerical simulation might be fine-tuned using a more complete representation of reflecting/absorbing boundaries.

Source Modelling

The source has been identified as a crucial contributor to the accuracy and resulting naturalness of voice synthesis. Despite this, the current implementation uses an extremely basic point source as a basis for impulse excitation. Development of the source could proceed in several ways. The sub-glottal system could be segmented and modelled. The immediate sub-glottal system (trachea) is very clear on the collected imaging, and the use of a combined head-neck coil means that the field of vision expands into the upper torso. Following the path downwards through the trachea leads to the bronchial tree. This is an asymmetrical, interconnected system of branches coupled to the lungs. Such a system is bound to represent interesting damping conditions, although it's not clear how difficult it might prove to achieve an accurate segmentation of its intricate geometry. The asymmetry of the lungs is particularly interesting, with the right lung typically larger than the left [43]. Given the influence of the branched sinuses in the nasal cavity [120, 59] and effect of asymmetry [88], it would be extremely interesting to incorporate changes in lung volume as a branched cavity.

Several interesting approaches could be taken to better representation of the glottis in these models. The static, closed-decoupled condition used is highlighted as a likely source of error. Beyond the glottal closure condition inherent to typical vocal fold vibration, the glottis is not predominantly closed. Indeed this is clear through the definition of the glottal open/closed quotients as characteristics of voice. Source behaviour does however encompass particularly complex behaviours and is an active area of research in fluid dynamics based voice research. The digital waveguide mesh would require an

effective niche to provide value in the face of more mathematically rigourous approaches. Since the focus of the technique is towards dynamic modelling, it could be possible for example to incorporate a form of dynamic boundary, where reflection and transmission is modulated across an appropriate geometrical analogue to reproduce the effect of rapidly changing glottal area or to investigate the acoustic correlate of changing contact quotients. Since such a dynamic operation is currently beyond the capabilities of three-dimensional simulation, the two-dimensional impedance-mapped technique might be considered, having demonstrated a strong tendency towards natural resynthesis. Such a model would provide an attractive extension to the one-dimensional analogue developed by Liljencrants [62] and the mechanical models of Barney et al. [42].

Alternatively, if a three-dimensional solution was preferred, several impulse responses could be recorded, for a range of glottal closure conditions to investigate their influence on the vocal tract transfer function from a stationary viewpoint.

An exciting option might be the development of a more appropriate injection mechanism. The use of a point source clearly constitutes an anthropomorphically limited implementation, especially considering that the behaviour of the vocal tract is predominantly one-dimensional. The combination of multiple receiver points to produce a desired spatial sensitivity has been demonstrated [39]. It is quite conceivable that the same approaches could be taken to combination of multiple injection points to produce a desired source directivity. Indeed, the nature of the source itself could be modulated by the condition of the glottal model previously suggested. This might create an exciting feedback system, whereby the glottal boundaries are influenced by the source waveform. Given that the use of simple cyclic source waveforms (such as the LF model demonstrated in this study) regularly results in unnatural voice synthesis, the use of such a 'self-interfering' source mechanism might introduce interesting variations and cues suggesting a natural human phonation.

The geometry of the laryngeal cavity is not simple. Significant recent interest has been paid for example to the role of the piriform recesses in

the vocal tract transfer function [64]. Now that a data set has been acquired which clearly shows the recesses, it would be trivial to artificially modify their size before simulation. This would provide an interesting means of exploring their acoustic contribution.

Nasal Tract

The lack of a nasal tract model constitutes one of the most fundamental omissions from the current model. High quality imaging has been acquired for several subjects. Segmentation was attempted but proper determination of the applicable propagation path is not straightforward. The nasal cavity is not an empty space like the vocal tract. Instead the conchae support a constricted network of soft tissue, nasal mucous and hair. Constrictions form a complicated case in flow conditions and the acquisition of appropriate reflective parameters is likely to prove difficult. Additionally, much of the nasal tract is cartiligious, complicating delineation of the cavity walls from the neighbouring passages. This is especially problematic for the coupling of the paranasal sinuses. While themselves easily determinable, the sinuses are connected to the cavity by ostia embedded in layers of cartilage, which is effectively invisible in the imaging. Despite the complications, the static nature of the nasal cavity would allow the same model to be coupled to all vocal tract models for a given subject. The acoustic phonetic contribution of the nasal cavity and its coupling with the vocal tract is thought to be significant. It is also extremely interesting to correlate aspects of nasal tract / paranasal sinus resonance with voice pedagogy. Singers are occasionally told to sing through the centre of the forehead for example. It would be extremely interesting to investigate whether this advise in some way correlates with resonant conditions in the ethmoidal sinuses. A complete, integrated model would be of significant value, but is complicated by the intricacy of the nasal tract and the difficulties inherent to its non-invasive imaging.

Teeth

By obtaining a more accurate model of the teeth, their geometries could easily be integrated into the existing graphical models. Coupled with appropriate reflective parameters, their inclusion would represent an important improvement to the model. The source of such data is immediately unclear, given the ethical considerations raised by X-Ray scanning. If casts or scans could be obtained from past dental records, it would be relatively trivial to generate a three-dimensional model of the upper/lower jaw, which could be effectively 'attached' to markers in the imaging. The inclusion of a dental model would also vastly simplify (and also increase the accuracy of) the segmentation process, given that a subjective removal of teeth will not be required.

A consideration of the reflective properties of the teeth might also prove interesting. At present the vocal tract boundaries are given a blanket reflective coefficient, which probably constitutes a serious approximation. Given that the teeth are often directly in the one-dimensional acoustic path, their contribution might prove significant. It would also be fascinating to incorporate the teeth (and their particular reflective behaviours) in an exploration of the reproduction of frication or plosives.

Tongue

The tongue (in combination with the jaw) perhaps provides the most significant articulation of the vocal tract geometry. Given that its range of motion and behaviours are well understood, it might prove straightforward to develop a physically modelled tongue in combination with a higher-dimensionality modelling technique. It would be particularly interesting to investigate the use of impedance mapping to modulate reflective behaviours across anticipated ranges of motion. As with the teeth, the tongue of course would feature its own characteristic reflective behaviours. An interesting first step might be to adjust the relationship between reflective parameters at the hard palate and the tongue surface, to observe any consequent acoustic correlation.

Dynamic

Development of the three-dimensional model into a dynamic model is the obvious progression for the simulation technique itself, as mirrored in the adaptation of the two-dimensional paradigm to the impedance-mapped case. This also satisfies an important justification of the use of digital waveguide schemes over more computationally cost-effective numerical implementations such as the finite-difference time-domain approach. The largest challenge encountered is likely to be in computational cost. Real-time operation is not feasible given current abilities, although dynamic offline operation is conceivable perhaps granted a reduction in system sampling rate and a stronger focus on optimisation of the implementation. An intermediate solution that might benefit from exploration is transition between impulse responses [121]. A means of moving effectively between target phones (perhaps guided by the dynamic imaging collected) might provide an effective method for dynamic voice synthesis although its physical analogue would of course be limited.

Further Speakers

Given that the current experimental procedure is established, repeating the scanning processes for a wider range of subjects would present a trivial and highly valuable extension to the current dataset.

Frication

At present, the system is only used to reproduce voiced stead-state vowels, based on convolution with a suitable source mechanism. While flow-based (turbulent) sound sources such as frication are not implicitly reproduced, it is possible to approximate them in physical models through noise injection at the point of constriction. In the model developed here, this could be implemented by the addition of a secondary source point at constriction and a corresponding impulse response recorded. By convolution with suitable noise signal the effect of frication could easily be reproduced.

Boundary Formulations

It is recognised that the boundary formulations used in the scope of this work constitute a significant approximation to the true reflective characteristics in the vocal tract. In the current model it is possible to assign different reflective properties to different regions of the vocal tract and the modular implementation allows for trivial replacement of boundary classes. Acquisition of more appropriate reflective parameters would permit a more accurate representation of boundary behaviours, which would possibly be reflected in a more accurate vocal tract transfer function.

Appendix A

The 1D Wave Equation

Many derivations of the one-dimensional wave equation are based on the case of a string. While this has the benefit of eliminating a geometrical degree of freedom, it can be confusing as the case of a transverse wave is not applicable to acoustic propagation in air. Instead a derivation based on gas particles is preferred. The following is largely based on the derivation of Fletcher [15].

Consider a volume of air, V, of infinitesimally small depth dx and face area S. It is clear that the volume is given by equation (A.1).

$$V = Sdx \tag{A.1}$$

This volume of air is free to move in the x axis and on this axis alone. For this reason, consider that the volume is mounted on four frictionless 'rails' as in Figure A.1.

A wave arrives at the position of the volume, causing it to move by a displacement ξ, and expand in depth by a distance E, as shown in Figure A.2.

The volume after expansion, V^+ is clearly (A.2).

$$V^+ = S(dx + E) \tag{A.2}$$

The rarefaction can hence be expressed as (A.3) by considering the change in displacement of the front and back faces of the volume over the course of a displacement in x.

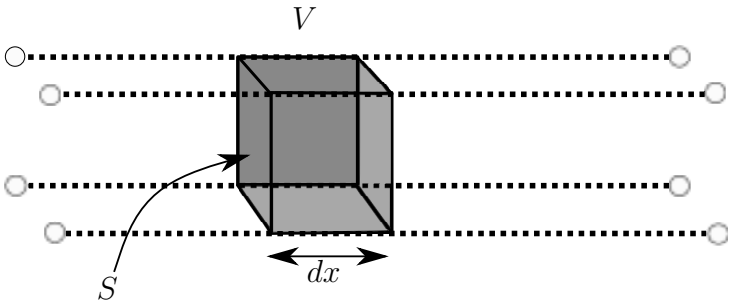


Figure A.1: Finite Acoustic Volume Moving in a Single Axis

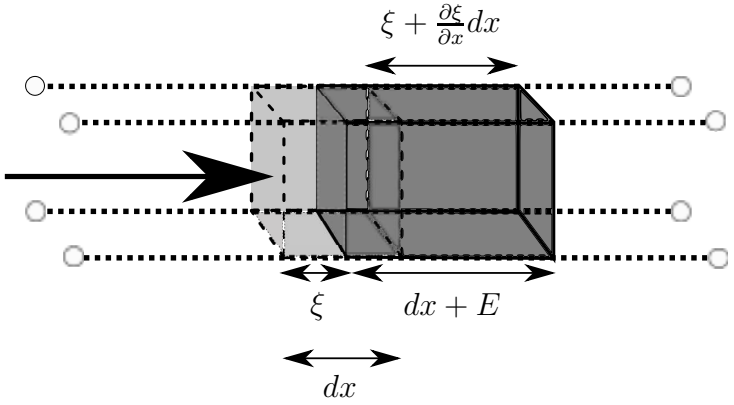


Figure A.2: Finite Acoustic Volume After Movement and Expansion

$$E = \frac{\partial \xi}{\partial x} dx \tag{A.3}$$

The new volume can be described in terms of this rarefaction as (A.5).

$$V^+ = S(dx + \frac{\partial \xi}{\partial x} dx) \tag{A.4}$$

$$= Sdx(1 + \frac{\partial \xi}{\partial x}) \tag{A.5}$$

Recalling the definition of the initial volume V^- in equation (A.6), express the change in volume, ΔV as (A.7).

$$V^{-} = Sdx \tag{A.6}$$

$$\Delta V = Sdx\frac{\partial \xi}{\partial x} \tag{A.7}$$

The bulk modulus of a given gas, K, describes the relationship between changes in pressure and volume as per (A.8).

$$K = -V\frac{\partial p}{\partial v} \tag{A.8}$$

Rearranging (A.8) to describe changes in pressure then results in (A.9).

$$dp = \frac{-Kdv}{V} \tag{A.9}$$

By substituting the change in volume (A.7) we find (A.10).

$$dp = \frac{-K\frac{\partial \xi}{\partial x}}{dx} \tag{A.10}$$

Consider Newton's second law (A.11), where F is force, m the mass of the gaseous volume and a the acceleration of the displacement ξ.

$$F = ma \tag{A.11}$$

Rewriting (A.11) for the change in the volume results in (A.12), where ρ is the density of the medium.

$$F = \rho Sdx\frac{\partial^2 \xi}{\partial t^2} \tag{A.12}$$

It is then possible to equate (A.12) with the force exerted by the increase in pressure (A.13), where $F_{\perp}$ represents the force resulting from pressure p acting in a normal direction to area A.

$$F_{\perp} = pA \tag{A.13}$$

It is then possible to rewrite (A.13) as (A.14).

$$F_{\perp} = -\frac{\partial p}{\partial x} dx S \tag{A.14}$$

Next, equate the two definitions of the acting force ((A.12) and (A.14)) to yield (A.15).

$$-\frac{\partial p}{\partial x} dx S = \rho S dx \frac{\partial^2 \xi}{\partial t^2} \tag{A.15}$$

Simplifying (A.15) then results in (A.16).

$$-\frac{\partial p}{\partial x} = \rho \frac{\partial^2 \xi}{\partial t^2} \tag{A.16}$$

Returning to (A.10), take the derivative of both sides with respect to x, to find (A.17).

$$\frac{\partial p}{\partial x} = -K \frac{\partial^2 \xi}{\partial x^2} \tag{A.17}$$

This equation can be substituted into (A.16) to give (A.19).

$$\rho \frac{\partial^2 \xi}{\partial t^2} = K \frac{\partial^2 \xi}{\partial x^2} \tag{A.18}$$

$$\frac{\partial^2 \xi}{\partial t^2} = \frac{\rho}{K} \frac{\partial^2 \xi}{\partial x^2} \tag{A.19}$$

Finally, introduce the Newton-Laplace equation (A.20) to give the final expression of the wave equation as (A.21).

$$c^2 = \frac{K}{\rho} \tag{A.20}$$

$$\frac{\partial^2 \xi}{\partial t^2} = c^2 \frac{\partial^2 \xi}{\partial x^2} \tag{A.21}$$

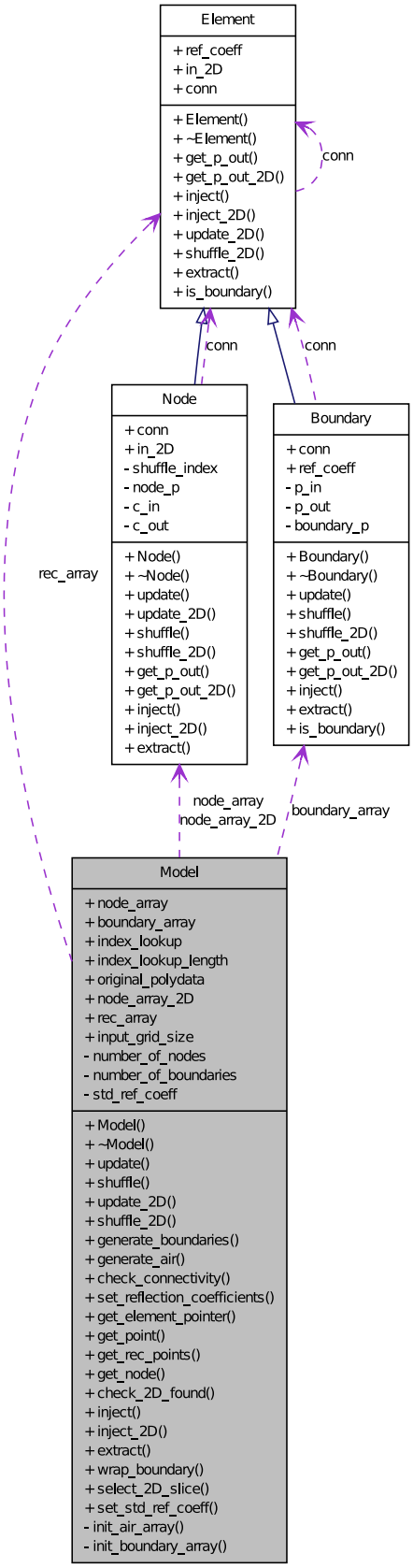
The example presented in Fig. A.2 represents an approximation of the complete case for wave propagation in a gaseous medium in that it is constrained to movement on a single axis, x. A similar expression can be defined describing such behaviour as a function of any coordinate system, as in (A.22) or for three Cartesian coordinates in (A.23)

$$\frac{\partial^2 \xi}{\partial t^2} = c^2 \nabla^2 \xi \tag{A.22}$$

$$\frac{\partial^2 \xi}{\partial t^2} = c^2 \left(\frac{\partial^2 \xi}{\partial x^2} + \frac{\partial^2 \xi}{\partial y^2} + \frac{\partial^2 \xi}{\partial z^2} \right) \tag{A.23}$$

Appendix B

Simulation Data Structure Collaboration Diagram


Element
+ ref_coeff
+ in_2D
+ conn
+ Element()
+ ~Element()
+ get_p_out()
+ get_p_out_2D()
+ inject()
+ inject_2D()
+ update_2D()
+ shuffle_2D()
+ extract()
+ is_boundary()
conn
conn
conn
Node
+ conn
+ in_2D
- shuffle_index
- node_p
- c_in
- c_out
+ Node()
+ ~Node()
+ update()
+ update_2D()
+ shuffle()
+ shuffle_2D()
+ get_p_out()
+ get_p_out_2D()
+ inject()
+ inject_2D()
+ extract()
Boundary
+ conn
+ ref_coeff
- p_in
- p_out
- boundary_p
+ Boundary()
+ ~Boundary()
+ update()
+ shuffle()
+ shuffle_2D()
+ get_p_out()
+ get_p_out_2D()
+ inject()
+ extract()
+ is_boundary()
rec_array
node_array
node_array_2D
boundary_array
Model
+ node_array
+ boundary_array
+ index_lookup
+ index_lookup_length
+ original_polydata
+ node_array_2D
+ rec_array
+ input_grid_size
- number_of_nodes
- number_of_boundaries
- std_ref_coeff
+ Model()
+ ~Model()
+ update()
+ shuffle()
+ update_2D()
+ shuffle_2D()
+ generate_boundaries()
+ generate_air()
+ check_connectivity()
+ set_reflection_coefficients()
+ get_element_pointer()
+ get_point()
+ get_rec_points()
+ get_node()
+ check_2D_found()
+ inject()
+ inject_2D()
+ extract()
+ wrap_boundary()
+ select_2D_slice()
+ set_std_ref_coeff()
- init_air_array()
- init_boundary_array()

Appendix C

Corpus Index

A large range of data was collected in the course of this study, as documented in Chapter 6. Simulation within the scope of this particular project focussed on simulation of a smaller subset of standard vowel sounds, as listed in Table 6.1, for the subjects given in Table C.1. For each subject, segmentation of each of the standard vowels was attempted. In some cases segmentation was not achieved due to motion artefacting. Tables C.2, C.3, C.4, C.5, and C.6 below, index the vowel sounds for which segmented data, graphical models and prebuilt numerical models are available. As with all the target sounds, extensive before/after/supine/standing benchmark audio is provided, as described in Chapter 6.

Pseudonym	Details
Jack	Male, English first language, classically trained singer (tenor range), phonetically trained.
Jill	Female, German first language, phonetically trained, no singing background.
Jasmine	Female, English first language, phonetically trained, classically trained professional singer (mezzo-soprano range).
Jim	Male, English first language, classically trained singer (tenor range), no phonetic training.
Jeff	Male, French first language, classically trained singer (bass range), no phonetic training.

Table C.1: Pseudonyms and backgrounds of subjects for MR imaging

Group	IPA	Co. Context
Vowel	/iː/	*neap*
	/ɪː/	*jib*
	/ɛː/	*red*
	/æː/	*anchor*
	/ɑː/	*hard*
	/ɒː/	*locker*
	/ɔː/	*port*
	/ʊː/	*foot*
	/uː/	*food*
	/ʌː/	*rudder*
	/ɜː/	*stern*

Table C.2: Segmented Data available for Jack, using coarticulatory contexts as per [4]

Group	IPA	Co. Context
Vowel	/iː/	*neap*
	/ɪː/	*jib*
	/ɛː/	*red*
	/æː/	*anchor*
	/ɑː/	*hard*
	/ɒː/	*locker*
	/ɔː/	*port*
	/ʊː/	*foot*
	/uː/	*food*
	/ʌː/	*rudder*
	/ɜː/	*stern*

Table C.3: Segmented Data available for Jill, using coarticulatory contexts as per [4]

Group	IPA	Co. Context
Vowel	/ɪː/	*jib*
	/ɛː/	*red*
	/æː/	*anchor*
	/ɑː/	*hard*
	/ɒː/	*locker*
	/uː/	*food*

Table C.4: Segmented Data available for Jasmine, using coarticulatory contexts as per [4]

Group	IPA	Co. Context
Vowel	/iː/	*neap*
	/ɪː/	*jib*
	/ɑː/	*hard*
	/ɒː/	*locker*
	/ɔː/	*port*
	/ʊː/	*foot*
	/uː/	*food*
	/ʌː/	*rudder*
	/ɜː/	*stern*

Table C.5: Segmented Data available for Jim, using coarticulatory contexts as per [4]

Group	IPA	Co. Context
Vowel	/iː/	*neap*
	/ɪː/	*jib*
	/ɛː/	*red*
	/æː/	*anchor*
	/ɑː/	*hard*
	/ɒː/	*locker*
	/ɔː/	*port*
	/ʊː/	*foot*
	/uː/	*food*
	/ʌː/	*rudder*
	/ɜː/	*stern*

Table C.6: Segmented Data available for Jeff, using coarticulatory contexts as per [4]

Appendix D

Supporting CD Materials

A supporting CD is delivered with this thesis, providing audio examples and a digital copy of the thesis.

The CD is designed to be used as a website and hence index.html should be the first file launched. Audio examples are provided as 48kHz 16-bit .wav PCM files. They are indexed and embedded within the website, although the original data files are still easily accessible in the **audio** folder. The structure of the CD is given in Fig. D.1.

Each subfolder within the **audio** folder contains examples of a certain type of synthesis, as per Table D.1. The contents of each consequent folder are separated into additional folders by subject pseudonym. Audio filenames begin with identification by the co-articulatory contexts given in Table 6.1. The second component of the filename describes what they are (as in *3Dsim*, *ImpMapped*, *1Dsim*). Electroglottographic waveform filenames always end in *_lx.wav* to clearly separate them from resynthesised audio examples.

1D	Results of simulation using the simple one-dimensional Kelly-Lochbaum model, based on data derived from the vocal tract models and convolved with the electroglottographic waveforms recorded during benchmarking
Z_mapped	Results of simulation using Mullen's two-dimensional impedance-mapped digital waveguide mesh model of the vocal tract, based on data derived from the vocal tract models and convolved with the electroglottographic waveforms recorded during benchmarking
3D	The results of full three-dimensional simulation of the vocal tract using the methodology explored in the course of this thesis. Simulated impulse responses are convolved with the electroglottographic waveforms recorded during benchmarking
3D_LF	As with **3D** except impulse responses are convolved with Liljencrants-Fant model source waveforms
egg	The electroglottographic waveforms used throughout the project are presented.

Table D.1: Index for audio materials on supporting CD

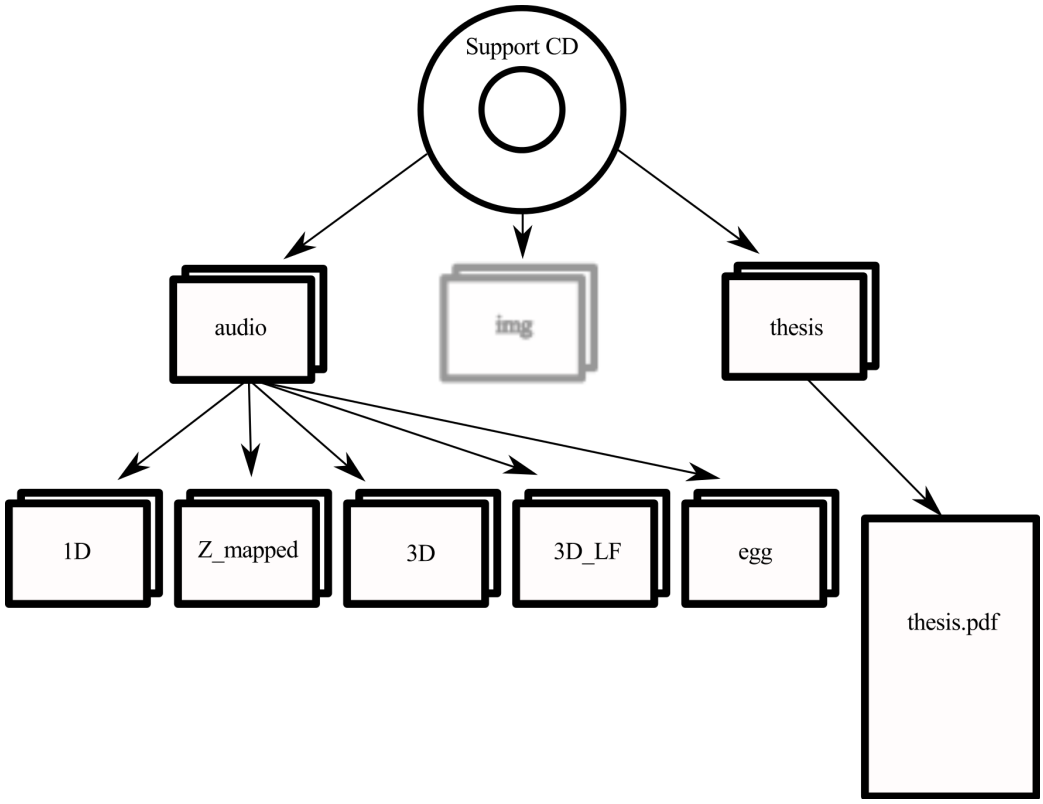


Figure D.1: Top level structure of supporting CD

References

[1] S. Ternström, "Does the Acoustic Waveform Mirror the Voice?" in *Logopedics Phoniatrics Vocology, vol.30, 3-4*, 2005, pp. 100–107.

[2] G. Peterson and H. Barney, "Control Methods used in a Study of Vowels," in *Journal of the Acoustical Society of America, 24*, 1952, pp. 175–184.

[3] D. W. McRobbie, E. A. Moore, and M. J. Graves, *MRI from Picture to Proton*. Cambridge Univ Pr, 2003.

[4] D. M. Howard and D. T. Murphy, *Voice Science Acoustics and Recording*. Plural Publishing, 2008.

[5] D. Blackstock, *Fundamentals of Physical Acoustics*. John Wiley and Sons, Inc, 2000.

[6] D. Murphy, A. Kelloniemi, J. Mullen, and S. Shelley, "Acoustic modeling using the digital waveguide mesh," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 55–66, 2007.

[7] D. T. Murphy and M. Beeson, "The KW-Boundary Hybrid Digital Waveguide Mesh for Room Acoustics," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 552–564, Feb. 2007.

[8] M. Kim and G. P. Scavone, "Domain Decomposition Method for the Digital Waveguide Mesh," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2009, pp. 21–24.

[9] T. Arai, "The replication of Chiba and Kajiyama's mechanical models of the human vocal cavity," *Journal of the Phonetic Society*, vol. 31.

[10] "Waseda Talker WT-7RII - Anthropomorphic Talking Robot - Takanishi Lab, Waseda University, Japan," http://www.takanishi.mech.waseda.ac.jp/top/research/voice/index.htm, 2012.

[11] M. Kob, "Physical Modeling of the Singing Voice," Ph.D. dissertation, University of Technology, Aachen, 2002.

[12] G. Rosen, "Dynamic analog speech synthesizer," Ph.D. dissertation, Massachusetts Institute of Technology, Dept. of Electrical Engineering, 1960.

[13] J. Epps, J. R. Smith, and J. Wolfe, "A novel instrument to measure acoustic resonances of the vocal tract during phonation," *Measurement Science and Technology*, vol. 8, no. 10, p. 1112, Oct. 1997.

[14] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, "Fundamentals of Acoustics," 1999.

[15] N. H. Fletcher and T. D. Rossing, *The physics of musical instruments*. Springer Verlag, 1998.

[16] J. Mullen, "Physical modelling of the vocal tract with the 2D digital waveguide mesh," Ph.D. dissertation, Department of Electronics, University of York, Apr. 2006.

[17] S. B. Shelley, "Diffuse Boundary Modelling in the Digital Waveguide Mesh," Ph.D. dissertation, University of York, 2007.

[18] J. Liljencrants, "End Correction at Flue Pipe Mouth," http://www.fonema.se/mouthcorr/mouthcorr.htm, 2006.

[19] H. Levine and J. Schwinger, "On the Radiation of Sound from an Unflanged Circular Pipe," *Physical Review Online Archive (Prola)*, vol. 73, no. 4, pp. 383–406, Feb. 1948.

[20] M. Atig, J. P. Dalmont, and J. Gilbert, "Termination impedance of open-ended cylindrical tubes at high sound pressure level," *Comptes Rendus Mécanique*, vol. 332, no. 4, pp. 299–304, 2004.

[21] M. Kaltenbacher, M. Escobar, S. Becker, and I. Ali, "Numerical simulation of flow-induced noise using LES/SAS and Lighthill's acoustic analogy," *Int. J. Numer. Meth. Fluids*, vol. 63, no. 9, pp. 1103–1122, 2010.

[22] M. Karjalainen and C. Erkut, "Digital Waveguides versus Finite Difference Structures: Equivalence and Mixed Modeling," in *EURASIP Journal on Applied Signal Processing, 2004:7, 978-989*, 2004.

[23] K. Stroud and D. Booth, *Advanced Engineering Mathematics*, 4th ed. Palgrave Macmillan, 2003.

[24] S. Bilbao, *Wave and Scattering Methods for Numerical Simulation*. John Wiley & Sons, London, 2004.

[25] K. Kowalczyk and M. van Walstijn, "Formulation of Locally Reacting Surfaces in FDTD/K-DWM Modelling of Acoustic Spaces," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 891–906.

[26] R. Rabenstein, S. Petrausch, A. Sarti, G. De Sanctis, C. Erkut, and M. Karjalainen, "Blocked-based physical modeling for digital sound synthesis," *Signal Processing Magazine, IEEE*, vol. 24, no. 2, pp. 42–54, Mar. 2007.

[27] J. O. Smith, "Physical Audio Signal Processing, December 2008 Edition," http://ccrma.stanford.edu/~jos/pasp/, accessed 24-09-09.

[28] G. R. Campos and D. M. Howard, "On the Computational Efficiency of Different Waveguide Mesh Topologies for Room Acoustic Simulation," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 1063–1072, Aug. 2005.

[29] M. Karjalainen and C. Erkut, "Digital waveguides versus finite difference structures: Equivalence and mixed modeling," *EURASIP Journal on Applied Signal Processing*, pp. 978–989, 2004.

[30] K. Kowalczyk and M. van Walstijn, "Formulation of a Locally Reacting Wall in Finite Difference Modelling of Acoustic Spaces," in *International Symposium on Room Acoustics, Seville, 10-12 September*, 2007.

[31] V. Välimäki, J. Pakarinen, C. Erkut, and M. Karjalainen, "Discrete-time Modelling of Musical Instruments," *Reports on Progress in Physics*, vol. 69, pp. 1–78, 2006.

[32] A. Bamberger, R. Glowinski, and Q. H. Tran, "A Domain Decomposition Method for the Acoustic Wave Equation with Discontinuous Coefficients and Grid Change," *SIAM Journal on Numerical Analysis*, vol. 34, no. 2, pp. 603–639, 1997.

[33] J. Mullen, D. M. Howard, and D. T. Murphy, "Real-Time Dynamic Articulations in the 2-D Waveguide Mesh Vocal Tract Model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 577–585, Jan. 2007.

[34] J. B. Schneider, C. L. Wagner, and O. M. Ramahi, "Implementation of transparent sources in FDTD simulations," *Antennas and Propagation, IEEE Transactions on*, vol. 46, no. 8, pp. 1159–1168, August 1998.

[35] F. Fontana and D. Rocchesso, "Signal-theoretic characterization of waveguide mesh geometries for models of two-dimensional wave propagation in elastic media," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 2, pp. 152–161, Feb. 2001.

[36] K. Kowalczyk and M. van Walstijn, "Room Acoustics Simulation Using 3-D Compact Explicit FDTD Schemes," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 34–46, Jan. 2011.

[37] L. Savioja and V. Valimaki, "Interpolated rectangular 3-D digital waveguide mesh algorithms with frequency warping," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 783–790, Nov. 2003.

[38] J. C. Strikwerda, *Finite difference schemes and partial differential equations*. Society for Industrial Mathematics, 2004.

[39] A. P. Southern, "The synthesis and auralisation of physically modelled soundfields," Ph.D. dissertation, University of York, 2010.

[40] H. Traunmüller, "Conventional, biological and environmental factors in speech communication: A modulation theory," *Phonetica*, vol. 51, no. 1-3, pp. 170–183, 1994.

[41] J. Liljencrants, *PhD Thesis: Speech Synthesis with a Reflection-Type Vocal Tract Analog*. KTH, Stockholm, 1985.

[42] A. Barney, A. De Stefano, and N. Henrich, "The Effect of Glottal Opening on the Acoustic Response of the Vocal Tract," *Acta Acustica United with Acustica*, vol. 93, pp. 1046–1056, 2007.

[43] C. S. Sinnatamby, *Last's Anatomy - Regional and Applied*, eleventh ed. Elsevier: Churchill Livingstone, 2006.

[44] H. Traunmüller and A. Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults," *Unpublished Manuscript*, 1995.

[45] G. Fant, *Acoustic Theory of Speech Production With calculations based on X-Ray Studies of Russian Articulations*. Mouton de Gruyter, 1970.

[46] K. Stevens, *Acoustic Phonetics*, ser. Current Studies in Linguistics. Cambridge, Massachusetts: The MIT Press, 1998.

[47] M. Speed, D. Murphy, and D. M. Howard, "Acoustic Coupling in Multi-Dimensional Finite Difference Schemes for Physically Modeled Voice Synthesis," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2009.

[48] B. H. Story, "A comparison of vocal tract perturbation patterns based on statistical and acoustic considerations," *The Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. EL107–EL114, 2007.

[49] M. Mrayati, R. Carré, and B. Guérin, "Distinctive Regions and Modes: A New Theory of Speech Production," in *Speech Communication, Volume 7, Issue 3*, 1988, pp. 257–286.

[50] L. J. Boë and P. Perrier, "Comments on 'distinctive regions and modes: a new theory of speech production' by M. Mrayati, R. Carre and B. Guerin," *Speech Communication*, vol. 9, no. 3, pp. 217–230, 1990.

[51] V. Pagneux, N. Amir, and J. Kergomard, "A study of wave propagation in varying cross-section waveguides by modal decomposition. Part I. Theory and validation," *The Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2034–2048, 1996.

[52] N. Amir, V. Pagneux, and J. Kergomard, "A study of wave propagation in varying cross-section waveguides by modal decomposition. Part II. Results," *The Journal of the Acoustical Society of America*, vol. 101, no. 5, pp. 2504–2517, 1997.

[53] T. Chiba and M. Kajiyama, *The Vowel, Its Nature and Structure*. Tokyo: Kaiseikan, 1958.

[54] P. Badin, K. Motoki, N. Miki, D. Ritterhaus, and M. T. Lallouache, "Some geometric and acoustic properties of the lip horn," *Les Cahiers de l'ICP. Rapport de recherche*, no. 4, pp. 3–16, 1995.

[55] K. Motoki, P. Badin, and N. Miki, "Measurement of acoustic impedance density distribution in the near field of the labial horn," in *Third International Conference on Spoken Language Processing*. ISCA, 1994.

[56] O. Fujimura and J. Lindqvist, "Sweep-Tone Measurements of Vocal-Tract Characteristics," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 541–558, 1971.

[57] J. L. Gauffin and J. Sundberg, "Acoustic properties of the nasal tract," *Phonetica*, vol. 33, no. 3, pp. 161–168, 1976.

[58] H. Matsuzaki, A. Serrurier, P. Badin, and K. Motoki, "Time-domain FEM Simulation of Japanese and French Vowel /a/ with Nasal Coupling," in *Spring Meeting of the Acoustical Society of Japan*, 2008.

[59] J. Dang, K. Honda, and H. Suzuki, "Morphological and acoustical analysis of the nasal and the paranasal cavities," *The Journal of the Acoustical Society of America*, vol. 96, p. 2088, 1994.

[60] S. Maeda, "The Role of the Sinus Cavities in the Production of Nasal Vowels," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1982, pp. 911–914.

[61] G. Fant and Q. Lin, "Glottal source-vocal tract acoustic interaction," *Speech Transmission Laboratory Quarterly Progress and Status Report*, vol. 1, pp. 13–27, 1987.

[62] J. Liljencrants, "Speech Synthesis with a Reflection-Type Vocal Tract Analog," Ph.D. dissertation, KTH, Stockholm, 1985.

[63] G. Fant, K. Ishizaka, J. Lindqvist-Gauffin, and J. Sundberg, "Subglottal formants," KTH, Stockholm, Tech. Rep. 1, 1972.

[64] J. Dang and K. Honda, "Acoustic characteristics of the piriform fossa in models and humans," *Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 456–465, Jan. 1997.

[65] P. Birkholz and B. J. Kröger, "Vocal tract model adaptation using magnetic resonance imaging," in *7th International Seminar on Speech Production (ISSP'06)*, 2006, pp. 493–500.

[66] C. Westbrook, *MRI at a glance*, 2nd ed. Wiley-Blackwell, 2010.

[67] J. L. Flanagan and L. Cherry, "Excitation of vocal tract synthesizers," vol. 45, no. 3, p. 764, Mar. 1969.

[68] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[69] G. Fant, "The LF-model revisited. transformations and frequency domain analysis," *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, vol. 2, p. 3, 1995.

[70] S. Maeda, "A Digital Simulation Method of the Vocal Tract System," *Speech Communication*, vol. 1, pp. 199–229, 1982.

[71] M. M. Sondhi, "Resonances of a bent vocal tract," *The Journal of the Acoustical Society of America*, vol. 79, no. 4, pp. 1113–1116, 1986.

[72] S. S. Narayanan, A. A. Alwan, and K. Haker, "An articulatory study of fricative consonants using magnetic resonance imaging," vol. 98, no. 3, p. 1325, Sep. 1995.

[73] O. Engwall and P. Badin, "Collecting and analysing two- and three-dimensional MRI data for swedish," Tech. Rep. 3-4, 1999.

[74] T. Frauenrath, W. Renz, J. Rieger, A. Goemmel, C. Butenweg, and T. Niendorf, "High spatial resolution 3D MRI of the larynx using a dedicated TX/RX phased array coil at 7.0T."

[75] T. Frauenrath, A. Goemmel, C. Butenweg, M. Otten, and T. Niendorf, "3D mapping of vocal fold geometry during articulatory maneuvers using ultrashort echo time imaging at 3.0T."

[76] T. Frauenrath, T. Niendorf, and M. Kob, "Acoustic method for synchronization of magnetic resonance imaging (MRI)," *Acta Acustica united with Acustica*, vol. 94, no. 1, pp. 148–155, Jan. 2008.

[77] K. Nishikawa, K. Asama, K. Hayashi, H. Takanobu, and A. Takanishi, "Development of a talking robot," in *Intelligent Robots and Systems, 2000. (IROS 2000). Proceedings. 2000 IEEE/RSJ International Conference on*, vol. 3. IEEE, 2000, pp. 1760–1765 vol.3.

[78] S. Rösler and H. W. Strube, "Measurement of the glottal impedance with a mechanical model," *The Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 1708–1716, 1989.

[79] K. Motoki, N. Miki, and N. Nagai, "Measurement of sound-pressure distribution in replicas of the oral cavity," *The Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2577–2585, 1992.

[80] B. H. Story, "Physiologically-Based Speech Simulation Using an Enhanced Wave-Reflection Model of the Vocal Tract," Ph.D. dissertation, University of Iowa, May 1995.

[81] G. Rosen, "Dynamic analog speech synthesizer," *Acoustical Society of America*, vol. 30, no. 3, pp. 201+, Mar. 1958.

[82] J. L. Kelly and C. C. Lochbaum, "Speech Synthesis," in *Fourth International Congress on Acoustics*, 1962, pp. 1–4.

[83] D. M. Howard, S. Ternström, and M. Speed, "Natural Voice Synthesis: The potential relevance of high-frequency components," in *3rd Advanced Voice Function Assessment International Workshop*, May 2009.

[84] J. Smith, "Digital waveguide synthesis - online book," http://www-ccrma.standford.edu/~jos/wg.html.

[85] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 537–554, 1996.

[86] J. Mullen, D. M. Howard, and D. T. Murphy, "Waveguide physical modeling of vocal tract acoustics: flexible formant bandwidth control from increased model dimensionality," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 964–971, May 2006.

[87] H. Matsuzaki and K. Motoki, "FEM Analysis on Acoustic Characteristics of Vocal Tracts Shape with Different Geometrical Approximation," in *Proceeds of the 6th International Conference on Spoken Language Processing, Vol. 3*, 2000, pp. 897–900.

[88] K. Motoki, "Three-Dimensional Acoustic Field in Vocal-Tract," in *Acoustical Science and Technology, 23, 4*, 2002, pp. 207–212.

[89] T. Vampola, J. Horacek, and J. G. Svec, "Fe modeling of human vocal tract acoustics. part i: Production of Czech vowels," *Acta Acustica united with Acustica*, vol. 94, no. 3, pp. 433–447, May 2008.

[90] T. Vampola, J. Horácek, J. Vokrál, and L. Cerný, "FE Modeling of Human Vocal Tract Acoustics. Part II: Influence of Velopharyngeal Insufficiency on Phonation of Vowels," *Acta Acustica united with Acustica*, vol. 94, pp. 448–460, 2008.

[91] P. Šidlof, "Fluid-structure interaction in human vocal folds," Ph.D. dissertation, Charles University in Prague, Faculty of Mathematics and Physics.

[92] O. Engwall, "Assessing MRI measurements: Effects of sustenation, gravitation and coarticulation," *Speech production: Models, Phonetic Processes and Techniques*, pp. 301–314.

[93] M. Stone, G. Stock, K. Bunin, K. Kumar, M. Epstein, C. Kambhamettu, M. Li, V. Parthasarathy, and J. Prince, "Comparison of speech production in upright and supine position," vol. 122, no. 1, pp. 532+, Jul. 2007.

[94] B. H. Story, I. R. Titze, and E. A. Hoffman, "The relationship of vocal tract shape to three voice qualities," *The Journal of the Acoustical Society of America*, vol. 109, no. 4, pp. 1651–1667, 2001.

[95] B. Story, "Comparison of Magnetic Resonance Imaging-Based Vocal Tract Area Functions Obtained from the Same Speaker in 1994 and 2002," in *Journal of the Acoustical Society of America, 123 (1), January*, 2008, pp. 327–335.

[96] S. Fuchs, R. Winkler, and P. Perrier, "Do Speakers' Vocal Tract Geometries Shape their Articulatory Vowel Space?" in *Proceedings of the 8th International Seminar on Speech Production, ISSP'08*, Y. L. Rudolph Sock, Ed., Strasbourg, France, Dec. 2008, pp. 333–336.

[97] D. Aalto, J. Malinen, M. Vainio, J. Saunavaara, and P. Palo, "Estimates for the Measurement and Articulatory Error in MRI Data from Sustained Vowel Production," in *17th International Congress on Phonetic Sciences*, Aug. 2011.

[98] "Optoacoustics — Fiber Optic Microphones and Sensors," http://www.optoacoustics.com/.

[99] J. Malinen and P. Palo, "Recording speech during MRI: Part II," in *Proceedings of the 6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA2009)*, 2009.

[100] T. Lukkari, J. Malinen, and P. Palo, "Recording speech during magnetic resonance imaging," in *Proceedings of the 5th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA2007)*. Citeseer, 2007.

[101] D. Aalto, J. Malinen, P. Palo, O. Aaltonen, M. Vainio, R. P. Happonen, R. Parkkola, and J. Saunavaara, "Recording speech sound and articulation in MRI," *Proc. Biodevices, Rome*, 2011.

[102] B. R. Gerratt, "Formant frequency fluctuation as an index of motor steadiness in the vocal tract," *Journal of speech and hearing research*, vol. 26, no. 2, p. 297, 1983.

[103] W. J. Schroeder, K. M. Martin, and W. E. Lorensen, "The design and implementation of an object-oriented toolkit for 3D graphics and visualization," in *VIS '96: Proceedings of the 7th conference on Visualization '96*. Los Alamitos, CA, USA: IEEE Computer Society Press, 1996.

[104] W. Schroeder, K. Martin, and B. Lorensen, *The visualization toolkit*. Prentice Hall PTR, 1998.

[105] W. J. Schroeder, K. Martin, L. S. Avila, and C. C. Law, *The VTK user's guide*. Kitware, 2001.

[106] K. Inoue, "Extraction of Vocal Tract Area Function from Three-Dimensional Magnetic Resonance Images using Digital Waveguide Mesh," Master's thesis, 2008.

[107] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[108] R. Lyons, "Reducing FFT scalloping loss errors without multiplication [DSP tips and tricks]," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 112–116, Mar. 2011.

[109] E. Joliveau, J. Smith, and J. Wolfe, "Tuning of vocal tract resonance by sopranos," *Nature*, vol. 427, no. 6970, p. 116, 2004.

[110] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," *Preprints-Audio Engineering Society*, 2000.

[111] R. Ben-Hador and I. Neoran, "Capturing Manipulation and Reproduction of Sampled Acoustic Impulse Responses," in *Audio Engineering Society Convention 117*.

[112] "Matlab multi-channel audio project," http://pa-wavplay.sourceforge.net/, Jan. 2011.

[113] "PortAudio - an Open-Source Cross-Platform Audio API," http://www.portaudio.com/, Jan. 2012.

[114] J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," pp. 1858–1861.

[115] J. A. Laird, "The physical modelling of drums using digital waveguides," Ph.D. dissertation, University of Bristol, Nov. 2001.

[116] T. Arai, E. Maeda, N. Saika, and Y. Murahara, "Physical models of the human vocal tract as tools for education in acoustics," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, p. 2345, 2002.

[117] O. Engwall, "Tongue Talking - Studies in Intraoral Speech Synthesis," Ph.D. dissertation, KTH, Stockholm, 2002.

[118] J. Mullen and D. Murphy, http://www-users.york.ac.uk/~dtm3/vocaltract.html, Feb. 2012.

[119] "WINE - run windows applications on linux, BSD, solaris and mac OS x," http://www.winehq.org/, Feb. 2012.

[120] J. Dang, C. H. Shadle, Y. Kawanishi, K. Honda, and H. Suzuki, "An experimental study of the open end correction coefficient for side branches within an acoustic tube," vol. 104, no. 2, pp. 1075–1084, Aug. 1998.

[121] G. Kearney, "Auditory scene synthesis using virtual acoustic recording and reproduction," Ph.D. dissertation, Trinity College Dublin, Jan. 2010.