Diphthong Synthesis using the Three-Dimensional Dynamic Digital Waveguide Mesh

Amelia Jane Gully

PhD

University of York

Electronic Engineering

September 2017

Abstract

The human voice is a complex and nuanced instrument, and despite many years of research, no system is yet capable of producing natural-sounding synthetic speech. This affects intelligibility for some groups of listeners, in applications such as automated announcements and screen readers. Furthermore, those who require a computer to speak—due to surgery or a degenerative disease—are limited to unnatural-sounding voices that lack expressive control and may not match the user's gender, age or accent. It is evident that natural, personalised and controllable synthetic speech systems are required. A three-dimensional digital waveguide model of the vocal tract, based on magnetic resonance imaging data, is proposed here in order to address these issues. The model uses a heterogeneous digital waveguide mesh method to represent the vocal tract airway and surrounding tissues, facilitating dynamic movement and hence speech output. The accuracy of the method is validated by comparison with audio recordings of natural speech, and perceptual tests are performed which confirm that the proposed model sounds significantly more natural than simpler digital waveguide mesh vocal tract models. Control of such a model is also considered, and a proof-of-concept study is presented using a deep neural network to control the parameters of a two-dimensional vocal tract model, resulting in intelligible speech output and paving the way for extension of the control system to the proposed three-dimensional vocal tract model. Future improvements to the system are also discussed in detail. This project considers both the naturalness and control issues associated with synthetic speech and therefore represents a significant step towards improved synthetic speech for use across society.

Contents

List of Figures

List of Tables

Acknowledgements

I have been very lucky throughout the course of this thesis to have an extremely knowledgeable and supportive supervision team, and my deepest thanks go to Damian Murphy and Helena Daffern, and formerly David Howard, for their guidance and generosity throughout this project. I would also like to thank all the members of the Audio Lab, past and present, for their friendship, encouragement and insight.

A heartfelt thanks to Professor Keiichi Tokuda, and all the members of his lab, for their support during my visit in summer 2016; especially to Takenori Yoshimura whose patience with explaining deep neural networks in a foreign language seemingly knows no bounds.

Thanks are due to the Engineering and Physical Sciences Research Council for funding this PhD, and the Japan Society for the Promotion of Science for making me a Summer Programme Fellow, allowing me to visit Japan for three months to visit Nagoya Institute of Technology and work with Professor Tokuda.

I owe a great debt to my parents, without whose support this endeavour would not have been possible, and to my brother for motivation: if he weren't so good at everything I wouldn't have had to complete a PhD to outdo him. Finally, to Joe, for his support, patience, and delicious curries.

Declaration

I hereby declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this or any other University. All sources are explicitly stated and referenced. I also declare that some parts of the work in this thesis have been presented previously, at conferences and in journals, in the following publications:

- **Perceived naturalness of a 3D dynamic digital waveguide mesh model of the vocal tract**, A. J. Gully and D. M. Howard, presented at the 11th Pan-European Voice Conference (PEVOC-11), Florence, Italy, 31^{st} August - 2^{nd} September, 2015.
- **Articulatory text-to-speech synthesis using the digital waveguide mesh driven by a deep neural network**, A. J. Gully, T. Yoshimura, D. T. Murphy, K. Hashimoto, Y. Nankaku and K. Tokuda, presented at INTERSPEECH 2017, Stockholm, Sweden, 20^{th}–24^{th} August 2017.
- **Diphthong synthesis using the dynamic 3D digital waveguide mesh**, A. J. Gully, H. Daffern and D. T. Murphy, *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 2, pp. 243–255, February 2018.

Part I

Introduction

Chapter 1

Introduction

> It would be a considerable invention indeed, that of a machine able to mimic speech, with all its sounds and articulations. I think it is not impossible.
>
> Leonhard Euler, Letter to Friederike Charlotte of Brandenburg-Schwedt, 1761

Speech is the most natural form of communication among humans. We are not just the only animals on the planet that speak, but the only ones *capable* of doing so, due to the unusually low position of the human larynx [1]. Research suggests that humans may have been speaking to each other 1.75 million years ago [2], long before the development of a written language. Speech is a fundamental part of what makes us human, allowing the development of mental processes that otherwise remain impaired [3]. Crucially, it facilitates the co-operation and sharing of knowledge that has led to humans becoming the dominant species on the planet, despite our outwardly fragile appearance. As Pieraccini [4] puts it, "our unprotected bodies are built for language."

Recognising, understanding and synthesising natural speech is full of challenges, and these are aptly summed up by Bryson [5, p. 83]:

> "People don't talk like this, theytalklikethis. Syllables, words, sentences run together like a watercolor left in the rain. To understand what anyone is saying to us we must separate these noises into words and the words into sentences so that we might in our turn issue a stream of mixed sounds in response. If what we say is suitably apt and amusing, the listener will show his delight by emitting a series of uncontrolled high-pitched noises, accompanied by sharp intakes of breath of the sort normally associated with a seizure or heart failure. And by these means we converse. Talking, when you think about it, is a very strange business indeed."

It is this "strange business" that researchers have been investigating since Kratzenstein's resonators in 1779 and von Kempelen's 'talking machine' in 1791 [6]—with models becoming increasingly complex throughout the twentieth century [7]—in the hope of creating machines that can communicate as naturally with humans as we communicate with one another.

At present, human-machine communication relies upon speech synthesis methods that, while largely intelligible for most groups of listeners, lack naturalness. This is irritating for many listeners, but has been shown to directly affect intelligibility for some vulnerable groups such as the elderly [8] and those with dyslexia [9]. Furthermore, synthetic speech is known to degrade faster than natural speech in noise [10] which limits its use in certain environments. With the increased prevalence of synthetic speech in daily life, such as in smartphone assistants and public transport announcements, the development of more natural-sounding synthetic speech is a research priority.

One intuitive approach to reproducing natural speech is to attempt to recreate the system that produces it. This approach is known as physical modelling, as it aims to emulate the physics of the underlying vocal system. This research project aims to improve upon current physical modelling approaches to speech synthesis, by introducing a detailed three-dimensional geometry capable of dynamic movement between articulations. Physical models offer the most potential for natural-sounding synthesised speech, as a model of the

complete human vocal system is capable of any vocal output, including the aforementioned "uncontrolled high-pitched noises" which convey important information to the listener, despite their lack of linguistic content. However, research into physical modelling of the vocal tract is still in its relative infancy, and this project represents a step in the direction of a more complete, and therefore more human-sounding, physical vocal tract model.

People without a voice—perhaps lost due to degenerative disease, trauma or surgery—lose a vital means of communication that current speech synthesisers cannot perfectly reproduce. Even the most natural-sounding synthesis systems lack the flexibility of expression inherent in natural speech, and may not even match the gender or age of the person using them. These issues are more easily resolved in physical models, and the eventual goal of this area of study—of which this thesis represents only a small part—is to reunite patients with their *own* voice.

1.1 Hypothesis

1.1.1 Hypothesis Statement

The hypothesis upon which this thesis is built is as follows:

A detailed, three-dimensional dynamic digital waveguide mesh model of the vocal tract can produce more natural synthetic diphthongs than the two-dimensional dynamic digital waveguide mesh model.

1.1.2 Description of Hypothesis

Detailed, three-dimensional dynamic digital waveguide mesh

The digital waveguide mesh (DWM) is a numerical acoustic modelling technique which simulates acoustic wave propagation through a domain. The model may be constructed in any number of dimensions, with three dimensions offering the most realistic models, which can capture complicated ge-

ometrical detail, but with higher computational expense than one- or two-dimensional models. A *dynamic* DWM technique permits the effective shape of the modelled domain to change during the simulation, and has not previously been implemented in three-dimensional DWM models.

Vocal tract model

The human vocal tract runs from the larynx to the lips, and comprises a bent, soft-walled tube with a complex shape. The vocal tract acts as a filter, with a transfer function that changes as the tract moves to produce different vocal sounds. A vocal tract model is a system that aims to not only reproduce a realistic vocal tract transfer function, but also to accurately model the transitions between vocal tract shapes.

Natural synthetic diphthongs

Diphthongs are vowel sounds which are produced by moving the vocal tract. As such, they are useful for testing dynamic vocal tract models, which produce synthetic output that can be compared with real human speech. Naturalness is an important measure of synthetic speech quality, and cannot be measured objectively; instead it must be assessed by asking multiple human listeners to perform subjective judgement of the synthetic speech sounds.

1.2 Novel Contributions

The research on which this thesis reports has resulted in the following novel contributions to the field:

- A three-dimensional digital waveguide model of the vocal tract, based on three-dimensional magnetic resonance imaging data of the vocal tract, that is capable of movement during the simulation;
- A novel means of simulating the volume of air outside the mouth in

a digital waveguide mesh vocal tract model, making use of the head volume and approximately anechoic locally reacting walls;

- Validation of the proposed model, in objective and perceptual terms, compared with the existing two- and three-dimensional digital waveguide mesh vocal tract models;
- An objective and perceptual comparison of the proposed model with a state-of-the-art three-dimensional finite element method vocal tract model [11];
- A deep neural network based approach to controlling a two-dimensional digital waveguide mesh model based on input text, and producing intelligible speech output.

1.3 Statement of Ethics

The experiments presented in this thesis, and the management of corresponding data, were approved by the University of York Physical Sciences Ethics Committee, with reference numbers Gully151104 and Gully150217.

1.4 Thesis Layout

The remainder of this thesis is split into two parts, laid out as follows:

Part II: Literature Review

Chapter 2 introduces the acoustic background required for the rest of the thesis. This includes general acoustic concepts, the acoustics of ducts, the anatomy of the vocal tract, and the source-filter model, before going on to describe the different categories of sound made by the vocal system and how they are produced.

Chapter 3 presents an overview of the field of text-to-speech synthesis and the current state of the art. The different methods of synthesising a speech signal are described, followed by the applications of such synthesisers for assistive technology and commercial applications. Finally, approaches for evaluating synthetic speech are presented.

Chapter 4 describes the research into physical models of the vocal tract upon which the main contribution of the thesis is based. The implementation of each physical modelling approach—including the digital waveguide mesh—is presented first, and comparisons made between techniques. The application of these methods to vocal tract models is then described, and simulation details highlighted where relevant.

Part III: Original Research

Chapter 5 builds upon the digital waveguide mesh introduced in Chapter 4, and describes development and refinement of the MRI-based, three-dimensional dynamic digital waveguide mesh vocal tract model. Objective comparisons are made against recordings and established dynamic two-dimensional and static three-dimensional digital waveguide vocal tract models. Finally, the production of diphthongs using the proposed model is described, and the outputs are compared to recordings and the dynamic two-dimensional digital waveguide mesh; where available, outputs are also compared to those of a dynamic three-dimensional finite element method vocal tract model.

Chapter 6 In order to assess naturalness, perceptual tests are required, and this chapter first highlights some important test design decisions. Next, two pilot perceptual tests are introduced, which are used to determine the naturalness of a simplified three-dimensional digital waveguide mesh vocal tract model, and to asses the perceptual effects of using different sampling frequencies in the model. Finally, the naturalness of the proposed model is compared to that of established techniques.

Chapter 7 introduces a novel control method for the digital waveguide

mesh, based on a deep neural network, which uses written text as input. This chapter first presents the necessary background information on neural networks and the adaptations required to the digital waveguide mesh. This is followed by descriptions of the implementation and results of a first study using the technique, and discussion on improvements to the model.

Chapter 8 summarises the results of this study, and draws conclusions based on the original hypothesis, before presenting areas for future research.

Part II

Literature Review

Chapter 2

Acoustics of Speech

This chapter considers the essential background in acoustics required throughout the remainder of this thesis. The interested reader is directed to Stevens' book Acoustic Phonetics [12], which goes into each of the following topics, and other related areas, in greater detail than is possible here.

2.1 Acoustic Quantities

Sound waves act as local fluctuations in the pressure of a medium, which are transmitted to the ear in the form of longitudinal waves. The ear amplifies these small local fluctuations into what is known as sound. The local acoustic pressure p is defined as the difference between between the equilibrium pressure $\mathcal{P}_0$ at a point (x, y, z) in 3D space, and the instantaneous pressure $\mathcal{P}$ at the same point [13, p. 114]:

$$p(x, y, z) = \mathcal{P}(x, y, z) - \mathcal{P}_0(x, y, z) \tag{2.1}$$

As sound waves pass through a medium, the pressure variation at point (x, y, z) produces a time-varying waveform, $p(x, y, z, t)$.

Acoustic impedance Z is a measurement of how much a medium resists the flow of acoustic energy. It may be defined as the ratio of acoustic pressure,

p, to the particle velocity in the medium, v, as follows [13, p. 286]:

$$Z = \frac{p}{v} \tag{2.2}$$

Impedance Z is analogous to electrical resistance, p is analogous to voltage and v is analogous to current, so (2.2) is sometimes referred to as the acoustic Ohm's law.

Finally, the *specific acoustic impedance* Z_{medium} is a characteristic of a particular medium and can be useful in determining wave propagation between different media. It is defined as [13, p. 286]:

$$Z_{medium} = \rho_{medium} c_{medium} \tag{2.3}$$

where ρ_{medium} is the density of, and c_{medium} the speed of sound within, the medium. Both quantities are affected by temperature, so this must be taken into account when calculating Z_{medium}. Throughout the remainder of this thesis, the subscript 'medium' will be replaced with the name of the medium in question. For example, for air at 20°C, $c_{air} = 343\ \mathrm{m\,s^{-1}}$ and $\rho_{air} = 1.21\ \mathrm{kg\,m^{-3}}$, so $Z_{air} = 415\ \mathrm{Pa\,s\,m^{-1}}$ [13, p. 528]. The specific acoustic impedance of water at the same temperature is $1.48 \times 10^6\ \mathrm{Pa\,s\,m^{-1}}$.

2.2 The Wave Equation

As sound travels as longitudinal waves through the air, it is governed by differential equations relating to the wave speed. Assuming that planar sound waves propagate linearly along the x axis provides the 1D wave equation [14, p. 122]:

$$\frac{\partial^2 p}{\partial t^2} = c^2 \frac{\partial^2 p}{\partial x^2} \tag{2.4}$$

where $p = p(x, t)$ is some function of time t and distance x and c is the speed of sound in the propagation medium. A full derivation of 2.4 can be found in [14].

D'Alembert discovered that the wave equation could be solved using a *travel-*

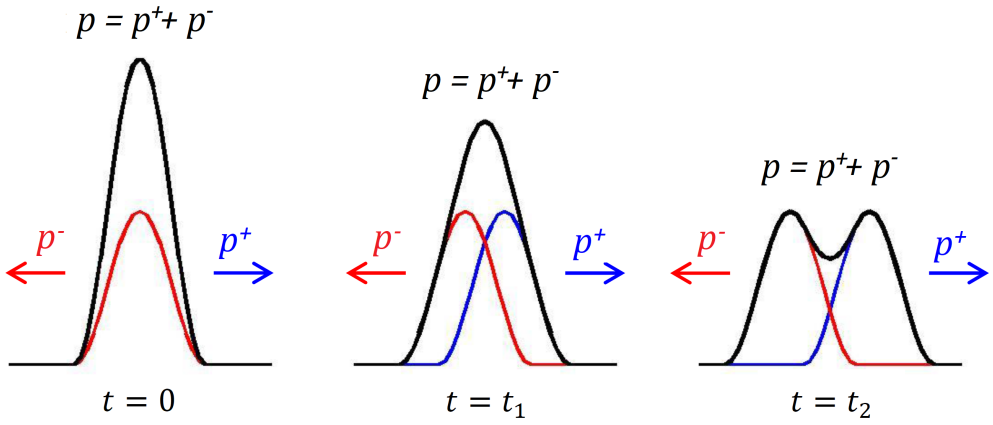


Figure 2.1: The travelling-wave solution to the wave equation. An initial input at $t = 0$ propagates through a 1D domain via travelling wave components p^- and p^+ which propagate left and right, respectively, throughout the domain as time goes on. Variable p is the sum of the travelling wave components.

ling wave solution, where $p(x,t)$ consists of two arbitrary twice-differentiable functions representing the right- and left-going components of the wave. This principle is illustrated in Figure 2.1 and described by the following equation:

$$p(x,t) = p^+(x - ct) + p^-(x + ct) \tag{2.5}$$

where p^+ and p^- are the right- and left-going travelling-wave components respectively. The function $p(x,t)$ may represent any physical variable, such as displacement on a string or velocity in air. The physical variable can be represented as the sum of p^+ and p^- at any point, which has important implications for sound synthesis because the propagation of the travelling wave variables can be represented using delay lines. The flexibility of this approach is explored further in Section 4.1.2.

The one-dimensional wave equation (2.4) can be extended to describe the behaviour of 2D (2.6) and 3D (2.7) acoustic waves [15]:

$$\frac{\partial^2 p}{\partial t^2} = c^2 \left[\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} \right] \tag{2.6}$$

$$\frac{\partial^2 p}{\partial t^2} = c^2 \left[\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} + \frac{\partial^2 p}{\partial z^2} \right] \tag{2.7}$$

where y and z represent the second and third spatial dimensions respectively, and p is a function of t and all relevant spatial variables x, y and z, depending on the dimensionality.

2.3 The Acoustic Duct

The two- and three-dimensional wave equations describe wave propagation in free space. However, within the vocal tract sound propagates in a tube. The acoustics of such a system can be approximated as a cylindrical duct. In such a system, the volume velocity, U, is typically used in place of particle velocity, v, in (2.2). Volume velocity is defined as the particle velocity multiplied by the cross-sectional area of the duct [16].

When describing the acoustics of a tube, it is common to assume that wave propagation in the tube is *planar*—i.e. that the wavefront propagates with the same speed throughout the tube—and therefore described by 2.4. This assumption holds for circular tubes while the highest frequency of interest, f_{lim}, satisfies the following [17, p. 84]:

$$f_{lim} < \frac{c}{1.71d} \tag{2.8}$$

where d is the diameter of the tube. The upper frequency for which this approximation holds in the vocal tract is usually given as 4000 Hz [18, p. 25] which is sufficient to model much of the behaviour of the vocal tract.

The *characteristic* acoustic impedance, Z_x, is calculated as follows:

$$Z_x = \frac{Z_{medium}}{A_x} \tag{2.9}$$

where Z_{medium} is the specific acoustic impedance of the medium from (2.2), and A_x is the cross-sectional area (CSA) of the tube at position x. For a tube with a circular cross section, $A_x = \pi r_x^2$, where r_x is the radius of the tube

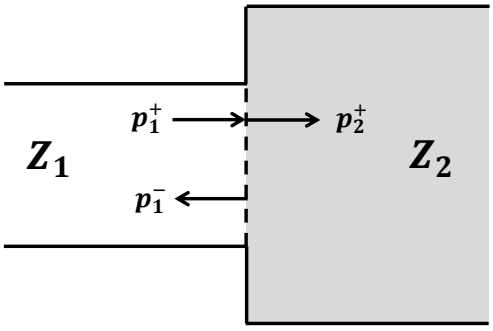


Figure 2.2: Scattering at an impedance discontinuity, after [16, p. 561]. Z_1 and Z_2 represent the characteristic acoustic impedances in tube 1 and 2 respectively. Right-going pressure wave component p_1^+ is incident on the boundary, and some is transmitted into tube 2 as right-going pressure component p_2^+, whereas some is reflected back into tube 1 as left-going pressure component p_1^-.

at position x. Therefore, a smaller cross sectional area results in a greater opposition to the flow of acoustic energy. The characteristic impedance of (2.9) allows the vocal tract to be approximated as a cylindrical duct of varying cross-section.

When a plane wave is incident upon an impedance discontinuity, for example caused by a change in the cross-sectional area of the duct, *scattering* occurs. This process is illustrated in Figure 2.2. The pressure must remain continuous across the discontinuity due to laws of conservation of energy and mass, as follows:

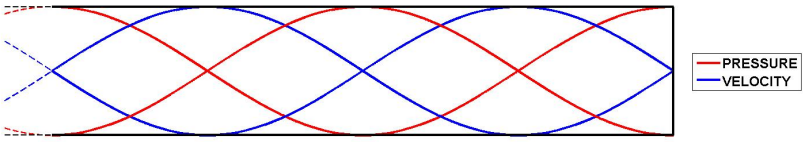
$$p_1^+ + p_1^- = p_2^+ \tag{2.10}$$

Following the derivation in [16, p. 561] eventually leads to the scattering relationships:

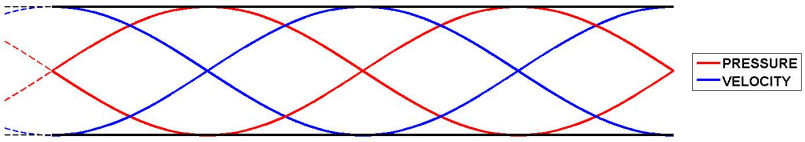
$$p_2^+ = (1 - R_{1,2})p_1^+ \tag{2.11}$$

$$p_1^- = rp_1^+ \tag{2.12}$$

where $R_{1,2}$ is a reflection coefficient at the boundary of tube sections 1 and 2, relating to the impedances of the first and second tube sections, Z_1 and



(a) Acoustics of a closed tube end



(b) Acoustics of an open tube end

Figure 2.3: Pressure and velocity at open and closed tube ends. At a closed tube end (top), pressure has a maximum and velocity has a minimum; at an open tube end pressure has a minimum and velocity has a maximum.

Z_2, as follows:

$$R_{1,2} = \frac{Z_2 - Z_1}{Z_1 + Z_2} \tag{2.13}$$

Some level of reflection will always occur at a discontinuity such as that in Figure 2.2, except where $Z_1 = Z_2$.

The end of the duct may be open or closed, and the type of termination determines the resonant frequencies that the duct is able to support. At a closed tube end, velocity is necessarily zero as air particles are unable to move immediately next to a rigid wall, while pressure is at a maximum. This is displayed in Figure 2.3a. In contrast, at an open tube end, pressure is at a minimum and velocity at a maximum as air may flow freely out of the duct, as seen in Figure 2.3b, although as expected at any impedance discontinuity, some pressure will be reflected back into the tube at an open end. Therefore, the type of termination defines the standing waves that can form in the tube and hence the resonant frequencies. The most applicable duct configuration for the case of vocal tract modelling is open at one end (the mouth) and closed at the other (the glottis), and the wavelength λ_n and resonant frequency f_n of the nth standing wave that forms in such a tube is

related to the tube length L by the following relationships [19, p. 43]:

$$\lambda_n = \frac{4L}{2n-1} \tag{2.14}$$

$$f_n = \frac{(2n-1)c}{4L} \tag{2.15}$$

Within the vocal tract, these resonances act to filter the input sound, which is described in detail in Section 2.6. The resulting resonant peaks that occur in the speech spectrum are known as *formants*.

2.4 The Vocal Tract Transfer Function

The typical adult male vocal tract is approximately 17 cm long, with a variable cross-sectional area of up to about 9 cm^2 [20]. As discussed in the previous section, resonances in the vocal tract have the effect of filtering any input signal. In the vocal tract, a volume velocity source produces an input signal u_{in}. The output of the system is a pressure wave p_{out}. The *vocal tract transfer function* (VTTF), $H(\omega)$ is given by [21]:

$$H(\omega) = \frac{P_{out}(\omega)}{U_{in}(\omega)} \tag{2.16}$$

where $P_{out}(\omega)$ and $U_{in}(\omega)$ are the Fourier transforms of the time-domain signals p_{out} and u_{in} respectively.

A vocal tract in a neutral position is approximately equivalent to a duct of constant cross-sectional area, so (2.15) applies and the VTTF features evenly-spaced peaks; for an adult male these peaks typically occur at 500, 1500, 2500 ... Hz [12], and the peaks correspond to formant frequencies in the speech spectrum. The effect of different vocal tract configurations on the frequency of these peaks will be discussed in Section 2.7.

The nasal cavity is an additional duct which is coupled to the vocal tract approximately halfway along its length, and acts as a side branch to the main vocal tract duct. A side branch is a cavity and as such has its own resonances.

Acoustic energy at or near the branch resonant frequencies will be absorbed into the side branch, causing a dip in the VTTF at these frequencies. This dip is known as an *antiresonance*.

One final concept that becomes relevant in acoustic ducts the size of the vocal tract is the *viscosity* of the medium, in this case air. Viscosity introduces losses into the VTTF, particularly at higher frequencies. The viscosity coefficient μ for air at 20°C is 1.85×10^{-5} Pa s [13, p. 528]. The effect of viscosity on the output of the vocal tract will be discussed in detail in later chapters.

2.5 Vocal Tract Anatomy

An outline of the vocal tract, representing a midsagittal slice through the head, is provided in Figure 2.4. The vocal tract is the tube that goes from the glottis, 1, to the lips, 11 and 12. Point 4 is the velum or soft palate, which acts as a flap that may be raised or lowered to control coupling between the vocal tract and the nasal tract, 7. The section of tract below the velum is referred to as the pharyngeal cavity, and the section of tract in front of the velum is known as the oral cavity.

2.5.1 Articulation

The shape of the vocal tract is governed by at least the 14 anatomic structures highlighted in Figure 2.4. However, many of these structures are fixed and do not move during the production of speech. The articulators that move to shape the sound output are listed below along with a brief description of their range of movement, following the control model developed at Haskins Laboratories [22].

Jaw the jaw moves on a hinge, so its motion describes an arc. Some lateral movement may occur, but this is assumed to be negligible.

Hyoid the hyoid is the cartilaginous structure that houses the larynx, within

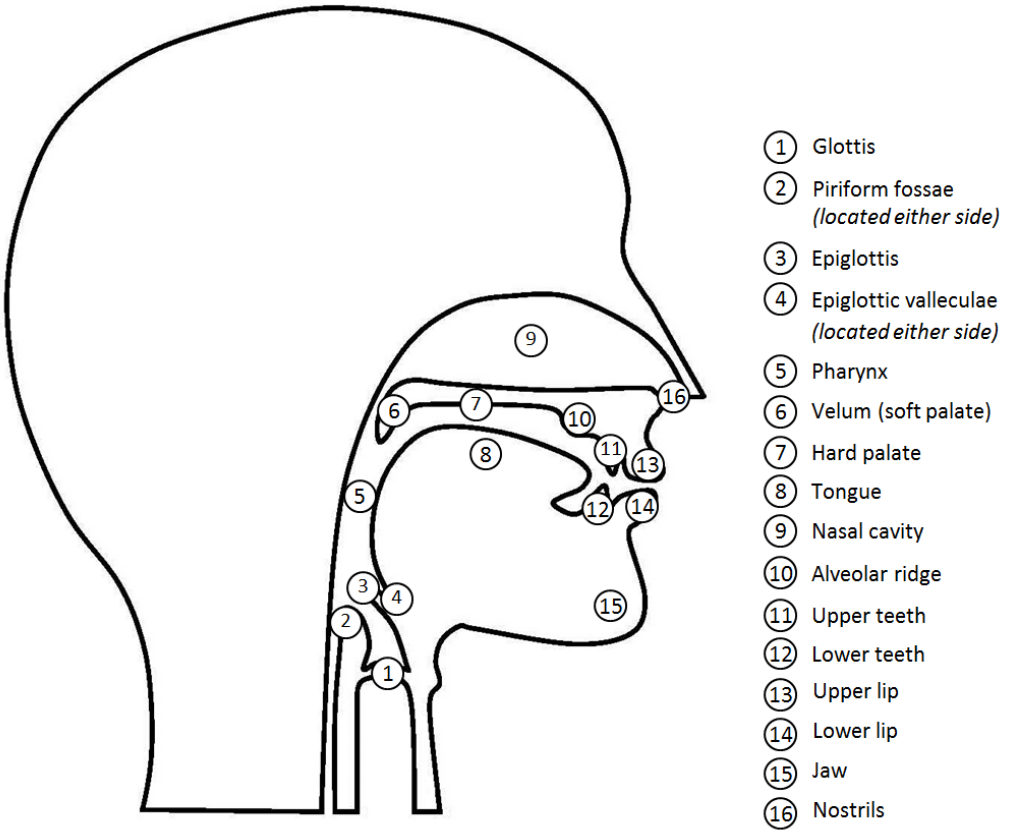


Figure 2.4: Articulators and points of interest in the vocal tract.

which the glottis is located. The hyoid can move up and down, changing the length of the vocal tract.

Lips the position of the lower lip can be specified relative to the jaw, as jaw motion will affect absolute lower lip position. Both the upper and lower lips may also change shape independently of other articulators, such as during lip rounding.

Tongue body the position of the tongue body may also be specified relative to the jaw, and can also move independently up and down, forwards and backwards.

Tongue tip the tongue tip is quite mobile but is attached to the tongue body so can be specified relative to that; and may additionally move forward and backward, up and down, and curl backwards.

Velum the tip of the velum may be considered to move in a straight line between its extreme open and closed positions.

The positions of all the other articulators may be determined from the above by imposing the properties of the relevant tissues onto the areas of the vocal tract illustrated in Figure 2.4.

2.5.2 Determining Vocal Tract Shape

The advent of medical imaging techniques has made it possible to look inside the head and neck and directly determine the shape of the vocal tract, with increasing accuracy as technology improves. A brief overview of the techniques used for vocal tract imaging and tracking of articulators is provided below.

X-Ray The earliest vocal tract studies to use medical imaging, such as [23], used x-ray images of the head and neck, providing the sagittal outline of the vocal tract. Relationships were developed to obtain approximate 3D cross-sectional areas from the 2D image (these are summarised in

[24]). Due to concerns with exposing subjects to radiation, x-ray studies have been superseded, but they were of considerable use in early vocal tract model development such as [25]. Recent related techniques include computed tomography (CT) scans which are based on x-rays. These provide very precise 3D detail [26], but due to the radiation involved they are not available for vocal tract imaging except where there is a medical need.

Ultrasound Ultrasound imaging techniques have been used with some success to determine tongue shape [27] by placing the transducer under the chin. However, the ultrasound signal does not provide useful information once it reaches a tissue-air boundary, and so cannot image the entire vocal tract, limiting its use for vocal tract studies. It does, however, provide some of the most detailed and dynamic tongue movement data available.

Magnetic Resonance Imaging (MRI) MRI is capable of producing detailed 3D scans of the vocal tract, and as such has been used extensively for vocal tract analysis and modelling since the early 1990s. MRI works by exciting hydrogen atoms (i.e. protons) which give off a detectable signal as they return to rest [28], so tissues with different relative hydrogen contents appear on the resulting image with varying brightness. Bone and teeth have a similar hydrogen content to air, making it difficult to identify the location of the teeth within the airway in MRI vocal tract data; some solutions to this have been proposed e.g. [29]. MRI data is used regularly for vocal tract modelling (e.g. [11], [30], and the work in this thesis). There are some shortcomings with the method, and these are described in Section 5.1.3.

Electropalatography Electropalatography uses a specialised sensor that attaches to the roof of a subject's mouth in a retainer-like device, to measure the location and duration of tongue contact with the palate [31]. As the technique does not offer information about the entire vocal tract and requires custom devices for each participant, it is not

commonly used as a data source for vocal tract modelling.

Articulography Articulography uses small magnets attached to the major vocal tract articulators and sensors to detect their positions. This information, although spatially sparse, contains important timing information about articulator movements, and can be used to train statistical speech synthesisers (e.g. [32]) or form control parameters for vocal tract models [33]. There are two types of articulography: *permanent magnet articulography* [32], which uses permanent magnets and outputs a mixture of magnet positions which can be used as input to a machine learning system; and *electromagnetic articulography* (EMA) which detects individual articulator positions [34] in three dimensions.

2.6 Source-Filter Model of Speech

The human vocal system can be conceptualised as a pair of connected, but independent systems: the *source*, and the *filter*. This simplification was originally proposed by Gunnar Fant [23], and although it has several shortcomings—as will be discussed later in this section—it has proven to be a very useful approximation to the real behaviour of the vocal system, and has led to many developments in speech analysis and synthesis. The basic principle of the source-filter model is illustrated in Figure 2.5.

2.6.1 The Voice Source

The vocal system is capable of producing several types of source signal. Those relevant to English, and many other languages, are the periodic signal produced by the glottis called *phonation*, and noise. The two source types may be produced individually or simultaneously, depending on the utterance.

Phonation is a periodic signal produced by the oscillation of the vocal folds, which rapidly collide, producing a wide-band sound, before being forced open by subglottal pressure, and then colliding again due to their own internal

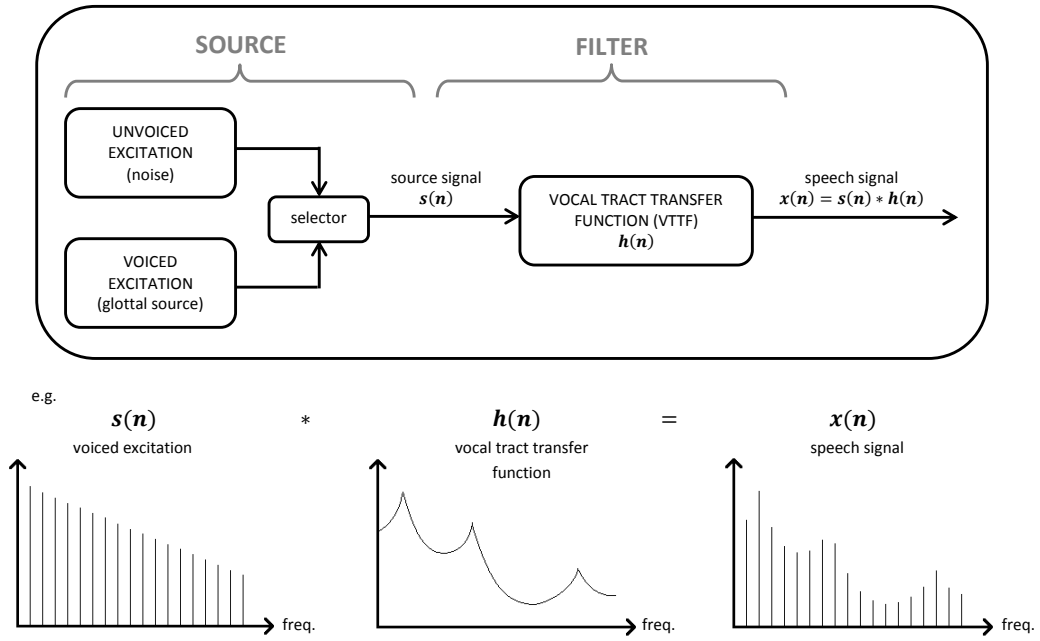


Figure 2.5: Source-filter model of speech, after [23].

forces. This process repeats to produce a periodic waveform representing the air passing though the glottis opening, known as the glottal flow waveform. A more detailed explanation of the anatomy and the phonation process is given in [12, p. 56]. The glottal flow waveform correlates with the open area of the glottis at any given time. The glottal area can be indirectly measured using a device called an electrolaryngograph (normally abbreviated to Lx) which consists of two electrodes placed on the outside surface of the throat at the larynx [35]. This device measures the resistance between the two electrodes, which increases when glottal open area is larger, producing a periodic signal that correlates with glottal area and hence glottal flow. An example Lx output trace is shown in Figure 2.6.

A number of models of the source waveform exist, notably the Liljencrants-Fant (LF) model [36] and the Rosenberg model [37], which allow synthetic glottal waveforms to be produced and studied. Each of these models produce a time-varying glottal flow waveform controlled by several input parameters such as frequency and amplitude [36].

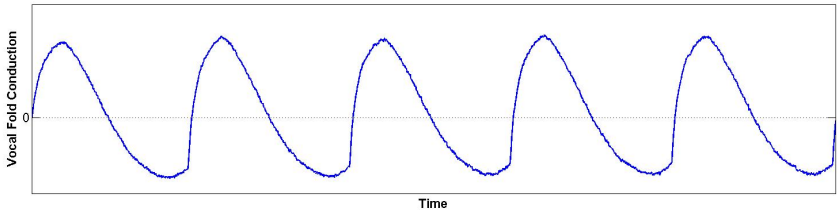


Figure 2.6: Electrolaryngograph (Lx) trace

The production of noise in the vocal tract occurs when the airway is sufficiently constricted that the Reynolds number, Re, exceeds a certain critical value, producing turbulent airflow. The Reynolds number is given by [12, p. 28]:

$$Re = \frac{vh\rho_{medium}}{\mu} \tag{2.17}$$

where v is the velocity of the air particles, ρ_{medium} is the density of the medium, i.e. air (1.14 $\mathrm{kg\,m^{-3}}$ at body temperature [18, p. 35]), μ is the viscosity of air (1.94×10^{-5} $\mathrm{kg\,m^{-1}\,s^{-1}}$ at body temperature [12, p. 27]), and h is the "characteristic dimension", i.e. the narrowest dimension of the constriction. The Reynolds number is a dimensionless quantity. The critical value of Re for the vocal tract airway is given variously as 1800 [38] or 2000 [12, p. 28], and when the flow in the vocal tract exceeds this value, turbulence is produced. It is apparent from (2.17) that the Reynolds number is proportional to the air velocity, which increases as it passes through a narrow constriction. Therefore if the constriction is sufficiently narrow, the velocity will be sufficiently high to exceed the critical Reynolds number, producing turbulence and hence noise downstream of the constriction. Constrictions below 30 $\mathrm{mm^2}$ in the vocal tract are often sufficient to produce turbulence [39, p. 84]. Where turbulent noise is produced in a relatively steady duct configuration it is known as *frication*, and when it is produced as a consequence of the duct rapidly moving from a closed to an open configuration it is known as *plosion*. The use of frication and plosion in the generation of speech sounds will be explored in Section 2.8.

In synthesis applications, turbulent noise is usually approximated as a band-pass-filtered white noise source injected at, or slightly downstream of, the

point of constriction. Phonation is usually approximated by a source model, such as the LF or Rosenberg models described above, or based upon a recorded signal such as Lx, and injected at the glottis location. Both phonation and noise sources are reasonably wide-band signals, making them suitable for filtering by the rest of the vocal apparatus to produce a wide range of sounds.

2.6.2 The Voice Filter

As described in Section 2.4, an acoustic duct has a number of resonances that form peaks in the VTTF and have a filtering effect on any input signal. The vocal tract is an irregular soft-walled duct, with a shape—and therefore a transfer function—that varies depending on the position of the articulators.

The filter part of the source-filter model is typically assumed to include the frequency-dependent effect of radiation at the mouth, which essentially acts as a filter with a 6 dB per octave roll-on characteristic [12, p. 128].

The effects of different articulator positions on the VTTF in relation to the sounds of spoken English are described in Sections 2.7 and 2.8. In addition to articulator positions, a number of other aspects have an effect on the VTTF. The nasal cavity may act as a side branch to the main vocal tract when the velum is open, introducing antiresonances in the VTTF as described in Section 2.4. Likewise, during nasal consonants like "mmm", the oral tract is occluded and forms a side branch to the main airway passing through the nasal tract. Other side branches in the vocal tract are known to introduce additional antiresonances. For example the *piriform fossae*, which are two small chambers located at the bottom of the pharyngeal cavity on either side, typically introduce antiresonances in the 4.5–5 kHz region [40], and the *epiglottic valleculae*, two small cavities at either side of the base of the epiglottis, also introduce antiresonances in a similar frequency region [30].

In addition to geometrical considerations, other characteristics of the vocal tract also affect the VTTF. The vocal tract walls are not acoustically hard but yielding, which increases the formant frequencies and bandwidths, in

particular for the first formant [12, p. 157]. Note that the impedance of the vocal tract is unlikely to be constant along its length, and may change between articulations—for example, as muscle tension in the tongue varies—which will affect the degree to which the yielding walls affect the formant frequencies and bandwidths. Coupling with the area outside the mouth, and with the subglottal system through the partially-open glottis, effectively increases the length of the vocal tract and reduces the formant frequencies [12, p. 154]. Finally, viscous and thermal losses at the walls of the vocal tract introduce losses proportional to the square root of frequency [41]. There are, therefore, a number of loss mechanisms related to the vocal tract that span the whole frequency range and affect the VTTF beyond the effects of geometry alone.

Where the source signal is located at the glottis, such as during phonation, the filtering effect of the entire vocal tract is taken into account. Where the source is located further forward, such as where frication occurs between the tongue tip and the teeth, the filtering effect is primarily due to the short section of vocal tract anterior of the constriction, as the high impedance of the narrow constriction effectively isolates the noise source from the posterior part of the vocal tract [12, p. 176]. In practice, however, this so-called 'back cavity' does have some effect on the VTTF and must be included to correctly model the consonant in question [12, p. 182].

2.6.3 Shortcomings of the Source-Filter Model

The major shortcoming of the source-filter model is its assumption of independence between source and filter: that the source is produced, and *then* it is filtered, and no interaction between the two processes takes place. There is ample evidence to suggest that this is not the case in natural voice production. For example, Stevens [12] describes a number of ways by which the mass-loading of the vocal tract on the glottis alters the glottal flow waveform, and shows that the effect is articulation-dependent, with a greater effect on the waveform during a constriction or closure in the vocal tract.

In addition to the presumption of independence, the influence of the *subglottal tract*—the airway below the larynx, including the trachea and the lungs—is usually neglected when implementing the source-filter model. The subglottal tract, however, has its own resonances and acts as a side branch to the vocal tract, albeit one whose coupling is more complicated during phonation due to the time-varying glottal opening. Coupling of the vocal tract to the subglottal system can modulate formant frequencies by up to 30 Hz and, when the glottis is open such as during "hhh", can shift the first formant frequency by as much as 100 Hz [12, p. 166].

Despite the shortcomings of the source-filter model, it is a useful starting point from which to understand the acoustics of speech. The vocal tract model proposed in this work is based largely upon the presumption that the source and filter *can* be considered independent, although discussion of how interactions between the two can be incorporated or approximated are included throughout.

2.7 Vowels

A *phoneme* is the smallest unit of speech which, if changed, affects the meaning of an utterance [39, p. 124]. For example, changing the "a" in "pat" to an "o" changes the meaning of the word, so "a" and "o" are phonemes. Phonemes cannot always be unambiguously represented by normal alphabetic characters, so they are represented using a special alphabet developed by the International Phonetic Association (IPA). Phonemes are denoted in text by slashes, e.g. /a/, to further differentiate them from normal written characters. The IPA symbols used to denote the phonemes of spoken English will be introduced throughout this chapter with pronunciation examples. There are two classes of phoneme, vowels and consonants, which have formal definitions beyond those commonly used among non-experts.

Vowels are a subset of the complete set of phonemes, and are defined in two ways. The phonetic description—relating to the acoustics of the vowels and their production—is that vowels are formed with a relatively open, stable

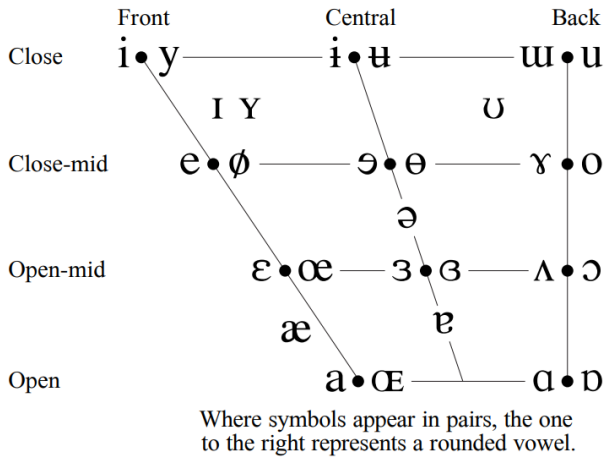


Figure 2.7: Description of vowels in terms of openness and frontage, from [43] under Creative Commons Attribution-Sharealike 3.0 Unported License (CC-BY-SA).

vocal tract configuration, with phonation as the voice source [18, p. 17]. The phonological definition—relating to the use of vowels within a language—is that vowels are *syllabic*; that is, they form the 'peak' of a syllable [39, p. 103]. Both definitions must be satisfied for the phoneme to be considered a vowel; one subset of phonemes known as the *semivowels* meet the phonetic definition of vowels but are not syllabic, and as such are classed as consonants and described in the next section.

2.7.1 The Vowel Quadrilateral

Vowels are characterised by their formant values, with the ratio of the first three formants critical for identification [42]. Higher formants are associated with naturalness and individual voice characteristics. The IPA provides a chart of all vowels describing their production, in terms of the 'openness' of the airway at its narrowest point, and the 'frontage' of the tongue i.e. how far forward in the vocal tract the narrowest constriction is formed. The IPA's diagram illustrating the complete range of vocal tract positions for vowels is provided in Figure 2.7, and is known as the *vowel quadrilateral*.

The edges of the vowel quadrilateral are determined by the *cardinal vowels*,

first presented in [44, pp. 36–37]. Each speaker will have their own distribution of vowels within the F1-F2 space, but the cardinal vowels indicate the extremes of vowel production, and all other vowels can be assumed to lie within the outline formed by the cardinal vowels. There are some arguments against the cardinal vowel system as originally proposed, especially that it "confuses articulatory and auditory properties" [39, p. 64], and that the parallel horizontal lines forming the vowel quadrilateral misrepresent the relative tongue height at different locations along the vocal tract and ignore the differences in tongue shape [39]. Nevertheless, the vowel quadrilateral serves as a useful approximation to vowel articulations and their distinction.

The vowels produced in spoken English are given in Table 2.1, along with their openness-frontage descriptions from Figure 2.7, and examples of their pronunciation. All pronunciation examples assume received pronunciation (RP), a standard British English accent sometimes known as the "BBC" British accent.

2.7.2 Static Vowels

During many vowel articulations, the vocal tract remains approximately static—that is, the articulators are held in a constant position—for the duration of the phoneme. Such vowels are known as monophthongs, with 'mono' indicating that one articulation is required. Monophthongs may be held for a long period without altering their meaning; for example, the /a/ in 'bad' may be long or short without changing the meaning of the word. The monophthongs used in English are given in the first section of Table 2.1, and typical values of their first three formants are also provided. These values are obtained from the speech of an adult male, but provided the ratio of formants is approximately correct, their absolute frequencies can vary without affecting vowel identification [42], as they do in the speech of women and children.

It is important to note that while monophthongs *can* be held static, this is not actually common in normal speech. For the most part, the vocal tract moves continuously and the 'static' articulations may be considered more

Movement	Type	Vowel	Formant values			Example
			F1	F2	F3	
static	monophthong	i	270	2290	3010	s**ea**
		ɪ	390	1990	2550	k**i**t
		ɛ	530	1840	2480	h**ea**d
		æ	660	1720	2410	b**a**d
		ə	-	-	-	**a**bout
		ɜ	490	1350	1690	st**ir**
		ɑ	730	1090	2440	h**ar**d
		ɒ	-	-	-	l**o**t
		ʌ	640	1190	2390	m**u**d
		ɔ	570	840	2410	l**aw**
		ʊ	440	1020	2240	f**oo**t
		u	300	870	2240	y**ou**
dynamic	diphthong	eɪ	-	-	-	d**ay**
		aɪ	-	-	-	h**igh**
		ɔɪ	-	-	-	b**oy**
		eə	-	-	-	f**air**
		ɪə	-	-	-	n**ear**
		ʊə	-	-	-	j**ur**y
		əʊ	-	-	-	sh**ow**
		aʊ	-	-	-	now
	triphthong	aʊə	-	-	-	h**our**
		aɪə	-	-	-	f**ire**

Table 2.1: English vowels, with pronunciation examples. Formant values are taken from [42] for an adult male where available and are not given for dynamic vowels, which transition between the values provided for static vowels.

like a target that the articulators approach [39, p. 255], before moving on to approach the next target. Nevertheless, static vowels provide a useful starting point for the development and evaluation of speech synthesis systems.

2.7.3 Dynamic Vowels

Dynamic vowels are those for which the vocal tract moves from one articulation to another during a single phonologically-defined vowel. They are therefore distinct from sequences of neighbouring monophthongs.

In English, the most common dynamic vowels are *diphthongs*, with the 'di' indicating that two articulations are involved. For example, the diphthong /eɪ/ in 'd**ay**' starts with an /e/ similar to that in 'b**e**d' and changes to an /ɪ/, like that in 'k**i**t', over the duration of the vowel. Any attempt to hold the /eɪ/ will result in either a long /e/ followed by a transition to /ɪ/, or a short /e/-/ɪ/ transition with a held /ɪ/; diphthongs are inherently dynamic in nature and while the transition can be slowed, they cannot be held in the same sense that monophthongs can. The diphthongs in RP English are given in the second part of Table 2.1, and the formant values can be assumed to transition between those of the corresponding monophthongs.

There are also a few examples of triphthongs (requiring three articulations) in English. Again, these are held phonologically distinct from three consecutive monophthongs or a monophthong-diphthong combination, with the difference usually manifest in the relative timing of the articulations. The English triphthongs are presented in the final section of Table 2.1.

Diphthongs and triphthongs are useful in the study of vocal tract models as they are dynamic—illustrating the potential of models to move between articulations—but with the relatively simple production method associated with vowels. As the majority of 3D vocal tract modelling techniques (including [30], [45] and the model proposed here) cannot yet reliably model consonant production, diphthongs are regularly used as demonstrators of the dynamic capabilities of a modelling approach.

2.8 Consonants

The term consonant encompasses all the phonemes produced by the human vocal system that do not fit the definition of vowels given in the previous section. Consonants typically—but not always—feature greater constriction in the vocal tract than vowels. There are a number of different mechanisms for the production of consonants, so the definition of different types of consonant is more involved than for vowels. As above, descriptions of the consonants will be split according to whether the articulations are static or dynamic.

2.8.1 Voice-Manner-Place

Consonants are defined using three descriptors: *voice*, *manner* and *place*. *Voice* refers to whether voicing (phonation) takes place at the glottis; for example, /s/ and /z/ have the same articulator positions, but /z/ is voiced and /s/ is not. Consonants usually feature a constriction along the vocal tract—sometimes more than one—and *place* refers to the location of this constriction. Finally, *manner* refers to the way in which the consonant is produced, and the manners applicable to English consonants are described in the following sections. Figure 2.8 is a chart produced by the IPA illustrating the majority of consonant sounds that can be produced by the human vocal tract[1] in terms of voice, manner and place. A list of the consonants used in English is given in Table 2.2 with voice-manner-place descriptors and pronunciation examples.

Some sources, e.g. [39], consider the voice-manner-place system to be insufficient for the description of consonant production, as the 'place' descriptor incorporates information about both location and articulator shape and position. For example, [39] proposes an additional descriptor, *stricture*, to describe the shape of the tongue at a constriction. However, the voice-manner-place system is considered sufficient for the current study, which uses only the subset of English consonants, avoiding much of the potential confusion

[1] Other sounds, such as the clicks used in languages such as Zulu, are defined elsewhere by the IPA and can be found at [43].

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

Figure 2.8: Description of consonants in terms of voice, manner and place, from [43] under Creative Commons Attribution-Sharealike 3.0 Unported License (CC-BY-SA).

that the stricture description was introduced to alleviate.

2.8.2 Static Consonants

Static consonants are those where the articulation can be maintained and the consonant will continue to be produced. As noted above, speech is inherently dynamic and these static configurations are unlikely to be held for any significant time during natural running speech; however, the static articulation acts as a target vocal tract shape and provides a convenient means of study.

The first class of static consonants are those most similar to vowels, and are called the *approximants*. The English approximants can be found in the first section of Table 2.2. Approximants, like vowels, are produced with a constriction in the airway that is insufficiently narrow to produce turbulence. A subset of the approximants are the *semivowels*, /j/ and /w/ in English, whose articulations alone would classify them as vowels (they are almost identical to the articulations for /i/ and /u/ respectively [39, p. 85]) but which are not syllabic like true vowels. The remaining English approximants are /ɹ/, which represents the 'r' sound in '**r**ight' and '**wr**ong' (note that /r/ represents a "rolled" r sound which is less common in English, /ɹ/ is used in most cases), and /l/, which is known as a *lateral* approximant as the tongue

Consonant	Manner	Place	Voice	Example
ɹ	approximant	alveolar	voiced	**wr**ong
j		palatal	voiced	**y**et
l		alveolar (lateral)	voiced	**l**ight
w		labial-velar	voiced	**w**on
m	nasal	bilabial	voiced	**m**ore
n		alveolar	voiced	**n**ice
ŋ		velar	voiced	ri**n**g
f	fricative	labiodental	unvoiced	**f**at
v		labiodental	voiced	**v**iew
θ		dental	unvoiced	**th**ing
ð		dental	voiced	**th**is
s		alveolar	unvoiced	**s**oon
z		alveolar	voiced	**z**ero
ʃ		postalveolar	unvoiced	**sh**ip
ʒ		postalveolar	voiced	vi**s**ion
h		glottal	unvoiced	**h**ot
p	plosive	bilabial	unvoiced	**p**en
b		bilabial	voiced	**b**ack
t		alveolar	unvoiced	**t**ea
d		alveolar	voiced	**d**og
k		velar	unvoiced	**k**ey
g		velar	voiced	**g**et
ʔ		glottal	unvoiced	foo**t**ball (‘dropped’ t)
ʧ	affricate	postalveolar	unvoiced	**ch**ur**ch**
ʤ		postalveolar	voiced	ju**dge**

Table 2.2: English consonants, with voice-manner-place descriptors and pronunciation examples.

forms a central obstacle around which the airflow must split laterally.

Another class of consonants with production characteristics similar to vowels are the *nasal* consonants. During the production of nasals, the oral tract is occluded, and the velum is lowered, so that the main airway passes through the nasal tract and the occluded oral tract acts as a side branch. The nasal tract has a fixed geometry, so the formants of the system are similar for all nasal consonants, and have higher damping than those of the oral tract [39, p. 252]. The location of the occlusion in the oral tract determines the length of the side branch, and hence the antiresonance characteristic of each nasal consonant: for /m/ this occurs around 1000–1200 Hz, and for /n/ at around 1600–1900 Hz [12]; for /ŋ/ the closure occurs close to the velum so no side branch is formed except very briefly when the occlusion is released. Like vowels, nasals and approximants typically feature phonation at the voice source.

The final class of static consonants are the *fricatives*, which are associated with frication noise as described in Section 2.6.1, and as such are produced by a constriction in the vocal tract narrow enough to generate turbulence. Most fricative consonants are produced by forcing the turbulent jet of air against an obstacle—such as the upper teeth or the hard palate—which produces significantly higher sound pressure levels (up to 30 dB higher [12, p. 101]) than are produced by constricted flow alone. This is easily demonstrated by comparing an /s/ or /ʃ/ sound with the turbulent noise made by simply pursing the lips and blowing: considerably more energy is required in the latter case to achieve a similar noise level. As noted in Section 2.6.2, the place of articulation—in this case, the location of the constriction—affects how much of the vocal tract is involved in the filtering of the noise source, and hence has a significant effect on the output sound.

The place of articulation for each fricative is defined as follows: *labiodental* fricatives /f/ and /v/ have a constriction formed by the upper teeth and lower lip; *dental* fricatives /θ/ and /ð/ require a constriction between the tip of the tongue and the teeth; *alveolar* fricatives /s/ and /z/ are produced with the tip of the tongue approaching the alveolar ridge; *postalveolar* fricatives /ʃ/

and /ʒ/ have a constriction between the tongue tip and the ridged area on the roof of the mouth posterior to the alveolar ridge; finally the *glottal* fricative /h/ occurs when air passes through a constriction formed by the partially-open vocal folds, and in a non-phonetic context is sometimes referred to as aspiration noise. Certain other fricatives may be present in different English accents or dialects, for example the voiceless *velar* fricative /x/ which is produced with a constriction between the body of the tongue and the velum, and occurs in the Scottish English pronunciation of the word "lo**ch**".

2.8.3 Dynamic Consonants

Dynamic consonants are those which inherently require movement of the articulators for their production. The first class of dynamic consonants are called *stops*, and are formed by a complete occlusion of the vocal tract and subsequent release of the closure. All stops used in English are egressive pulmonic stops—that is, they are formed using an out-going air stream originating at the lungs—and this class of stop consonants are called the *plosives* [39, p. 82] as they incorporate plosion as described in Section 2.6.1. During the closure, pressure builds up behind the occlusion, and when released the vocal tract forms a narrow constriction for a short time, producing a burst of turbulent noise. Particularly large or fast changes in the vocal tract shape during the release also introduce a transient burst of volume velocity preceding the noise burst [12, p. 117].

Plosives have three stages in their production: the *closure* phase, during which the articulators move towards the occluded position from rest or from the articulation of the previous phoneme; the *hold* phase, during which the occlusion is held; and the *release* phase, during which the occlusion is released and the articulators begin to move to the next position. In speech, the hold phase typically lasts 40–150 ms and the closure and release phases 20–80 ms [39, p. 82]. Formant transitions to and from those of the surrounding phonemes can be crucial in plosive identification, and this is discussed further in Section 2.9. Note that recent theories of speech control suggest that, while the closure location is important, the shape of the rest of the vocal tract is

not critical for plosive identification, and in fact the other articulators tend to transition from the shape associated with the previous vowel to that of the following vowel during production of the plosive [46].

As with fricatives, the location of the closure within the vocal tract determines the properties of the vocal tract filter and hence the characteristic frequencies associated with each plosive. A closure at the lips results in the *bilabial* plosives /p/ and /b/; a closure formed by the tongue tip pressing against the alveolar ridge produces *alveolar* plosives /t/ and /d/; a closure between the back of the tongue and the velum produces *velar* plosives /k/ and /g/; and a complete closure of the vocal folds results in the *glottal* plosive, commonly known as the "glottal stop" /ʔ/, which is heard regularly when the 't' is dropped in words like "foo**t**ball".

The next class of dynamic consonants are the *affricates*. Affricates are formed of a plosive followed immediately by a fricative, and act phonologically as a single consonant. There are two affricates used in English, both of which feature an alveolar plosive followed by a postalveolar fricative: /ʧ/ (unvoiced) and /ʤ/ (voiced).

There are three more types of dynamic consonant: *trills*, *taps* and *flaps*. These are described briefly here as they are not typically associated with English pronunciation but do occasionally occur. The trills are consonants that require constant vibration of an articulator against a surface, so despite being dynamic, trills can be sustained. Only a few articulators are sufficiently flexible to vibrate in this way. An example of a trill is /r/, which is commonly known as a "rolled r" sound. Taps and flaps are similar to trills in that an articulator strikes another, but a single strike rather than multiple strikes occur. Flaps occur when "one articulator strikes another in passing" [39, p. 86], while a tap is a "single deliberate movement to create a closure, tantamount to a very short stop" [39, p. 86]. Flaps may be heard in some articulations of "three" and "throw", where a flapped r is often used. Taps are heard, particularly for North American English, in words such as "atom", which become /ærəm/ [47].

2.9 Running Speech

It has already been noted that speech is inherently a dynamic signal, and that the articulators are rarely still; instead they move constantly between articulations. The individual phoneme descriptions given above serve as a convenient starting point for the study of speech, but such independent articulations are unnatural, and the context of a phoneme has a significant effect on its articulation in a number of ways. One way in which the context of a phoneme can influence the speech produced is by altering the timing. For example, in spoken English, vowels that occur before voiced fricatives and plosives are longer than those before voiceless consonants: compare the vowel lengths in 'feed' and 'feet', 'fad' and 'fat'. While this may appear "natural and inevitable" in English [39, p. 72], it is not necessarily the case in other languages. A number of other factors must be considered once analysis moves from single phonemes to syllables, words and phrases, and these are discussed below.

2.9.1 Coarticulation and Assimilation

The vocal articulators comprise muscles and other biological tissue that have inertia, and therefore a maximum speed at which they can respond to neuromuscular control signals. As a result, the articulation of phonemes is often affected by the articulation of their neighbours, in a process known as *coarticulation*. A common example is the different pronunciation of /k/ depending on whether it is followed by a front or back vowel: compare 'key' /ki/ with 'cork' /kɔk/ to hear the difference in /k/ as the tongue changes position to accommodate the following vowel. A summary of this process is given in [48, p. 9]:

> "[the vocal system] is producing its communicative artefact in 'real time'. It cannot move instantaneously from one target to the next. Rather than giving one phoneme an invariant articulation, and then performing a separate and time-consuming transition

> to the next, it steers a graceful and rapid course through the sequence. The result of this is coarticulation."

It is important to note that coarticulation is not due to 'lazy' speech but is inherent in the efficient use of vocal tract articulators having inertia. Furthermore, there may be some advantages [48, p. 9]:

> "the fact that the influence of a segment often extends well beyond its own boundaries means that information about the segment is available to perception longer than would be the case if all cues were confined inside its boundaries"

This is evident, for example, in the case of voiced plosives, where formant transitions from those of neighbouring vowels are often essential in phoneme identification [39, p. 255]. One way of understanding and modelling coarticulation, and running speech in general, is to consider the individual phoneme articulations described in the previous sections as 'targets', which the articulators aim to approach during speech, subject to the constraints of timing and the articulation of the surrounding phonemes.

Assimilation occurs when the coarticulatory effect is sufficient to actually change which phoneme is produced. A good example is "good morning" which, when spoken aloud, typically becomes /gʊbmɔnɪŋ/; that is, the /d/ that might be expected to occur is replaced with a /b/ to more easily accommodate the following /m/. Again, assimilation is not a product of 'lazy' speech but the human vocal tract compensating for its limitations due to the inertia of the articulators. As a result, coarticulation and assimilation effects must be accounted for in any vocal tract model aiming to reproduce natural speech.

2.9.2 Reduction and Casual Speech

Coarticulation and assimilation are effects that occur during all speech, even when carefully produced. *Reduction*, where words are "shortened in duration and produced with more centralised vowels" [47] is more likely in casual

speech, such as that between close friends in an informal setting, than in the formal speech recorded in a laboratory and typically used to study the vocal system. An example of reduction would be 'do you have to' becoming /djæftə/. A number of audio examples of reduction are available online [49].

In addition to increased reduction, casual speech has a number of other differences compared to careful speech. These are explored fully in [47] and include word choice, syntax, intonation, increased rates of assimilation, substitutions and deletions of phonemes (such as the flap /ɾ/ replacing /t/, as discussed above), and a greater variability in the production of certain sounds—for example, one cited study of casual speech reported 117 different pronunciations of 'that' and 87 of 'and'.

Casual speech comprises the majority of talking undertaken by people on a daily basis, so appreciating the nuances of casual speech is essential for the production of a natural-sounding speech synthesiser. This thesis focuses mainly upon individual phonemes and hence cannot yet address these concerns, however it is important to to be aware of the issues specific to casual speech for the development of speech synthesisers in the future.

2.9.3 Prosody

Running speech allows additional information to be encoded beyond the production of different phonemes. Variations in lexical stress, intonation, duration and volume are known as the *prosody* of an utterance, and are used to convey intent and emotional state. Prosodic features can also be used to clarify meaning, such as the rising inflection at the end of a question, which is recognised even if the utterance is arranged as a statement. Prosodic features are *suprasegmental*, which means they affect multiple segments of speech and may apply across an entire utterance.

Prosody encodes information at multiple linguistic levels, from the acoustic to the syntactic (phrase construction), semantic (intended meaning) and pragmatic (context-dependent knowledge) levels [50]. This is one reason why monotonous, robotic voices are so disturbing to humans: without variations

in pitch and timing due to prosody, much of the information usually available to listeners is removed.

2.10 Conclusion

This chapter has introduced the acoustics of speech production, from the basics of sound propagation to the complexities of running speech. The human voice may be described in terms of the signal produced, or in terms of the underlying mechanisms, and both of these aspects have applications to speech synthesis, as will be illustrated in the next chapter.

Despite phonemes being presented individually in this chapter, speech must be understood as a signal that is constantly changing, as the vocal tract moves from one articulation to another in order to produce a continuous stream of vocal information. Aspects of natural speech such as coarticulation and prosody must be considered in any system that aims to reproduce natural-sounding synthetic speech. The next chapter introduces synthesis systems that produce such continuous synthetic speech signals, and discusses the ongoing issues relating to the naturalness of the output speech.

Chapter 3

Text-to-Speech (TTS) Synthesis

The previous chapter described the acoustical principles relevant to the production of speech. These principles can also be leveraged in order to produce synthetic speech. For example, known acoustic characteristics corresponding to the speech signal can be reproduced by a synthesiser, or alternatively a model of the vocal system can be produced that will automatically produce appropriate acoustic output.

Text-to-speech (TTS) synthesis is perhaps the most common application of speech synthesis, and is defined as the conversion of written text into acoustic speech output. TTS synthesis is used for many applications including screen readers (where the text already exists as a book or website), smartphone voice assistants (where text is generated by a machine) and voice replacement technologies (where text is input by a user). Incorporation into a TTS system is the most likely end goal for any speech synthesis technique intended to integrate with modern technology. This chapter describes the different parts of a TTS system, introduces the most common TTS techniques and applications, and discusses the evaluation of such systems.

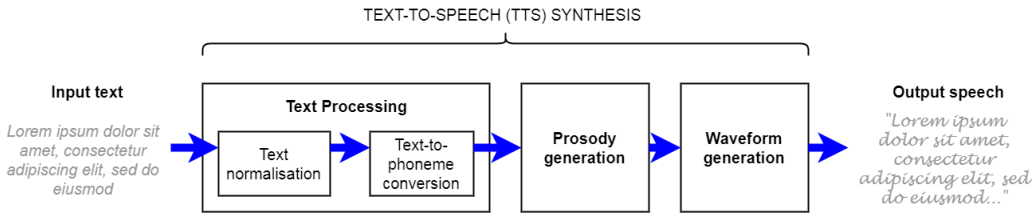


Figure 3.1: Components of a text-to-speech (TTS) synthesis system.

3.1 Components of TTS Systems

A TTS system must contain at least three basic components in order to convert written text to acoustic output, as illustrated in Figure 3.1: *text processing*, *prosody generation* and *waveform generation* [51]. Text processing converts the written text into a sequence of phonemes. Prosody generation refers to the generation of suprasegmental features such as pitch to illustrate meaning and intent. Finally, waveform generation concerns the creation of an acoustic waveform based on this phonemic and prosodic input information. Each of the three stages in a TTS system is subject to certain challenges.

The first stage, text processing, typically includes two intermediate stages: *text normalisation* and *text-to-phoneme conversion*. Text normalisation is the process of converting text components such as numbers and abbreviations into written words. This stage offers a number of challenges: depending on context, the abbreviation "Dr" might be written "doctor" or "drive", "747" might refer to an aeroplane ("seven-four-seven") or the number "seven hundred and forty-seven", "3/6" might be a fraction or a date, and the Roman numerals "VIII" might relate to "chapter eight" or "Henry the Eighth". There are many more such conflicts [52], and context-specific rules are required for each, to avoid mistranscription. Further challenges exist in languages such as Japanese that do not use spaces between words, with additional context-specific rules required to determine word boundaries.

After text normalisation, text-to-phoneme conversion is required, which also uses context-specific rules to obtain the correct pronunciation—consider "read" (past tense) and "read" (present tense), or the spelling and pronunciation of "cough" versus "plough". Tonal languages such as Mandarin Chinese also

require additional pronunciation rules. Both stages of text processing therefore require intricate and language-dependent rules, which may be hand-determined by an expert or, more commonly, automatically learned based on statistical information from a representative set of training data [52].

The next stage of a TTS system is prosody generation. As noted in Section 2.9, prosody can be essential for meaning and can also convey intent, emotion, and other paralinguistic information. Example prosodic features that should be considered by a TTS system include [52]: phrasing, lexical stress, pitch and intonation, duration, energy and voice quality; however, it may not be possible for every type of synthesiser to model every one of these features. Where prosody can be inferred from input text and punctuation—such as a rising inflection at the end of a sentence punctuated with a question mark—context-specific and language-specific rules can be developed or learned to determine the appropriate prosody. Outside of these cases, "default prosody" is often used [51], which incorporates the minimum variation in prosodic features wherever intended prosody cannot be clearly ascertained from input text. This approach prevents unnatural sounding output, as inappropriate prosody is obvious and disturbing to listeners, but results in speech that sounds unemotional and 'robotic' [51]. One approach that minimises the effect of incorrect prosody is to use a *limited domain* speech synthesiser, which only synthesises certain application-specific words or phrases. For example, transport announcements have a limited vocabulary and pre-determined phrasing, giving quite clearly-defined rules for prosody generation.

The text processing and prosody generation stages make up what is known as the *front end* of a TTS system [53]. The TTS front end receives written text as an input, and outputs a *linguistic specification*—a symbolic representation of the associated speech in terms of phonemes and prosody—for synthesis by the *back end* or waveform generation section of the system. There are, in general, two approaches used in the front end of a TTS system: *rule-based* and *data-driven* methods. Rule-based methods make use of hand-determined rules to determine the appropriate linguistic specification for any given input text. Rule-based methods are therefore either too simple for convincing

speech, or difficult to design, requiring significant time and linguistic expertise in the desired language. Data-driven methods, by contrast, make use of statistical or machine learning techniques, trained on speech recordings and their corresponding text, to learn transforms from text to linguistic specification [52]. While this approach does not require as much linguistic expertise or time-consuming identification of rules, it does require a large amount of reliable training data for every desired language, which may be difficult to obtain. Indeed, providing TTS systems for under-resourced languages has evolved into a large and complex field of study (see, e.g. [54]). There are a number of TTS front end systems in use, and the specifics of their implementation are beyond the scope of this work. The interested reader is referred to [52] which, although not yet complete, describes the various considerations of a TTS front end in detail, and provides examples using the popular Festival [55] front end system.

The final stage in a TTS system is the back end or waveform generation stage. This stage takes a symbolic representation of the desired speech as input, and outputs an acoustic speech signal. The generation of a speech-like acoustic output for a given phoneme has been a challenge since the earliest days of speech synthesis, and several of the most successful methods currently available are explored in the next section.

3.2 Waveform Generation Techniques

A number of different types of TTS system have been developed over the last 50 years. While the front end of each TTS system varies slightly, the focus of the current study is on waveform generation, so the back ends of the various TTS systems are considered in detail here. To be used as a TTS back end, a waveform generation technique must be capable of producing every phoneme in the language to be synthesised. It must also be able to accept the linguistic specification generated by the TTS front end as an input, and produce corresponding speech output. In the current context it is instructive to consider existing successful systems, as well as those currently

under development for which incorporation into a TTS system seems likely in the near future.

3.2.1 Formant Synthesis

Formant synthesis is the earliest of the modern TTS techniques, and is based on the source-filter model (Section 2.6), with the first such synthesisers introduced in the early 1950s [7]. Formant synthesis is essentially a subtractive synthesis algorithm, where a harmonically-rich source is passed through a filter representing the vocal tract transfer function (VTTF), obtained from the short-term spectrum of the speech signal. Peaks in the VTTF are known as *formants*, hence the name formant synthesis. The estimated VTTF is then reproduced, typically by a cascade of second-order band-pass filters—representing the formants—in series or in parallel [16, p. 455]. Estimating spectral parameters is not always easy, particularly when the fundamental frequency is high; where harmonics are spaced further apart, it is harder to resolve formant frequencies and bandwidths due to a lack of information in between harmonics.

The reproduction of the spectrum from short sections of the signal is possible if the vocal system is considered to be a linear time invariant (LTI) system system for the duration of the segment. This assumption is necessary for VTTF filter design, and is approximately true for short sections of speech on the order of $\sim$10 ms [56, p. 5], due to the inertia of the vocal tract articulators.

Improvements to formant synthesis vary according to the implementation details, such as the type and arrangement of filter units and the type of source signal used, as well as the introduction of antiformants due to nasal and subglottal resonances and the incorporation of period-to-period variations in the filter parameters [7]. An online demonstration of a simple formant synthesis technique for vowels can be found at [57].

Formant synthesis approaches have largely been superseded by concatenative approaches (see Section 3.2.2) in commercial speech synthesisers, although there is evidence to suggest that it remains more intelligible than other meth-

ods at high speeds, making it a preferred synthesis method for reading aids for the blind [4]. Additionally, formant synthesis is more controllable than concatenative synthesis, and this has been used with limited success in the generation of emotional speech [58]. One well-known example of formant synthesis still in use is the artificial voice of Stephen Hawking, who uses a formant synthesiser based on Dennis Klatt's DECTalk system [7]. From this example it is clear that formant synthesis approaches remain far from natural sounding.

3.2.2 Concatenative Synthesis

Motivated by the unnaturalness of formant synthesis, and with increasing computational power making more data-driven approaches feasible, in the 1980s focus shifted towards using segments of real speech for synthesis, a process known as *concatenative synthesis*. Instead of attempting to model the vocal *system*, concatenative synthesis takes a *signal*-based approach to waveform generation, aiming to replicate the speech signal by splicing together components of recorded speech in the time domain. Good concatenative synthesis remains the most natural TTS option currently available [59] because of this use of real recordings. The concept is simple, but the difficulty lies in selecting appropriate segments of speech and maintaining pitch, voice quality and other prosodic features across concatenated segments [52].

Each speech segment must be recorded from a real human talker, and therefore a practical concern is to keep the recording time and resulting database size to a minimum. For this reason, the type of segment must be selected with care to allow maximum flexibility. A natural choice might be a *phone* (an acoustic realisation of a phoneme), but using such static components of speech as the database element leads to difficulties in reproducing dynamic elements of speech such as the transitions between phones, and in maintaining a realistic prosodic contour.

The first widely-accepted concatenative synthesis method is known as *diphone synthesis*. A diphone is defined as the transition from the middle of

one phone to the middle of the next. Phones are more stable in the centre of their duration [52], so such an approach allows for smoother transitions between segments than can be obtained using phones alone. The use of dynamic, rather than static, segments also means that the transitions between phones are reproduced not just smoothly but realistically. The database of recorded speech segments should contain all possible diphones in the language under study: approximately the number of phones squared, although some combinations may be deemed impossible and excluded. It may also be desirable to include some context-specific phone variations to improve naturalness, increasing the database size further [52]. Diphones are typically recorded in a monotonous voice for consistency [58]. Diphone synthesis systems generally produce intelligible output, but as only one example of each diphone is stored, unnatural transitions and prosody remain a problem. An overall pitch contour may be imposed using signal processing techniques to improve prosody; this introduces some distortion, but remains more natural-sounding than formant synthesis [58].

Unit selection synthesis [60] is a more recent approach to concatenative synthesis. Unit selection permits the use of any length of speech segment in the database, for example phones [60], diphones, or even variable-length units [61]. The distinguishing feature of unit selection synthesis is that it uses automatic estimation procedures to select the most suitable segment from a database of recorded speech, which means that multiple versions of the same unit, in different prosodic contexts, can be stored and the most appropriate one chosen automatically for the current situation. Units may differ in length and linguistic value, and the automatic selection process chooses the most appropriate. As a result, unit selection synthesis can sound very natural, although when no suitable units are available it can also sound very unnatural [58].

Units are selected automatically by minimising two cost functions: the *target cost* (how well a candidate unit matches the desired segment) and the *concatenation cost* (how smoothly it joins with surrounding units) [59]. In [60], a feature vector is produced for the target and candidate units, containing ele-

ments such as pitch, duration, and discrete features such as vowel/consonant, voiced/unvoiced, and consonant type. The target cost is then calculated as a weighted sum of the differences between each of the features in the target and candidate unit vectors. Other methods, such as those described in [59], make use of acoustic distance measures which are linked to human perception to determine the target costs. The concatenation cost is typically a sum of sub-costs such as cepstral distances at concatenation points and differences in power and frequency [60], and point at which two units join may also be variable to promote the most natural join [62]. Neighbouring units in the database are assigned a concatenation cost of zero [60], promoting the use of naturally-occurring consecutive units where possible. This contributes to the high naturalness ratings achieved by unit selection systems.

Examples of both diphone and unit-selection synthesis are available online [55], where they can be compared with statistical parametric synthesis approaches (see Section 3.2.3). It is clear from this demonstration that unit-selection synthesis sounds significantly more natural than diphone synthesis.

Unit selection synthesis has a number of disadvantages. In order to obtain a large enough corpus of speech for a realistic output, and to minimise the number of bad joins, many hours of recordings are required and the resulting database can be very large, limiting the devices which can utilise it. Ideally, each unit would be recorded in every prosodic context, but it is simply impossible to record all possible variants and, where one context is under-represented in the database, bad joins and perceived unnaturalness are likely [59]. Similarly, the concatenated outputs are limited by the style and speaker of the original recordings, and in order to change speaker characteristics, or other aspects such as emotion [58], an entirely new set of recordings is required.

Despite these drawbacks, unit selection synthesis is the most common TTS technique in use today, for the simple reason that a good unit-selection system with a sufficiently large database remains more natural sounding than competing methods [59]. As a result, unit selection synthesis is found in most modern commercial speech synthesis systems, for example Siri on the

Apple iPhone and Amazon Alexa. However, due to the inflexibility of unit selection, in recent years the research focus has shifted towards the statistical parametric speech synthesis methods described in the next section.

3.2.3 Statistical Parametric Synthesis

In the simplest terms, statistical parametric speech synthesis (SPSS) represents the speech signal using *parameters* typically corresponding to pitch, duration and spectral information (although in theory any parameters could be used). This is combined with a training dataset comprising *statistical* information about those parameters in speech, such as mean values and variances. SPSS then uses a *generative model* to produce the speech parameters most likely to correspond to input text, essentially producing average parameter values from similar-sounding items in the training dataset [59]. As speech is actually generated from scratch rather than simply played back, the system has a lot of flexibility, as different talkers, emotions and even languages can be simulated by simply adjusting the synthesis parameters. Only the model is stored rather than the entire speech database, so SPSS also has a much smaller computational footprint than unit selection methods [59]. There are two major types of generative model used in SPSS systems: hidden Markov models (HMMs) and deep neural networks (DNNs). The overall outline of an SPSS system will be described first, and the details specific to each technique presented later in this section.

There are two stages in a SPSS system: training and synthesis. The training stage takes place off-line, and during this stage the generative model is provided with inputs and the corresponding correct outputs in order to learn the transform between the two. The input is the linguistic specification, as would be produced by the TTS front end, and for SPSS systems this is typically provided as as a vector of features for each phoneme in the input text, called a *linguistic feature vector*. The output is a set of parameters from which speech can be generated, such as source and filter parameters. Once the transform is learned, the model is saved and can then be used for the on-line synthesis stage, where speech output is produced from previously un-

seen input text. Speech is generated frame-by-frame, with frames typically on the order of 5 ms [63], and during each frame the synthesis parameters are assumed to be constant.

The key to the success of SPSS is the content of the linguistic feature vector. In addition to information about the current phoneme, the vector contains syllable-, word-, phrase- and utterance-level context information. A full list of features can be found in [10] and includes, for example, the position of the current syllable within the current word and phrase, the number of syllables in the preceding, current and next words, and the end-tone of the current phrase. As a result of this detailed context information, correct prosody is automatically learned by the model, provided the training database contains a sufficient range of contexts [53]. It is not possible for every context to be provided in the training dataset, and given the amount of detail captured by the context vector, each phoneme is likely to produce a unique context vector, so it is necessary that the model is also able to generalise to unseen contexts. Since SPSS methods generate speech from scratch, this generalisation is usually better than that which occurs in unit selection synthesis, which must either select the "best of a bad bunch" of candidate units, or introduce processing and hence distortion [59], when an unknown input encountered. The different approaches used by HMM- and DNN-based systems to determine the relevance of contextual features and generalise to unseen inputs are discussed later in this section.

Once output parameter sets have been produced by the generative model, they are passed to a synthesiser to produce the final output speech. Synthesis in SPSS systems is typically performed using a vocoder [53], which is similar to a formant synthesiser, so the parameters output by the model include information about the source (fundamental frequency, voicing information) and the filter (such as mel-frequency cepstral coefficients (MFCCs) or other spectral coefficients). Since SPSS produces averaged parameter values, a synthesis method must be chosen than behaves realistically even when parameters are interpolated or extrapolated. Linear predictive coding coefficients, for example, would be a poor choice as they do not guarantee a

stable system when interpolated [53]. In addition to the static parameter values output for each frame, the model also outputs dynamic parameters in the form of first and second derivatives of the parameter values, referred to as *delta* and *delta-delta* terms respectively, providing information about dynamic behaviour that is used to produce smoothly-varying parameter trajectories.

Hidden Markov Model (HMM) Synthesis

A hidden Markov model (HMM) is a finite-state automaton based on a Markov chain, which is a probabilistic model describing a sequence of events "in which the input sequence uniquely determines which states the [model] will go through" [64]. For the case of speech synthesis, the input sequence will be a series of linguistic feature vectors, and a *hidden* Markov model allows some *hidden* features that are not directly observable (such as the spectrum of speech that would be produced from a given input text) to be approximated based on variables that *are* observable (such as the current phoneme). A complete introduction to HMMs is beyond the scope of this work, but can be found in Chapter 6 of [64].

The field of automatic speech recognition (ASR) has used Markov chains to model speech for many years, and made use of this to develop high quality speech recognition algorithms using HMMs [65]. A similar approach for speech synthesis was proposed in [66], and has since become very popular, with [10] reporting that about 76% of papers at the major international speech technology conference INTERSPEECH in 2012 made use of HMM-based speech synthesis.

An overview of the speech synthesis process using HMMs is provided in Figure 3.2. The system takes a sequence of phoneme-level linguistic feature vectors as an input, and a corresponding sequence of context-specific phoneme-level HMMs is concatenated to produce an utterance-level HMM. Each phoneme-level HMM may consist of several states—typically 3 or 5 [59]—allowing various parts of the phoneme such as onsets and transitions to be modelled. An explicit state duration model [10], based on the mean

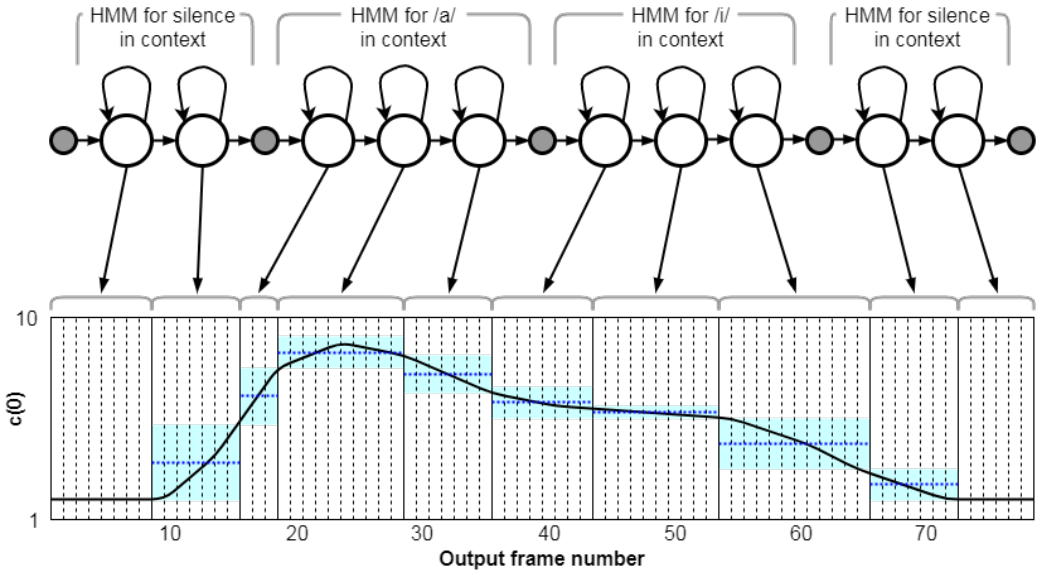


Figure 3.2: Generation of a parameter trajectory for a simple utterance /ai/ using HMM speech synthesis, using the first mel-frequency cepstral coefficient $c(0)$ as an example parameter. Vertical dotted lines represent frame boundaries. Horizontal dotted lines and shaded areas represent the mean and variance, respectively, of the Gaussian probability density function for each state. Image adapted from [10].

values of state duration distributions in the training data, is imposed[1] to determine how long the model stays in each state and therefore how many output frames are produced by each state. A maximum likelihood parameter generation algorithm then selects synthesis parameters based on the current state, while also taking into account delta and delta-delta coefficient values, to produce the smoothly-varying parameter trajectory illustrated in Figure 3.2.

The HMM method produces an individual HMM for each context-specific phoneme, but this introduces two issues. Firstly, not all context features are likely to be relevant in every phoneme realisation, so too many highly-specific models with very small training datasets will be produced. Secondly, the models will not generalise to contexts not represented in the training

[1]Technically, imposing an explicit duration model makes this a hidden *semi*-Markov model (HSMM), as true HMMs do not include duration modelling, but following convention in the literature this model will still be referred to as an HMM.

data. Both of these issues are addressed by *parameter tying*, where models that behave similarly in certain contexts are grouped together. To use an example from [53], if the phoneme preceding [ʃ] has no effect on its acoustic realisation but the next phoneme does, then the same model of [ʃ] can be used in all the contexts [aʃt], [ɪʃt], [ɛʃt], etc., and another, different model of [ʃ] can be used for all the contexts [aʃə], [ɪʃə], [ɛʃə] and so on. In practice this tying is performed by a decision tree that is automatically produced based on the available training data; a larger database allows more fine distinctions to be made [53]. Since this process is performed automatically, only contextual information relevant to acoustic realisation is retained, reducing complexity and providing a wider range of training data for each model. The decision tree also addresses the generalisation problem, as similar contexts can be identified and averaged to produce a good approximation to the true output parameters.

Although HMM-based synthesisers are not as natural-sounding as unit selection synthesisers and make several substantial approximations in their representation of the speech waveform, they have been sufficient to produce a very successful range of synthesisers with a number of clear advantages over unit selection synthesis. These are detailed in [10] and summarised below.

Variable talker characteristics: By training HMMs to produce an 'average' voice first, speaker adaptation can be performed using a very small training dataset to adapt the model to a new speaker. This has been applied for personalising voices for assistive technology applications [67] as well as facilitating a much wider range of voices than concatenative synthesisers can produce with their large database requirements [68].

Expressive synthesis: As speech is generated 'from scratch', rather than by replaying samples, the system is much more flexible and can be used to generate expressive speech more easily than unit selection synthesis [69].

Multi-language support: The HMM model itself is language-independent, so the only change required is to the relevant contextual features which

will be language-specific; given a suitable training database and suitable contextual features the automatic training procedure will 'learn' any language required [10]. Furthermore, HMMs permit cross-lingual speaker adaptation—providing a speaker of one language with their own voice in another language—by adapting an 'average' voice in one language with data from a speaker of another [70], and changes in dialect without the need for specific additional training data [71].

Singing synthesis: From an HMM point of view, singing may be considered as just another language [10], so with suitable context information and training data, singing can also be modelled using the HMM paradigm, such as the reasonably successful HMM-based singer 'Sinsy' [72].

Memory requirements: As only the model is stored, rather than the training database, HMM synthesis systems require significantly less memory than unit-selection systems which must store the whole speech database [10]. Furthermore, HMM synthesisers outperform unit selection synthesisers when the database size is small [53], allowing performance to be traded off against memory requirements depending on the application.

There are also a number of disadvantages to the HMM-based SPSS approach, and these are summarised in [59] as follows:

Vocoder: The vocoder used to synthesise speech from HMM-generated parameters makes a number of significant simplifications that affect the quality of the output speech. For example, a very simple pulse-train is typically used as the source model, producing "buzzy" output [59]; additionally the vocoder is based on a source-filter model of speech which, as discussed in Section 2.6, does not describe several important interactions in the vocal tract and contributes to the unnatural character of the synthesised speech.

Acoustic model accuracy: HMMs make a number of assumptions that may not be applicable for speech, for example that the input is a first-order Markov process, where the probability of the current state depends only on the probability of the previous state [64]. Furthermore,

HMMs require that state-output probabilities are conditionally independent on a frame-by-frame basis [59], which is not true for speech.

Over-smoothing of parameter trajectories: SPSS techniques essentially perform averaging of source and filter parameters, which leads to over-smoothing and a "muffled" quality in the output speech [59].

Several approaches have been attempted to minimise the problems listed above. For example, some of the quality issues associated with the vocoder have been somewhat alleviated by using an LF source model as opposed to a pulse train [73], and similar small improvements have been made to various aspects of the vocoder. The over-smoothing problem has been addressed by the inclusion of global variance [74], which matches the variance of the output parameters to those of the input speech and hence attains more realistic parameter contours. Finally, the acoustic modelling concern has been alleviated by the development of DNN-based SPSS systems, which are described in the next section.

Deep Neural Network (DNN) Synthesis

Neural networks have been used to synthesise speech since the 1990s, but it was not until the advent of fast parallel processors such as graphical processing units that the use of *deep* architectures became a viable option for speech synthesis [63]. The application of DNNs to speech synthesis was motivated by the need for a better acoustic model of speech than HMMs. The use of decision trees in HMM-based speech synthesis makes some complex context-dependencies (such as XOR) difficult to model [63], and breaks the training data down into small subsets with similar contexts, which can lead to over-fitting due to insufficient data. By contrast, the DNN architecture uses only one large model, rather than multiple smaller phoneme- and context-specific models, to transform linguistic feature vectors into synthesis parameters, and can therefore model complex dependencies and use all the available data for training.

A DNN speech synthesis system uses the same linguistic feature vectors and

synthesis parameter vectors as the HMM system described above: only the transform from one to the other is different. Unlike the HMM system, a single model is trained which accepts any linguistic feature vector and produces the corresponding output in terms of synthesis parameters. As before, the weights of the DNN model are trained by providing pairs of input and output feature vectors obtained from the training dataset, so it is necessary to use a dataset with sufficient context examples. The model is initialised with random weights, and then trained by altering these weights in order to minimise a mean-square error criterion until the model converges. Optimisation techniques such as stochastic gradient descent methods [63] are used to minimise this criterion.

As a single generative model is used for all possible phonemes and contexts, the training time for a DNN system is significantly longer than for an HMM system, and the resulting transforms are much harder to interpret [63]. Furthermore, no explicit parameter tying takes place in the DNN model. Nevertheless, experiments indicate that listeners prefer DNN-based synthesis over an equivalent HMM-based system [63], and all the advantages of flexibility that apply to HMM systems also apply to DNN systems. One study [75] indicates that the increase in naturalness is due to the removal of the decision tree structure, and the modelling of the output on a frame-by-frame, rather than state-by-state, basis. Techniques first used for HMM systems, such as the global variance technique described in the previous section, have also been used to further improve DNN systems [76].

The most recent advances in DNN-based speech synthesis have moved away from the traditional SPSS architecture, and take a more direct approach to modelling the speech waveform. For example, in [77], a DNN produces the parameters of a filter, rather than MFCCs, and these are then used to directly filter a pulse train or noise source to produce synthetic speech. This approach reduces the number of intermediate parametrisation steps, and hence may be more accurate, than MFCC-based approaches; initial results show that appropriate time-varying speech spectra are produced [77]. Google have recently gone a step further by introducing their WaveNet synthesiser

[78], which aims to reproduce *any* audio signal on a sample-by-sample basis using a convolutional neural network. Tests on TTS applications appear to suggest that WaveNet produces more natural sounding output, in both US English and Mandarin Chinese, than either state-of-the-art concatenative or SPSS methods. However, it should be noted that the paper [78] was not peer-reviewed, the statistical methods used to validate the naturalness scores are poorly reported, and the sample size for the perceptual tests is very small. Furthermore, the speech signals use a 16 kHz sampling rate—giving an 8 kHz usable bandwidth which removes a lot of the high-frequency energy associated with natural speech [79]—and although no run-times are reported in [78], anecdotal evidence suggests the system is very far from operating in real time. Nevertheless, such direct modelling approaches are sure to be an area of considerable future research interest.

While WaveNet represents the extreme of a purely data-driven approach to speech synthesis, it is widely believed that by incorporating knowledge about the vocal system into the synthesis method, natural-sounding synthetic speech can be achieved. This knowledge-driven approach is encapsulated in articulatory synthesis, explored in the next section.

3.2.4 Articulatory Synthesis

Articulatory speech synthesis is an umbrella term for a number of synthesis techniques that aim to approximate or model the physics of the vocal system, and to create synthetic speech by making articulations with the model in the same way humans generate speech. By modelling the entire articulatory system used for speech, theoretically any vocal sound can be synthesised, even in the absence of a recording or training signal.

It is generally agreed that articulatory synthesis, by virtue of modelling the actual human vocal system, offers the most potential for natural sounding speech of any of the methods discussed in this chapter [80]. Articulatory synthesisers have not yet been incorporated into TTS systems as they typically have a large computational cost, and suitable control methods to convert

text into articulatory parameters have not yet been developed. This section describes several approaches to the design of such a control method, and with ongoing increases in computing power it is anticipated that real-time articulatory synthesis will become the state of the art for TTS systems in the future.

The implementation of specific articulatory synthesis methods will be discussed fully in Chapter 4, however, relevant aspects of the models and their control systems are presented here in the context of TTS systems. In general articulatory synthesisers may be split into two broad types: vocal tract analogues, which approximate vocal tract behaviour by analogy, for example to electrical circuits; and physical models, which aim to reproduce the physics of the vocal system in its entirety. The computational expense of the former is significantly lower than that of the latter, so the majority of progress applicable to TTS has been made using vocal tract analogues.

Controlling Articulatory Synthesisers

In their review paper [81], Kroger and Birkholz define the three modelling tiers of an articulatory synthesiser:

1. the **control** model, a module for generating vocal tract movements;
2. the **vocal tract** model, a module to convert movement information into "a continuous succession of vocal tract geometries" [81]; and
3. the **acoustic** model, which generates a speech signal based on the resulting vocal tract shapes.

Much work has been done on (3), and more detail on these acoustic vocal tract models is provided in Section 4.2. It is the first two tiers of the articulatory synthesiser that are critical for driving articulatory models, using articulation data such as electromagnetic articulography (EMA) trajectories (see Section 2.5.2), and eventually text, as an input.

Some early work on articulatory synthesis in the 1970s achieved notable results before increasing computer power shifted the research focus towards

concatenative synthesis. Of particular interest are the models proposed by Flanagan et al. [82] and Coker [83], produced in 1970s, which took phoneme strings as input parameters, essentially forming complete rule-based articulatory TTS back ends. However—perhaps because TTS uptake at the time was poor due to a lack of sufficiently advanced front end systems—these models were never fully exploited.

Recently, work on control models for articulatory synthesisers have focused in more detail on theories of motor control during real speech. Two well-studied examples for articulatory synthesis are Story's vowel-substrate model [84, 46], where vocal tract movement is described as a series of vowel-vowel transitions upon which shorter-term consonant gestures are superimposed, and Birkholz et al.'s target-based model [85, 86], where articulations are defined in terms of articulator targets (some of which may be located beyond the limits of the vocal tract) and the time constant of a tenth-order dynamic system that models each articulator. These models are promising steps in the development of a articulatory TTS synthesiser informed by real human gestures, although as Birkholz states [87], "articulatory synthesis is not yet at a level of development where it is competitive for text-to-speech synthesis".

Hybrid Statistical-Articulatory Synthesis

In recent years, an alternative from of 'articulatory' synthesis has been developed, making use of statistical and machine learning methods to learn the transform between articulatory and acoustic parameters. These models are not articulatory in the sense that they directly model the vocal tract articulators; instead they use articulatory data, such as EMA trajectories, as input from which to produce speech output. To avoid confusion with synthesisers that model the vocal tract articulators, this statistical approach will henceforth be referred to as *hybrid statistical-articulatory synthesis*. The first such synthesiser appears to have been developed by Toda et al., [88] and was based on Gaussian mixture models. Since then, much like SPSS systems, HMMs [89] and recently DNNs [90] have been used to improve the articulatory-to-acoustic mapping. Similarly to SPSS systems, the acoustic

output is typically parametrised in terms of source and filter features, such as mel-frequency cepstral coefficients.

The development of hybrid statistical-articulatory synthesis systems was motivated by a desire to incorporate articulatory data into SPSS techniques for situations in which traditional SPSS training databases might be insufficient. For example, when a change of accent or articulatory style is desired compared to what is available in the database, the relevant articulatory changes may be known or predicted more easily than changes in the acoustic domain [90]. Presumably, hybrid statistical-articulatory synthesis is suitable for use with the articulatory control models described above, although such a combination has not yet been attempted. However, a more direct combination of statistical and articulatory techniques is possible by simply replacing the vocoder in an SPSS system with an articulatory synthesiser, as described in the next section.

Using SPSS to Control Articulatory Synthesis

One new and alternative approach to the development of complete text-to-articulation control models, or statistical articulatory-to-acoustic mapping, is to make use of the existing TTS capabilities of SPSS systems, which are known to generate smooth and realistic parameter trajectories (see Section 3.2.3). Because articulatory models are usually controlled by parameters relating to the size and shape of the vocal tract, interpolation between parameters should result in physically realisable vocal tract shapes and hence well-defined and stable vocal tract filter behaviour. This makes articulatory control parameters suitable for use in an SPSS system, with an articulatory synthesiser replacing the more usual vocoder. Instead of generating source and filter parameters, the generative model can be trained to produce control parameters for an articulatory synthesiser.

This approach was recently tested by the author [91], using a DNN-based synthesis system to control the parameters of the two-dimensional digital waveguide mesh vocal tract model described in Section 4.3. To the author's knowledge this was the first such combination of DNN-based and articulatory

synthesis. The model successfully generated intelligible speech despite very few constraints on the output parameters, illustrating the potential of this method for the control of articulatory vocal tract parameters. A detailed description of the system and results is provided in Chapter 7.

This combination of statistical and articulatory synthesis methods has a number of potential advantages over other synthesis methods. Firstly, by removing the vocoder, a primary source of unnaturalness associated with SPSS systems is removed [59], and replaced with a synthesis method believed to offer the most potential for naturalness [80]. Furthermore, the advantages of SPSS techniques described in Section 3.2.3 remain valid, with the added advantage that knowledge about the vocal tract such as size can be easily incorporated in order to transform talker characteristics.

3.3 TTS for Assistive Technology

Text-to-speech systems have a number of applications, but the area in which they have perhaps the most impact on the users' lives is in assistive technologies, designed to improve the quality of life for people with medical conditions.

3.3.1 Augmentative and Alternative Communication (AAC)

Patients who lose their voice, perhaps due to a degenerative disease or as a result of head or neck surgery, lose the most natural form of communication available to humans. The voice does not just convey linguistic information but information about our personality and emotional state [69], physical characteristics [92] and even subtle social indicators such as how we perceive our social status relative to that of others [93].

Patients with voice loss caused by degenerative and other chronic diseases are usually described as having augmentative and alternative communication

(AAC) needs. It is estimated that around 0.5% of the UK population has AAC needs [94], due primarily to nine conditions: dementia, Parkinson's disease, autism, learning disability, stroke, cerebral palsy, head injury, multiple sclerosis and motor neurone disease [94]. Depending on the severity of the condition and the abilities of the candidate, such patients may be assigned speech synthesisers known as *voice output communication aids* (VOCAs). Such synthesisers typically use TTS systems, although for many users the input of text is problematic and non-keyboard input devices, such as eye trackers, must be used. Recent research has proposed a number of alternative input paradigms that do not require text input—these are explored in Section 3.6—however, at present the vast majority of systems available to patients do require text as an input.

A further patient group that could benefit from synthetic speech are post-surgery patients—such as laryngectomy patients—from whom parts of the vocal apparatus have been removed, making natural speech generation difficult or impossible [32]. These patients have different needs to the patients described above as they can generally use the rest of their body normally, so typing out text for synthesis is very cumbersome. These patients typically use a device known as a voice prosthesis to replace the functionality of the larynx, and the rest of the vocal tract is used to filter this artificial source signal as usual. Synthetic speech systems have not typically been used for such patients due to the unsuitability of current text-based input methods. However, given a suitable mechanism to control a synthesiser using their normal vocal tract articulations, such patients could benefit from synthetic speech systems, and suitable devices are explored further in Section 3.6.

Both VOCAs and voice prostheses cause the patient to lose the naturalness of their voice, which can lead to feelings of stigmatisation [95] and loss of identity. There are therefore a number of approaches being developed to personalise TTS systems to a particular user. One such system is VocaliD [96], which is based on concatenative synthesis techniques. VocaliD takes a recording of the voice source signal from a patient and combines it with speech from a healthy 'donor' speaker of approximately the same age, gender

and geographical origin by inverse filtering to remove the contribution of the donor's source. A similar idea is used for the Speak:Unique system [97], which is an HMM-based system that uses voice banking to obtain a representative database of voices of all ages, genders and regional accents, and the most applicable are combined to produce a suitable average voice model matched to a particular target voice [67]. This approach was recently in the news when a sufferer of motor neurone disease was provided with a synthetic voice that matched his Yorkshire accent [98], a key aspect of his identity.

3.3.2 Other Applications

Although VOCAs and voice prostheses are obvious applications for synthetic speech, there are a number of other assistive technology applications that benefit from TTS functionality. One important example is screen readers for the visually impaired, which have a number of applications including auditory menus, audio books, and general information access. There is evidence to suggest that existing TTS systems used for this purpose are "more difficult to listen to" than natural speech [99], so more natural sounding synthetic speech based on articulatory synthesis may help to alleviate this problem. On the other hand, it may be more desirable that the synthetic speech is intelligible at high speed than natural at normal speeds, as many visually-impaired people choose to access auditory information with a faster-than-normal playback rate [4]. This illustrates some of the competing demands upon synthetic speech systems, and the importance of application-specific system design.

In addition to helping the visually impaired, TTS systems may also be used as reading aids for people with learning difficulties who struggle to take in information visually. With the increasing naturalness of synthetic speech systems, a further area of application is in virtual speech therapy for patients with speech disorders [100], allowing them to practice language skills at their leisure.

3.4 TTS for Commercial Applications

Assistive technologies have traditionally been a key area of application for synthetic speech systems, but with the increasing prevalence of technology in daily life, commercial applications for synthetic speech systems are becoming a significant driving force in the development of new synthesis techniques. Several broad application areas are summarised below, although this list should not be considered exhaustive. Many, although not all, of these applications also rely on speech recognition technology.

Human-computer interaction (HCI): recent development of smartphone assistants, such as Apple's Siri, has made TTS systems more prevalent in daily life. HCI extends beyond smartphones to interactions with all kinds of machines and has particular applications in robotics and artificial intelligence.

Real-time translation and language training: language training systems that can correct a student's pronunciation in their own voice are known to improve language learning [101], and real-time translation systems that allow users to communicate in another language are becoming increasingly prevalent [70].

Educational applications: beyond language education, speech synthesisers facilitate learning in a diverse range of areas including psychology, linguistics and voice science [102].

Creative and musical applications: such as singing synthesisers [72].

Hands-free information delivery: from transport announcements and in-car navigation systems to audio descriptions for visually-intensive tasks.

Although there are some examples of commercial systems that use HMM-based synthesis [10], and there has been significant recent interest in DNN-based synthesis methods from large technology companies such as Google

[78], at present commercial text-to-speech systems still typically use unit-selection methods, for reasons discussed previously. While unit selection usually sounds quite natural, occasionally bad joins are formed which can severely impact the perceived naturalness. There is evidence to show that, for certain groups of listeners such as the elderly [8], those with dyslexia [9] and non-native speakers, these unnatural joins impact intelligibility. Furthermore, synthetic speech is known to degrade faster in noise than natural speech [10], affecting intelligibility for all groups of listeners. Depending on the application, these issues may cause irritation or, in the hypothetical case of an emergency announcement made with a synthetic voice in a noisy environment, may actually be dangerous. It is therefore essential that intelligibility is evaluated in real use conditions during the development of commercial speech synthesis systems.

3.5 Evaluating Synthetic Speech

Early speech synthesis systems were primarily concerned with intelligibility; however, synthetic speech has improved to such a degree that intelligibility is, in many cases, as good as recorded speech, at least in the absence of noise [10], and attention has now turned to making synthetic speech as similar to human speech as possible. This concept is often termed *naturalness*, and in [103] it was concluded that intelligibility and naturalness are indeed separate dimensions in the assessment of synthetic speech quality, and are by far the most perceptually relevant dimensions for quality judgements. This section introduces the many aspects of naturalness that have been studied in the literature, and considers how best to measure the naturalness of the present system in order to address the hypothesis.

3.5.1 Naturalness

Naturalness is an *ill-defined* concept. Humans seem to innately know whether speech was produced by a human or a computer, but it is much harder

to define exactly how they make this judgement, or to assess naturalness objectively. There are a lot of factors that influence the perceived naturalness of an utterance, and there is evidence to suggest that the brain processes these in a "gestalt-like" way [104]; that is, the results of decisions about the naturalness of various different aspects of speech are combined in one overall judgement of naturalness, and it only takes one of these aspects to be slightly unnatural for the whole utterance to be judged unnatural. It has been shown that humans are unreliable in estimating the naturalness of a single acoustic property of speech such as intonation [105]. It is not inconceivable that this hypersensitivity to unnatural or unexpected characteristics of speech may have developed as an evolutionary tactic to identify intruders to social groups or tribes. Whatever the reason, it must be accounted for when considering how to assess the quality of synthetic speech.

As [106] notes, "there is no such thing as 'optimum' speech", and the internal standard to which humans compare synthetic speech to determine its naturalness "cannot be fully described by the physical proximity to natural speech". Humans appear to have an internal criterion for naturalness that is not well understood, and therefore is difficult to measure.

From the point of view of synthetic speech, naturalness may be split into two categories: *linguistic* and *acoustic* naturalness. While both of these factors contribute to one overall perceptual idea of naturalness, there is evidence that the brain processes these two different streams of information separately: information related to linguistic naturalness is "processed in a dorsal stream in the posterior part of the auditory cortex and caudal neural pathways," while information on acoustic naturalness is "processed in a ventral stream along the [superior temporal sulcus]" [104]. Historically, the synthesis systems described earlier in this chapter have been most concerned with linguistic naturalness, as the foremost concern of synthetic speech, after being intelligible, is to be understood. Linguistic naturalness remains crucial today, and is mostly related to the TTS front end and its rules for producing a linguistic specification from input text.

However, as speech synthesis systems improve, acoustic naturalness is becom-

ing more and more important. Acoustic naturalness refers to the characteristics of speech that do not convey linguistic information, but instead carry information about the talker. These aspects include features like average pitch of the voice, breathiness, nasality, and other such individual characteristics. Acoustic naturalness effectively encompasses everything that occurs in the human vocal system to produce a natural utterance if the linguistic content is assumed to be perfectly natural. As a result, acoustic naturalness is primarily determined by the TTS back end, or waveform generation technique, and as such is most pertinent to the current study.

Evaluating Acoustic Naturalness

Due to the various dimensions of naturalness in synthetic speech, some of which are still not well understood, measuring acoustic naturalness presents a significant challenge. Ideally, some objective measure would be used to provide an overall naturalness rating, facilitating direct comparison between synthesis systems.

Recently, attempts have been made to develop a suitable objective naturalness measure. For example, in [106], a detailed study of the factors influencing perceived speech quality—and by extension naturalness—is performed, with the aim of combining these factors into a single quality rating. The authors conclude that with sufficiently rich and varied training data based on many different synthesis methods, an objective measure may be developed, but that such a measure will be non-linear in nature. Other objective measures of speech quality have been tested in [107], for example signal-to-noise ratio and linear predictive coding (LPC) based methods, as well as the perceptual evaluation of speech quality measure put forward by the ITU-T [108]. However, [107] found these measures to be unsuitable for predicting the results of subjective quality tests: that is, the results did not match with human judgements of naturalness.

Another approach to objective naturalness measurement might be to compare the waveform generated by a speech synthesiser to that of recorded speech, but as [105] notes:

> "it *is* possible to make a direct comparison between acoustic characteristics of a target utterance (what the synthesiser has been asked to produce) and acoustic characteristics of the same utterance spoken by a human speaker. [...] However, this ignores the fact that speech is highly variable (utterance-to-utterance, speaker-to-speaker, etc.) and that there are often many acceptable ways of producing a single utterance [...] The perceived quality of a synthetic speech utterance is clearly not, therefore, simply a matter of the degree to which the physical characteristics of the utterance match the physical characteristics of one natural speech utterance."

Indeed, while synthetic voices that are personalised to match a specific voice are desirable in the context of AAC devices or natural machine translation, waveform similarity is not a suitable measure of global naturalness as it serves only to determine how well one specific recording is reproduced by the synthesis system.

In the absence of a suitable objective measure for the assessment of naturalness, subjective methods must be considered instead. Indeed, even if a suitable objective measure existed, the nature of the concept of naturalness would surely require subjective tests to be performed alongside such assessments. As [52] puts it, "The only real synthesis evaluation technique is having a human listen to the result. Humans individually are not very reliable testers of systems, but humans in general are". Therefore, it is necessary to consider the design of a subjective test that will best assess the naturalness of a synthesised speech sample.

The first consideration in designing any test for naturalness must be how naturalness is to be quantified. Although it is difficult to determine how naturalness is perceived by humans, the term appears to be sufficiently well understood that it is acceptable to ask even non-expert listeners whether a speech sample sounds natural, such as in [103], or whether it was produced by a human or a computer, as was done in [109]. Many studies have used similar techniques, which, depending on how the question is presented, may result

in a simple binary decision (natural / unnatural), or a position on a discrete or continuous scale between the two extremes. Additional measures that have been proposed for the assessment of naturalness include likeability and believability [51]. These factors are of particular concern in a full synthesis system which involves long-term interaction: as [51] puts it, "if people do not like a synthesiser's voice [...] they will soon cease to use it."

Instead of directly asking participants how natural or pleasant a speech sample sounds, several papers have proposed alternative approaches to measuring naturalness. For example, [110] takes the position that naturalness is equivalent to ease of understanding, a somewhat naïve view given the many non-linguistic information sources encoded in a truly natural speech signal. However, extensions of this idea, such as [111, 112] and [113], equate naturalness with the degree of cognitive load required to process speech, with the implicit assumption that humans are "tuned" to natural speech and will therefore require minimum cognitive load to process it. These studies quantify cognitive load by measuring reaction times to questions about the origin of the samples presented, and find that synthetic speech does indeed require a longer reaction time (although the system under test was DECTalk, a system which is significantly less natural than those currently available). It was also found in [111] that test subjects could correctly identify a word in natural speech after hearing 67% of its duration, while 75% was required for a word of synthetic speech. Tests such as these could therefore be used to give quantitative values to the concept of naturalness.

It is important to consider the fact that naturalness should be measured on a *relative* basis [51]: it is difficult for a human to assign an absolute naturalness score to a sample without something to compare it to, and for most synthetic speech applications, the concern is whether a system sounds *more natural than* alternative systems and, perhaps one day, *as natural as* real human speech.

As naturalness is a fundamental aspect of the present hypothesis, the definitions and approaches summarised in this section are essential considerations when testing the proposed system, and will be explored further in Chapter 6.

3.5.2 Other Evaluation Criteria

The most commonly reported evaluation criterion for synthetic speech systems is *intelligibility*. As noted above, most synthesis systems are now sufficiently intelligible that naturalness is a better metric against which to judge synthesis quality, however intelligibility remains an important baseline for performance. Intelligibility is easily measured using, for example, word error rate (the proportion of words incorrectly identified by a listener).

As discussed in Section 3.3.1, an important aspect of speech synthesis systems for AAC applications is how closely a patient's own voice can be approximated. Such a measure is not necessary for every application of synthetic speech, but is very important for AAC users. To the author's knowledge, there is currently no agreed metric for measuring how successful such personalisation has been, but it is possible to measure acoustic similarity to the target waveform (if such a waveform is available) based on spectral distance measures such as errors in formant frequency and magnitude. A further concern specific to AAC applications is the practicality of the system: it must have a suitable input device based on the capabilities of the user, and be sufficiently portable.

One final consideration for synthetic speech systems are the hardware requirements, including the running time and computational footprint. As discussed in previous sections, a trade-off between simulation quality and computational expense may be necessary depending on the application.

The above is not intended to be an exhaustive list of evaluation criteria for TTS systems, but an overview of some aspects to consider depending on the intended application. Another application-specific consideration is whether text is a suitable input mechanism for the synthesiser in the first place.

3.6 Alternatives to TTS

As touched upon in Section 3.3.1, the requirement that text exists before synthetic speech is generated may severely limit some users, and there are

many situations where text is an unnatural or unnecessary intermediate step in the production of synthetic speech. This section discusses some of the alternatives to TTS. None of these methods appear to be widely available at present, but with the continued increase in computational power, they may come to supersede TTS for certain applications in the future.

One emerging area of research in the field of human-computer interaction (HCI) is known as concept-to-speech synthesis. Since the natural language generation component of an HCI system produces "an abstract representation of the sentence to be spoken" [114], converting this representation to text and then using a TTS system to synthesise speech is inefficient and is likely to introduce errors at the prosody-generation stage. While an ideal concept-to-speech system might map concepts directly to speech, in practice an intermediate TTS system is used [114], but knowledge is shared with the prosody-generation stage, which helps to resolve ambiguities [115].

Similar approaches bypassing text have been implemented for AAC applications. Such devices are typically referred to as brain-computer interfaces (BCIs), and may take several forms, all of which broadly aim to access neural signals controlling speech and use these to control a speech synthesiser instead. A review of BCI systems is given in [116]. Studies considering the development of suitable articulatory synthesis systems for integration with BCIs [117] find that synthesisers with around 10 degrees of freedom are suitable for control by BCI. Instead of directly accessing the neural pathways used to form speech, another approach known as 'silent speech' makes use of the articulatory movements directly, by permanently attaching sensors to the tongue and lips [32]. This data is then used to control an articulatory synthesiser.

One different approach to speech synthesis for AAC applications makes use of the acoustic signal from disordered speech, from patients with moderate to severe dysarthria occurring as a result of degenerative disease. This approach is called a voice-input voice-output communication aid [118], and uses speech recognition to build a message which is then synthesised, with mixed success.

3.7 Conclusion

The text-to-speech systems presented in this chapter range from highly simplified system-based approaches such as formant synthesis, to intricate signal-based approaches using machine learning techniques that require large databases of text and speech to learn transforms between the two. Each method has strengths and weaknesses, and one conclusion of this chapter is that, at present, no synthesis method can be considered the 'best' for every application.

Articulatory synthesisers offer a great deal of potential for natural-sounding synthetic speech that may one day fill this role, aiming as they do to reproduce the behaviour of the human vocal system itself. The next chapter describes the range of physical vocal tract models that may be used to inform articulatory synthesis, with the goal of producing natural-sounding synthetic speech.

Chapter 4

Physical Vocal Tract Models

In the previous chapter, different approaches to speech synthesis were introduced, with articulatory synthesis shown to offer significant potential for the production of natural-sounding synthetic speech. Physical models are the basis of articulatory synthesis, and this chapter considers the variety of approaches that are available for developing a physical model of the vocal system.

Physical models of the vocal system aim to reproduce the complete physical behaviour of the system, summarised in Chapter 2. This behaviour encompasses subglottal activity, oscillation of the vocal folds, wave propagation in the vocal tract, physical behaviour of the tissues comprising the vocal tract articulators, and radiation at the mouth and nose. In practice, these elements are often separated, following the source-filter model, for easier computation. This chapter discusses physical models of the filter, i.e. the vocal tract, ranging from the simplest one-dimensional model that approximates the vocal tract as a set of concatenated cylinders (see Section 2.4), to complex three-dimensional systems that accurately reproduce the detailed 3D geometry of the vocal tract.

In the 1960s to 1980s a number of studies investigated one-dimensional physical vocal tract models in some detail, and with a variety of control mechanisms, but interest in physical models appears to have waned with the advent

of concatenative synthesis. In more recent years, attention has returned to physical modelling approaches due to their potential for naturalness and intuitive gestural control, especially given the increased computational power now available. A large number of physical vocal tract models have been developed over the past 70 years, so this chapter is not exhaustive; it focuses primarily on those approaches that provide context for the proposed model or detail about parameters, control methods, loss mechanisms and other implementation details that are relevant to this proposed system. Note that mechanical models that reproduce the physics of the vocal tract are also available (e.g. [119]), but these do not provide implementation details relevant for the study of computational physical models, so are not considered here.

This chapter first provides an overview of the acoustic modelling methods used to simulate wave propagation in the vocal tract. The application of these techniques in the literature to increasingly complex vocal tract models is then described, before methods of evaluating such models, and the challenges that still exist for vocal tract modelling, are presented at the end of the chapter.

4.1 Physical Modelling Techniques

4.1.1 Transmission Lines

A section of an acoustic duct with constant cross-sectional area can be represented by an analogous electronic circuit, known as a *transmission line*, with resistances, capacitance and inductance that correspond to the viscous and thermal losses, compressibility, and inertia of the air in the duct respectively [18]. The equivalent structures are illustrated in Figure 4.1.

By considering the vocal tract as a series of short concatenated cylinders[1], as described in Section 2.4, it is possible to build up an equivalent elec-

[1]In practice non-circular cross-sections may also be implemented by including a "shape factor" [18] that quantifies the difference in the ratio of the cross-sectional perimeter and area to that of a circle.

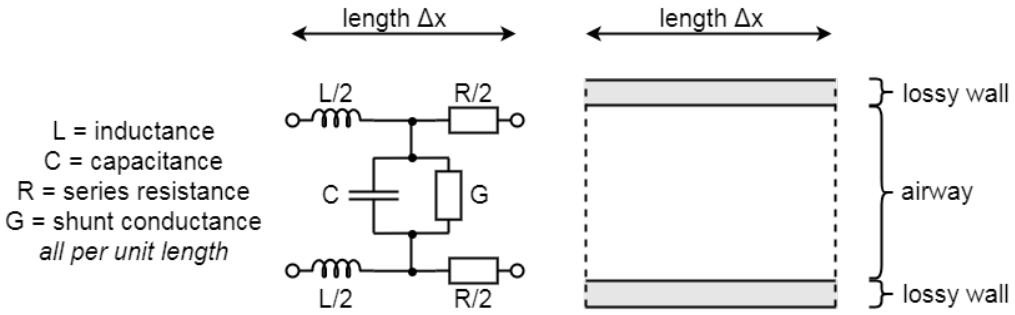


Figure 4.1: Electronic circuit (left) analogous to section of lossy acoustic duct (right) with fixed cross-sectional area and lossy walls, after [18].

tronic circuit of concatenated transmission line segments that represent the vocal tract. Radiation characteristics are simulated with a suitable lumped impedance at one end of the transmission line [82], and glottal flow simulated with a current source at the other, since in the analogy current is equivalent to velocity and voltage is equivalent to pressure. Additional losses can be implemented by including extra resistances, capacitances and/or inductances in each section or at specific points along the transmission line model. The result is a complete vocal tract model such as those discussed in Section 4.2.1.

4.1.2 Reflection Lines

A different approach, also based on the approximation of the vocal tract as a series of concatenated tubes, is known as the *reflection line* or wave-reflection model. The reflection line model makes use of the d'Alembert solution to the wave equation (2.5) to represent wave propagation in a duct as a sum of left- and right-going wave components, with reflection and transmission between tube segments of different cross-sectional areas taking place as illustrated in Figure 4.2. The parameter R_k, which controls the reflection between tube section k and tube section $k+1$, is calculated as follows [25]:

$$R_k = \frac{A_k - A_{k+1}}{A_k + A_{k+1}} \tag{4.1}$$

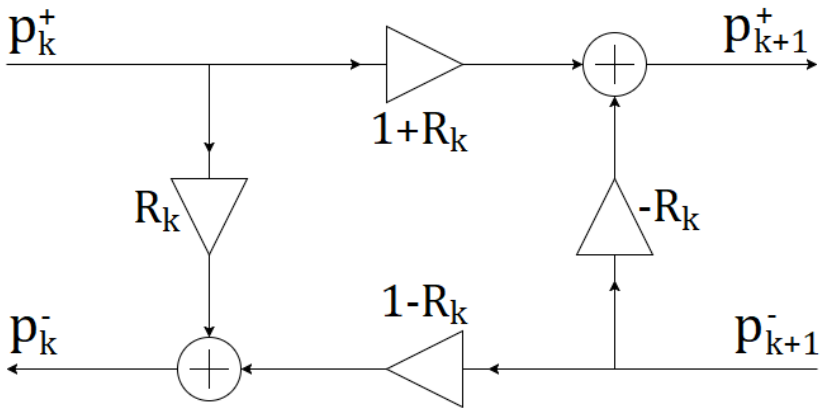


Figure 4.2: Scattering of wave variables in a reflection line model, after [25]. Superscript + represents the right-going and superscript − the left-going travelling wave components of acoustic pressure p in tube sections k and $k+1$.

where A_k is the cross-sectional area of the kth tube segment. As such, the reflection line model is completely defined by the cross-sectional areas of the vocal tract along its length. Unit delay operators are applied to travelling wave variables between scattering junctions to simulate wave propagation over time.

One key implementation difference between transmission and reflection line representations of the vocal tract is that in the former, the length of each section can be varied independently, while the latter is implemented using unit delays between scattering junctions, so section lengths are constant and are related to sampling frequency. However, the reflection line method has a much lower computational cost than the transmission line method [120], leading to its early popularity for speech synthesis applications [25].

4.1.3 Finite-Difference Time-Domain (FDTD) Models

The *finite-difference time-domain* (FDTD) approach is a method for approximately solving differential equations, and has been used in many fields including electromagnetics [121] and seismology [122]. When applied to acoustics, the FDTD method is used to approximate the wave equation (2.4)–(2.7),

so may be used to simulate wave propagation in any number of dimensions depending on the equation selected. For this reason, it may be applied to anything from a simple cylindrical-duct approximation of the vocal tract to a detailed 3D vocal tract model.

Recalling that the differential of a function $f(x, y)$ is defined as the limit where Δx and Δy tend to 0, a finite difference operation simply replaces the differential operator with a *difference* operator having a non-zero value of Δx (and/or Δy), thus approximating the differential with a series of discrete 'steps'. The value of the differential at any x can be approximated as the average of x and $x - \Delta x$ (backward difference), x and $x + \Delta x$ (forward difference) or $x - \frac{1}{2}\Delta x$ and $x + \frac{1}{2}\Delta x$ (centred difference).

Applying finite differences to the 1D wave equation simply requires setting:

$$\frac{\partial^2 p}{\partial x^2} \approx \frac{p_{k+1}^n - 2p_k^n + p_{k-1}^n}{\Delta x^2} \tag{4.2}$$

$$\frac{\partial^2 p}{\partial t^2} \approx \frac{p_k^{n+1} - 2p_k^n + p_k^{n-1}}{\Delta t^2} \tag{4.3}$$

where p_k^n represents the instantaneous acoustic pressure at time step $t = n\Delta t$ and spatial location $x = k\Delta x$. This represents a centred difference, applied twice to approximate the second differential. Substituting this approximation into (2.4) and rearranging to find the future sample p_k^{n+1} gives:

$$p_k^{n+1} = (2 - 2\lambda)p_k^n + \lambda^2(p_{k+1}^n + p_{k-1}^n) - p_k^{n-1} \tag{4.4}$$

where

$$\lambda = \frac{c\Delta t}{\Delta x} \tag{4.5}$$

is a dimensionless parameter known as the Courant number [123, p. 131]. The same process may be applied for higher dimensionality wave equations (2.6) and (2.7), yielding similar approximations for 2D and 3D wave propagation. For stability the Courant-Friedrichs-Lewy (CFL) condition must be satisfied:

$$\lambda \leq \frac{1}{\sqrt{D}} \tag{4.6}$$

where D is the number of dimensions in the system (i.e. 1, 2 and 3 for 1D, 2D and 3D systems respectively). In practice this places an upper limit on Δt for a given Δx, so FDTD models have an inherent lower bound on computational complexity for a given spatial resolution. Furthermore, FDTD schemes are prone to numerical dispersion error which affects the audio signal above approximately $0.1 \times f_s$ [124], placing further requirements on the minimum sampling frequency and hence computational cost.

4.1.4 Digital Waveguide Mesh (DWM) Models

The digital waveguide approach [125] is based directly upon the reflection line model described above, with acoustic variables modelled as the sum of bidirectional travelling wave components propagated through delay lines. The digital waveguide method can be extended to multiple dimensions in the form of the *digital waveguide mesh* (DWM). In the DWM technique, unit-length delay lines are connected at *scattering junctions*, similar to those illustrated in Figure 4.2, but with multiple input and output branches. The scattering junctions are arranged regularly, most commonly in a Cartesian grid.

The DWM algorithm consists of three stages, which must be completed at every time step, n, for every scattering junction, J, in the mesh [126]. The first stage is the scattering of travelling wave variables:

$$p_J(n) = \frac{2\sum_{i=1}^{N} Y_{J,J_{nei}} p^+_{J,J_{nei}}(n)}{\sum_{i=1}^{N} Y_{J,J_{nei}}} \tag{4.7}$$

where $p_J(n)$ is the acoustic pressure at scattering junction J at time step n, $Y_{J,J_{nei}}$ is the acoustic admittance of the waveguide connecting junction J to neighbouring junction J_{nei}, $p^+_{J,J_{nei}}(n)$ is the pressure incident at junction J from junction J_{nei} at time step n, and N is the number of neighbouring scattering junctions connected to junction J y waveguides: for a 2D rectilinear mesh, $N = 4$ and for a 3D rectilinear mesh, $N = 6$.

The second stage is to calculate the outgoing pressure from a junction into

its neighbouring waveguides. The junction continuity expression states that the total junction pressure is the sum of the incoming and outgoing travelling wave components:

$$p_J(n) = p^+_{J,J_{nei}}(n) + p^-_{J,J_{nei}}(n) \tag{4.8}$$

where $p^-_{J,J_{nei}}(n)$ is the pressure output by junction J into the waveguide connecting it to junction J_{nei}. Therefore, the outgoing pressure may be calculated as:

$$p^-_{J,J_{nei}}(n) = p_J(n) - p^+_{J,J_{nei}}(n) \tag{4.9}$$

The final stage is to introduce a unit delay between adjacent scattering junctions, so the outgoing wave variable from one junction becomes the input to a neighbouring junction at the next time step:

$$p^+_{J,J_{nei}}(n) = p^-_{J_{nei},J}(n-1) \tag{4.10}$$

Equations (4.7), (4.9) and (4.10) describe the complete DWM update procedure. The DWM has been shown to be equivalent to the FDTD technique operating at the CFL limit, and as such it is subject to the same issues of dispersion error. However, the DWM scattering equation (4.7) introduces the free parameter admittance, which presents a simple way of modelling heterogeneous media. This will be explored further in Chapter 5.

4.1.5 Transmission Line Matrix (TLM) Models

The *transmission line matrix* (TLM) method arose from transmission line models just as the DWM arose from reflection line models, and the two methods are also equivalent [127]; as such the TLM is also equivalent to the FDTD method at the CFL limit and subject to the same limitations as above. The TLM method uses different terminology to the DWM and replaces (4.7), (4.9) and (4.10) with matrix operations, but is otherwise identical. Details of the TLM method can be found in [127].

The FDTD, DWM and TLM methods are therefore all equivalent methods for modelling acoustic wave propagation in any number of dimensions, and the

choice of which to use is motivated by the requirements of the simulation. All three methods use a regularly spaced grid to discretise the domain of interest, commonly a Cartesian grid with spacing dependent upon the system sampling frequency and the stability condition (4.6). A Cartesian grid is conceptually simple to understand, but may lead to inaccuracy if the domain being modelled has a complicated shape, as non-rectilinear domain shapes can only be approximated.

4.1.6 Finite Element Method (FEM) Models

The *finite element method* (FEM) is another method for approximating differential equations based on splitting the simulation domain into smaller sections. However, unlike the above methods, in FEM there are no restrictions on the shape or uniformity of these 'elements', making them suitable for modelling complex domains and permitting greater simulation detail in areas of interest. Furthermore, the FEM permits frequency-dependent behaviour to be implemented, which is difficult to achieve in the purely time-domain FDTD, DWM and TLM methods. The consequence of this increased accuracy is significantly increased computation time [30].

As for the FDTD-based models described above, the basic concept of FEM is that the equations governing behaviour in a complex domain—such as the wave equation—can be calculated across small subsections of the domain and then combined to produce an approximate solution across the entire domain. The difference between FEM and the FDTD-based methods is that these small subsections of the domain may not all be the same shape and size, so the calculations required in each will differ slightly. Therefore, instead of a single operation performed multiple times throughout a domain as in an FDTD simulation, an FEM simulation produces a large set of simultaneous equations that must be solved. As a result the FEM is more accurate than FDTD-based methods but requires significantly more computation time.

The first stage of FEM simulation is to break the domain down into elements. These may be 1D, 2D or 3D depending on the dimensionality of the simula-

tion. The corners of each element are referred to as *nodes*, and neighbouring elements share nodes. The governing equation is solved for these small, simple elements, and the resulting functions are constrained to be continuous at the nodes (i.e. between neighbouring elements). A major constraint on the accuracy of FEM models is the distribution and size of the elements, so "meshing" a domain is an important and time-consuming stage in the simulation process. FEM models are typically quoted as requiring 8–10 nodes per wavelength of interest [45], so, like FDTD models, the desired bandwidth of a simulation is a consideration when determining the size of the elements and hence the complexity of the simulation. The FEM has been well studied for vocal tract modelling, and various FEM vocal tract models are presented in Sections 4.3 and 4.4.

4.1.7 Boundary Element Method (BEM) Models

If the problem domain has well-defined boundary conditions, the *boundary element method* (BEM) may be used. This converts the problem to a surface integral over the boundary, which is then solved for smaller, simpler elements—again, subject to continuity constraints—and combined to produce an approximate solution [128], much like in the FEM method. The BEM has similar advantages to the FEM in terms of modelling irregular shapes and frequency-dependent behaviour, and furthermore may be more efficient, as for a large 3D domain only the boundary is considered, resulting in many fewer elements than in FEM [128]. However, the BEM has only rarely been applied to vocal tract modelling, as will be discussed in Section 4.4.

4.1.8 Choosing a Modelling Approach

This section has discussed several of the different modelling approaches that have been used to produce physical or physically-informed vocal tract models. The choice of method is usually based upon the relative importance of two competing concerns: computational cost, and model accuracy and flexibility.

Several approaches, such as the transmission and reflection line approaches, are limited in that they can only reproduce 1D wave propagation, but have a low computational cost. More general time-domain acoustic modelling techniques, including FDTD, DWM and TLM, can be applied in any number of dimensions, so they can produce more detailed and accurate 2D or 3D vocal tract models at the expense of longer computation times. Finally, to accurately reproduce the geometry of a domain and incorporate frequency-dependent behaviour such as absorption by the vocal tract walls, methods such as FEM or BEM are required, but these have the highest computational cost of all.

4.2 One-Dimensional Vocal Tract Models

One-dimensional articulatory models treat the vocal tract as a series of—usually cylindrical—concatenated tubes, with one end corresponding to the glottis and the other to the lips. The cross-sectional area of the real vocal tract is determined by medical imaging of the vocal tract, with x-ray data typically used in older models (e.g. [25]) and magnetic resonance imaging (MRI) data used in more recent approaches (such as [129]).

4.2.1 Transmission-Line Vocal Tract Models

The earliest articulatory synthesisers represented the vocal tract as an electronic transmission line. One of the earliest such models is the system proposed by Dunn in 1950 [130], in which 25 transmission line segments are used to model the vocal tract. The cross-sectional areas (CSAs) of these segments are fixed at 6 cm^2, but a movable "tongue hump" is included which permits the identification of distinct vowels. Resistive losses due to viscosity and absorption by the vocal tract walls are deemed negligible, so each section is defined solely in terms of its inductance and capacitance. A periodic and white noise source are used at the glottis to produce voiced and unvoiced speech.

Dunn's model demonstrated the usefulness of the transmission line approach, and in 1953, Stevens, Kasowski and Fant [131] introduced a model of the vocal tract that offered variable CSAs along the tract and the option to vary the length of the tract by bypassing sections. Three source mechanisms accounted for phonation, turbulent noise, and transient plosive bursts, although the paper primarily focused on vowel production based on x-ray-derived tract shape information.

Hecker [132] was one of the first to add a nasal tract to the—by now quite complicated—transmission line models, linked to the oral tract by a velum section with variable CSA. This system also provided variable-length sections to account for lip rounding, and experimented with dynamically varying the model parameters—using linear interpolation between parameter sets—to produce dynamic output.

In 1975, Flanagan, Ishizaka and Shipley ([82] with the fundamentals of the model explained earlier in [18]) presented a vocal tract model that no longer relied upon physical electronic hardware but a computer simulation of its behaviour based on difference approximations. Most transmission line models introduced later are based to some degree on [82], which incorporated all the aspects of the above models in addition to a more realistic model of the yielding wall and losses due to sound radiation from the wall, relevant especially during the hold phase of voiced plosives. Furthermore, the noise source for turbulence was activated automatically when the critical Reynolds number was exceeded (see Section 2.6.1). Finally, the model was combined with a self-oscillating model of the vocal folds [133], providing a realistic glottal signal and the means to model natural source-tract interaction.

In 1976, Coker [134] introduced a control model for the Flanagan system [82] based purely upon articulator movement and glottal control parameters. Simulation run-times are provided, with approximately 6 minutes required to calculate 1 second of output speech. This model produced speech by rule from written text, making it an early TTS system.

The Maeda [135] model is still widely cited but is in fact simply a version of [82] that saves computational expense by injecting a glottal area signal

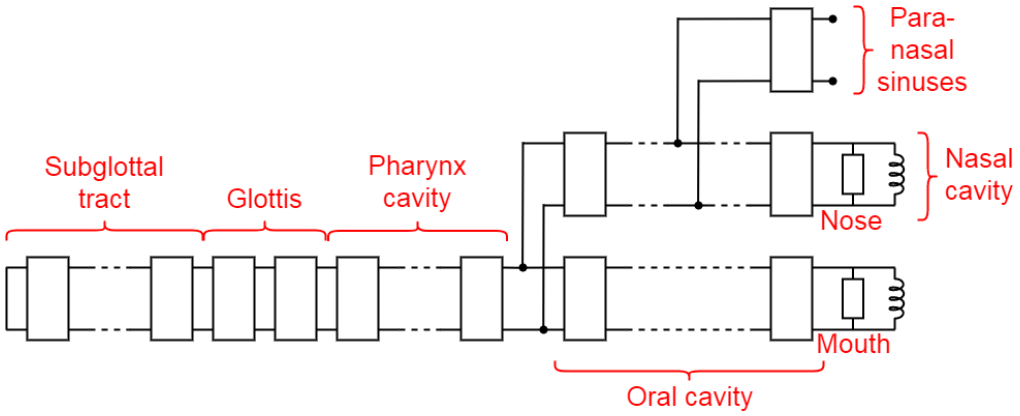


Figure 4.3: Combination of transmission line sections (each block represents a transmission line section similar to that of Figure 4.1) to form a complete vocal tract model incorporating subglottal tract, variable glottal opening, oral, nasal and paranasal cavities, following [137].

rather than using the true vocal cord model, and by ignoring turbulence, nasalisation and losses due to viscosity, heat conduction and wall radiation.

The model more recently introduced by Birkholz et al. [136, 137], which is used in VocalTractLab [138], is similar to [82] but adds paranasal sinuses which are modelled as Helmholtz resonators, as well as a section of subglottal tract. The system [137] is illustrated in Figure 4.3, with every block in the illustration representing a transmission line segment. Mechanical properties used in the simulations are given in [136] as follows: mass of vocal tract walls $M_{walls} = 21\ \mathrm{kg\,m^{-2}}$, resistance of vocal tract walls $R_{walls} = 8000\ \mathrm{kg\,m^{-2}\,s^{-1}}$, and stiffness of vocal tract walls $K_{walls} = 845000\ \mathrm{kg\,m^{-2}\,s^{-2}}$, all specified per unit area.

4.2.2 Reflection-Line Vocal Tract Models

Another early development in articulatory synthesis came about as a result of Kelly and Lochbaum's 1962 model [25], which uses a reflection line to model the vocal tract. A reflection line vocal tract model has several limitations compared to the transmission line, such as the necessity for fixed-length

tube sections and an inability to implement frequency-dependent losses, but the reduced computational expense made the Kelly-Lochbaum model very popular as a starting point for improved articulatory synthesisers.

The original Kelly-Lochbaum (KL) model was introduced in a very brief paper [25], but incorporated a number of features distinguishing it from a simple reflection line as introduced in Section 4.1.2. Damping was introduced—a factor of 1/128 applied to the backward-travelling wave component only—to produce more realistic formant bandwidths. The glottis was modelled as a 0.2 cm^2 area, and the lips were modelled with a radiation impedance of zero and some additional damping (not quantified). A nasal tract was also included, with more damping than the oral tract. Excitation was by periodic pulses at the glottis or noise at the point of constriction. Transitions between articulations (with shapes determined from x-ray data) were achieved using linear interpolation, and a number of rules were developed to model speech, including some coarticulatory effects.

In 1973 [22], a control method based on the articulators listed in Section 2.5.1 was produced, and this was further refined in [24] where it is noted that the transmission line model offers better automatic control of parameters such as turbulent noise amplitude, but at too high a computational cost for practical implementation.

In his doctoral thesis [41], Johan Liljencrants addressed a number of issues with the reflection line based vocal tract model, and introduced corrections to the scattering equations to more accurately account for dynamic tract shape changes. He also developed expressions for frequency-dependent loss factors, permitting more accurate modelling. A series loss of $3.626 \times 10^{-5} \times \sqrt{f/A}$ accounts for viscous losses, and a parallel loss of $1.610 \times 10^{-5} \times \sqrt{f/A}$ accounts for thermal losses, illustrating the proportionality of both terms to the square root of frequency f, as well as their dependence on cross-sectional area, A. Values for the speed of sound within, and density of, the vocal tract walls were provided as $c_{walls} = 1500\ \mathrm{m\,s^{-1}}$ and $\rho_{walls} = 1000\ \mathrm{kg\,m^{-3}}$ respectively, based on "body temperature water of 1% salinity". Noise was generated when the Reynolds number Re exceeded the critical value of 1800, and had

an amplitude proportional to $Re - 1800$.

Perry Cook's thesis [139] explored the application of the reflection line model to the singing voice, concentrating mainly on the natural fluctuations in the source signal that appear to be necessary for perceptual naturalness in the output. This culminated in the production of a real-time singing synthesis system.

Välimäki and Karjalainen [140] improved upon the underlying assumption of a piecewise cylindrical duct by using conical tube sections, which resulted in the reflection coefficients at tube boundaries being replaced with digital filters. Fractional delay filters were also used to allow the length of each tube section to vary, overcoming an often-cited weakness of the reflection line model.

Finally, Brad Story's doctoral thesis [141] sought to introduce a personalised element to synthesis, using MRI scans of specific subjects rather than the generic x-ray data of an anonymous subject that had been used previously. Following Liljencrants [41], the model incorporated yielding walls and frequency-dependent losses, while noting that "it is the *effect* of the viscous loss that is being modelled rather than the fluid dynamical loss mechanism itself". This study illustrated that even small changes in the CSA at a constriction can have a significant effect on formant frequencies, highlighting the necessity of accurate vocal tract imaging.

Several other approaches to 1D vocal tract modelling exist that fall outside the simple reflection/transmission line dichotomy, including hybrid time and frequency domain models such as [120]. However, these models do not introduce any new insights aside from those previously discussed in the models above.

4.2.3 Shortcomings of 1D Vocal Tract Models

Models that rely on a 1D approximation of the vocal tract for synthesis are inherently limited, and despite the improvements offered by modelling losses and branching in the vocal tract, a number of factors mean that there is

always an upper limit on accuracy for this type of simulation:

- The plane wave assumption on which 1D models are built is only valid up to around 5 kHz (see Section 2.3), and ignores cross-modes in the tract which may occur at frequencies as low as 3.92 kHz [30].
- Models are typically unable to temporarily split the airway in order to produce lateral articulations such as /l/.
- The bend in the vocal tract is not typically accounted for, which introduces errors above 5 kHz [142].
- The radiation models used to simulate acoustic behaviour at the lip end of the vocal tract are approximate and do not usually take into account the shape of the lips, known to affect accuracy even below 4 kHz [143].
- Models typically assume that the cross-section is the same shape along the length of the tract and that neighbouring tubes are joined along their centre line. Both of these assumptions have been shown to affect the accuracy of the transfer function, particularly above 5 kHz [142, 144].

In order to rectify these shortcomings, it is necessary to consider vocal tract models in multiple dimensions.

4.3 Two-Dimensional Vocal Tract Models

The 1D vocal tract models described above provide a good first approximation to the physics of the vocal tract. However, by increasing the dimensionality of the vocal tract model to 2D, it is possible to address at least the first two issues discussed above. While 2D vocal tract models have received limited attention in the literature—focus tends toward either the lower computational cost of 1D models or the improved accuracy of 3D models—several studies have demonstrated the flexibility of 2D vocal tract models.

One early 2D vocal tract model was that of Thomas [145], which sought an FEM approximation to the Navier-Stokes equation. The Navier-Stokes equation is more complicated than the wave equation and incorporates fluid dynamic properties such as viscosity, and as such [145] represented an ambitious and computationally demanding simulation that has not yet been attempted in 3D. However, the approach was successfully used to simulate a diphthong. Approximately 100 hours of computation time was required to generate one second of output speech.

At the other end of the computational complexity scale, Mullen introduced a static 2D DWM model [146], and later a dynamic 2D DWM model [147], capable of producing diphthongs in real time. The dynamic method makes use of a heterogeneous implementation of the DWM that will be described in detail in the next chapter. This method was extended upon in [33] and [148] to simulate plosive and nasal consonants with some success.

One issue affecting 2D vocal tract models is that they are not as accurate as 3D models, and the modelling approach in [149] sought to address this by 'tuning' a 2D FEM vocal tract model to better match the acoustic output of a 3D FEM model. The tuning process included scaling the CSAs of 2D models (effectively narrowing the tract), altering the boundary admittances along the length of the model in proportion to the ratio between 2D and 3D CSAs, altering the parameters of the glottal source, and iteratively optimising the 2D radiation model to match that of 3D. The tuned 2D models showed formant values closer to those of 3D models than non-tuned 2D models did, but comparisons against recorded speech were not made. Acoustic parameter values given for this simulation were $c_{air} = 350\ \mathrm{m\,s^{-1}}$ and $\rho_{air} = 1.14\ \mathrm{kg\,m^{-3}}$.

A recent 2D modelling approach [150] is based on an FDTD simulation, but makes used of a method known as the immersed boundary method (IBM) to specify a vocal tract boundary at any location, not just Cartesian grid points. This presumably has the effect of increasing the accuracy of the simulation, although detailed results are not provided in this proof-of-concept study. The authors state a desire to apply the approach in 3D, but this has not yet been attempted.

Despite a number of advantages over 1D models, in particular the ability to model cross-modes in vocal tract and simulate lateral articulations, 2D models do not usually address concerns to do with accurate radiation modelling, irregularity of cross-sectional shapes, and in most cases, a bent vocal tract. It is necessary to increase the dimensionality still further in order to address these issues.

4.4 Three-Dimensional Vocal Tract Models

With the advent of more powerful computers, 3D models of the vocal tract became feasible in the early 1990s. One of the earliest models [151] took full advantage of the 3D geometry, obtained from x-ray images and assumed to be symmetrical, to form a 3D BEM vocal tract model and a comparable FEM model. A complex, frequency-dependent wall impedance of $14000 + j16\omega$ $\mathrm{kg\,m^{-2}\,s^{-1}}$ and a model of diffraction due to the head and shoulders was included. It is not clear why this line of study was not further pursued, but the large computational cost—alluded to, but not specified in [151]—is likely to have been a factor. A later paper [152] from the same team investigates the deformation of the BEM models to match a target vocal tract shape. Aside from this, the majority of early 3D studies retained simplified cross-sections and straight tubes, but began to explore 3D features such as changing the cross-section shape [153] or varying the impedance on different surfaces [154], more accurately representing real vocal tract properties.

An early time-domain approach was the TLM method explored by El-Masri et al. [155, 156] which used a rectangular approximation of the vocal tract duct, and later attempted to model obstacles such as teeth [157], although very little detail is given about the simulations save that the vocal tract walls were considered acoustically hard. A simulation time of around 29 hours to calculate 100 points of a transfer function is quoted in [156]. Throughout the 1990s and early 2000s, 3D vocal tract models continued to develop, often using FEM models with some simplification of the true 3D shape and detailed radiation characteristics, such as [158, 159, 142, 160]. These studies largely

demonstrated the predicted increase in accuracy of 3D models over 1D models, particularly for frequencies above 5 kHz, but retained highly simplified radiation models and cross-sectional shapes, and/or modelled the vocal tract walls as acoustically hard.

Towards the end of the 2000s, detailed 3D vocal tract models with more realistic, yielding wall behaviour began to be developed. These models were used successfully in an FEM approach to simulate Czech vowels [161, 162] and simulation details were provided: $\rho_{air} = 1.2\ \mathrm{kg\,m^{-3}}$, $c_{air} = 353\ \mathrm{m\,s^{-1}}$, and $Z_{walls} = 83666\ \mathrm{Pa\,s\,m^{-1}}$. A radiation volume was included, although the shapes of the lips and head were not yet incorporated. In [161], parts of the tract were manipulated—the tonsils were removed—to determine the acoustic effect of changing this part of the vocal tract shape prior to surgery. Removing the tonsils was found to affect formant frequencies by as much as 450 Hz depending on the vowel studied.

The effect of different parts of the vocal tract on the VTTF is explored further in Takemoto et al.'s work [30], which is an FDTD model of the tract based on MRI data with the teeth accounted for [29]. In this model, simulated VTTFs are compared to those measured from a 3D-printed model of the same vocal tract geometry, so yielding walls are not included in the model. As the vocal tract is not *in situ*, simulation parameters $c_{air} = 346.7\ \mathrm{m\,s^{-1}}$ and $\rho_{air} = 1.17\ \mathrm{kg\,m^{-3}}$ are used, corresponding to the temperature in the measurement room. The walls are given a normal absorption coefficient of 0.004, corresponding to a wall impedance of $Z_{walls} = 405096\ \mathrm{Pa\,s\,m^{-1}}$, which appears to have been obtained empirically, and may be too low to model the hard 3D-printed walls based on the values obtained in other studies. Nevertheless, the simulations show good agreement with measured VTTFs, and various side branches are then systematically occluded to determine their effects on the VTTF. Results suggest that the piriform fossae have a significant effect, introducing antiresonances around 4 kHz, altering formant frequencies and amplitudes, and exhibiting interaction between the left and right fossae. Similar but much smaller effects were noted for the epiglottic valleculae, and some vowel-dependent effects were also noted for the inter-

dental spaces. Transverse resonance modes were observed at 3.92 kHz and above. This paper re-established FDTD as a suitable approach for vocal tract modelling, and simulations times of 60 minutes per 50 ms, i.e. 20 hours per second of output, were reported.

Wang et al. produced several further FDTD vocal tract studies [163, 164], using c_{air} and ρ_{air} values from [30] (although these papers appear to model the vocal tract *in situ*, so these values may not be appropriate) and $c_{walls} = 1500\ \mathrm{m\,s^{-1}}$ and $\rho_{walls} = 1000\ \mathrm{kg\,m^{-3}}$. This gives $Z_{walls} = 1.5 \times 10^6\ \mathrm{Pa\,s\,m^{-1}}$, where the specific acoustic impedance of the vocal tract walls, Z_{walls}, is calculated from ρ_{walls} and c_{walls} following (2.3). In [164], simulations are compared to recordings and a mean absolute error of around 6% is reported for the first four formants, for a single vowel, although details of the comparison recordings are not provided. Simulation times are given as 100 hours [163] and 13.3 hours [164] for one second of output. Dynamic behaviour is simulated in [164] by calculating the VTTFs of intermediate vocal tract shapes in a sequence, and interpolating between them, however the simulation model itself is not truly dynamic.

A 3D DWM method for vocal tract modelling was introduced in [165] for cylindrical vocal tract analogues, and extended to the detailed MRI geometry in [166, 167]. This study was particularly notable for using multiple sets of MRI data for different participants to inform the models, finding some differences between participants which appeared to be related to the quality of the MRI data. A reflection coefficient of 0.99 was used at the vocal tract walls. The simulations were found to outperform 1D and 2D DWM vocal tract models in terms of the accuracy of formant frequencies compared to recordings of natural speech.

The most extensive 3D model developed in recent years has been the outcome of a project called EUNISON [168], where a time-domain FEM technique has been developed [45] permitting the synthesis of diphthongs. This model has been used to determine the effects of simplifications in the vocal tract, including head geometry [21], effect of the lips [143], and the tract geometry itself [144], concluding that accuracy in lip and tract geometry is essential for high-

frequency accuracy in the VTTF, and that the head may be approximated as a sphere with equivalent volume without having too detrimental an effect on the output. Simulations use $c_{air} = 350\ \mathrm{m\,s^{-1}}$, $\rho_{air} = 1.14\ \mathrm{kg\,m^{-3}}$, and $Z_{walls} = 83666\ \mathrm{Pa\,s\,m^{-1}}$. Simulation times vary depending on the complexity of the model studied, but for detailed MRI-based geometry with a radiation volume, simulation times are around 60 [149] to 80 [143] hours for 20 ms events, which means several thousand hours would be required to calculate a second of output speech. To the author's knowledge, the EUNISON model has not been compared to recordings of natural speech to determine its accuracy. Like all the models described in this section, the EUNISON model also lacks a nasal and subglottal tract, and while FEM permits frequency-dependent loss modelling, this has not yet been incorporated. Consonant modelling has occasionally been considered [169], but a reliable model of their production has yet to be developed.

A number of other 3D vocal tract models have been considered over recent years, most of which are similar to the models described above. Fricative production has briefly been considered in the context of a TLM model ([170]), although results of these simulations are not forthcoming. Development of vocal tract airway models has proceeded largely in parallel with FEM models of the vocal tract articulators (e.g. [171, 172]), but it is only very recently that airway and articulator models have been combined [173] in a proof-of-concept study. Simulation times are not given but are assumed to be very high given the additional computations required to specify articulator behaviour. It is anticipated that future vocal tract airway models may be combined with parametrised articulator models in a similar way to provide a gestural control interface for simulation of speech. However, at present the 3D airway models are insufficiently developed to allow this.

4.5 Evaluating Physical Vocal Tract Models

As discussed in Section 3.5, the challenges inherent in evaluating TTS methods make it difficult to produce an overall metric for synthetic speech quality.

In the case of vocal tract models, which are typically not incorporated into TTS systems and—as in the case of 3D models—may not yet even be capable of producing all types of phoneme, these problems are compounded: as word-, phrase- and sentence-level outputs are unlikely to be available, there is already an element of significant unnaturalness to speech segments presented in isolation.

Perceptual tests of naturalness and intelligibility remain relevant in assessing the quality of vocal tract model output, as the ultimate goal for the models must be the generation of intelligible, natural-sounding speech. However, where 3D vocal tract models produce synthetic speech based on a specific person's vocal tract geometry, it becomes possible to assess the accuracy of the simulation objectively by comparing the synthesised output to the recorded natural speech. As noted earlier, there is some danger in basing a quality criterion for synthetic speech on similarity to a specific audio signal, as it limits the drawing of general conclusions about synthesis quality. However, objective similarity to recorded audio does provide a means to verify 3D vocal tract shapes and the accuracy of the physical modelling technique. At present, with 3D physical models still not fully developed, such validation remains a necessary step in assessing the quality and accuracy of the output speech.

Means for comparing synthetic and natural speech are generally focused on the frequency domain, and are generally based on how well the formant frequencies are reproduced, as well as spectral details such as the antiresonances caused by the piriform fossae. Naturalness also requires realistic behaviour in the high frequency regions above 5 kHz, and adequate damping of formants, but this has not yet been well-studied because, as noted above, frequency-dependent losses are generally lacking in state-of-the-art physical vocal tract models. Likewise, the effect of the nasal cavity or subglottal system on modelled spectra are not yet known.

A priority for the development of 3D vocal tract models is the reproduction of consonants. Once this is possible, words and phrases may be generated, and subjected to the same measures of speech quality used to assess TTS

systems. Until such a time, the quality of the output speech must be assessed by a combination of objective measures of similarity to a target signal, and perceptual measures of the acceptability of isolated word fragments such as diphthongs, keeping in mind the potential weaknesses of these methods.

4.6 Challenges for 3D Vocal Tract Modelling

In order to produce the most natural-sounding output, it is necessary to use 3D vocal tract models. However, as described in the previous sections, there are a number of issues in 3D vocal tract models that are not yet solved. For instance, the models generally lack a nasal cavity, and means of producing fricative or plosive consonants, rendering them incapable of generating synthetic speech at present. Additionally, the models lack realistic frequency-dependent loss mechanisms, affecting the spectrum of the output especially in the higher frequencies. These issues are expected to be solved in the next 5–10 years, inspired by the processes used for 1D models described in Section 4.2.

A pressing issue for 3D models is computational expense. As discussed above, the most accurate models require very long run times rendering them unusable for synthetic speech applications in the foreseeable future. However, the human ear and brain will ultimately decide whether the produced signal is acceptably natural, and it seems likely that a perceptual threshold exists beyond which increases in accuracy have minimal audible effect. A perfectly accurate physical model may be useful for the study of speech production and acoustics. However, focus on *perceptual* acceptability, rather than perfect reproduction of the system physics, is required if vocal tract models are to be made usable for synthetic speech applications.

Finally, control of articulatory synthesisers, which was touched upon in Section 3.2.4, becomes a real concern for 3D vocal tract models which have many thousands of discrete parameters. It will be necessary to develop relationships and translations between gestural control parameters, such as the movement of an articulator, into appropriate changes in the individual

parameters for each modelling element.

4.7 Conclusion

This chapter has introduced the wide variety of approaches that are available in the production of a physical model of the human vocal tract. While one-dimensional systems such as the reflection line and transmission line analogues of the vocal tract offer low computational expense and hence minimal computation times, inherent simplifications in one-dimensional models make them of limited use in the production of natural-sounding synthetic speech. Two-dimensional modelling approaches offer some improvements over one-dimensional approaches, but again require simplification to the geometry and wave propagation modelling that reduce the naturalness of the output signal.

Three-dimensional vocal tract models are required if truly natural-sounding synthetic speech output is to be produced. The remainder of this thesis will make use of the digital waveguide mesh (DWM) in order to construct a 3D vocal tract model. The DWM has several advantages over the other methods introduced in this chapter: the use of a regular grid makes wave propagation simpler to interpret and less computationally expensive than FEM and BEM which use an irregular grid; and the existence of admittance as a free parameter in the modelling equation (4.7) allows dynamic behaviour to be implemented more intuitively than in the FDTD method, as described in the next chapter.

Three-dimensional models of the vocal tract require significant computational expense, making them unsuitable for use in most practical speech synthesis applications at present. However, such models offer the greatest potential for achieving the goal of natural-sounding synthetic speech in the future.

Part III

Original Research

Chapter 5

Dynamic 3D DWM Vocal Tract

The previous chapter illustrated the significant potential of physical models for synthesising natural-sounding speech. A number of different physical modelling approaches were introduced in their general form, including the digital waveguide mesh (DWM). In this chapter, the DWM is used to produce a novel physical model capable of reproducing the detailed 3D vocal tract geometry obtained from MRI scan data, and dynamically changing the shape of the vocal tract during synthesis. The DWM method was chosen for this purpose as it uses wave variables and a regular discretisation of the simulation domain that are intuitively easy to understand and interpret. As illustrated in Section 4.1.4, the DWM method also features admittance as a free parameter, facilitating the implementation of heterogeneous, and hence dynamic, models as this chapter will illustrate.

The dynamic nature of the modelling approach means that dynamic phonemes, as described in Chapter 2, can be implemented. Since the model does not currently incorporate the simulation of turbulence, or a nasal tract, diphthongs are used to illustrate the ability of the model to perform dynamic articulations.

This chapter will first introduce the existing vocal tract models upon which the proposed model builds, and will then describe the implementation and refinement of the proposed model before performing objective evaluation of

the resulting simulations.

5.1 Homogeneous DWM Vocal Tract Models

In the homogeneous DWM, admittance Y is constant throughout the mesh, so (4.7) simplifies to:

$$p_J(n) = \frac{2}{N} \sum_{i=1}^{N} p^+_{J,J_{nei}}(n) \tag{5.1}$$

with the remaining steps of the DWM algorithm following (4.9) and (4.10). A homogeneous digital waveguide in one dimension is not suitable for vocal tract modelling, but in higher dimensions the shape of the domain can be used to represent the shape of the vocal tract.

5.1.1 Homogeneous 2D DWM Vocal Tract Model

A 2D vocal tract model was introduced in [146], using a 2D DWM with a width governed by the area function of a specific vowel, as illustrated in Figure 5.1. This model was shown in [146] to more accurately reproduce formant frequencies than a 1D Kelly-Lochbaum type model (see Section 4.1.2). Furthermore, varying the reflection coefficient at the sides of the mesh was shown to have a direct effect upon formant bandwidth, allowing the model to approximate realistic formant bandwidth values. This model was therefore successful in the simulation of static vowel sounds. However, as the shape of the domain was governed by the area function, any changes between vowels over the course of a simulation required the addition or removal of scattering junctions, affecting continuity at the junctions and resulting in discontinuities in the output signal as illustrated in Figure 5.2. For this reason, the homogeneous 2D DWM vocal tract model is unsuitable for the synthesis of dynamic speech sounds and running speech.

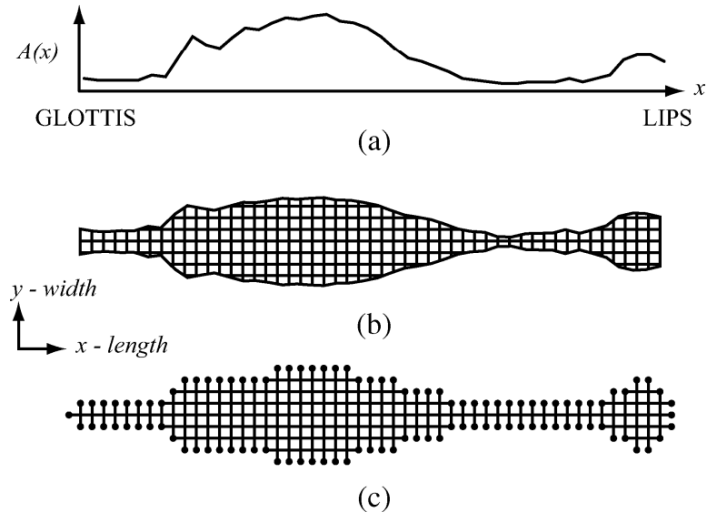


Figure 5.1: Widthwise 2D DWM vocal tract model, based on the area function for vowel /i/, from [147].

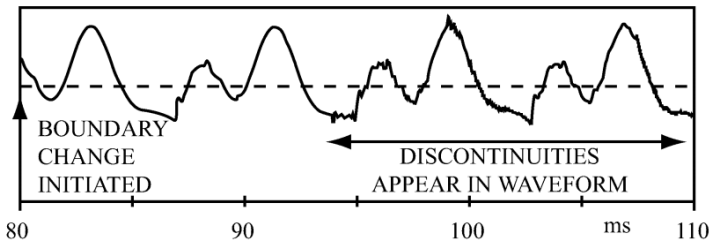


Figure 5.2: Discontinuities that appear in simulated signal after changing 2D DWM model shape, from [147].

5.1.2 Homogeneous 3D DWM Vocal Tract Model

A further development in homogeneous modelling was the 3D homogeneous DWM vocal tract model proposed by Speed [166, 167]. In this model, a complete 3D vocal tract geometry was obtained for a number of subjects using MRI scans, and converted into a Cartesian grid of scattering junctions located within the vocal tract airway. A small radiation volume, which incorporated the shape of the lips, was included, with an approximately anechoic boundary implemented at the edge of this volume to model sound propagation into the free field.

The simulation results from the 3D homogeneous DWM model were shown to more accurately reproduce formant frequencies and spectral characteristics of recorded speech than 1D and 2D DWM simulations, illustrating the importance of correctly reproducing the detailed 3D vocal tract geometry, for static vowel sounds. Furthermore, the study found differences in simulation accuracy between subjects, which has important implications for the study of 3D vocal tract models, which in the literature tend to only consider one subject. However, as the vocal tract shape was mapped to the shape of the domain, the 3D homogeneous DWM model is, like the 2D homogeneous model, unsuitable for the simulation of dynamic speech sounds.

5.1.3 A note about MRI data

The homogeneous 3D DWM vocal tract model, and the other models presented in this section, make use of MRI data to inform the vocal tract shape. It is therefore important to note some details and issues associated with MRI scans of the vocal tract.

One issue with MRI scans, noted previously in Section 2.5.2, is that teeth and bone appear the same as air in the resulting images, making it very difficult to determine the edges of the vocal tract airway in the area around the teeth. A method to incorporate the teeth into MRI vocal tract data was presented in [29], but requires dental casts for the subject which are not always available, as is the case for the MRI subject used in this study.

A further problem with MRI data is that a supine position is required during the scan, which may lead to differences in articulation compared to normal speech, due to the effect of gravity. Additionally, the noises made by the scanner disturb auditory feedback, so participants cannot hear their own voices. Finally, the low speed of the 3D scanning procedure means that articulations must be held for an unnaturally long time. These effects may all contribute to *hyperarticulation* [174], such that MRI data is not truly representative of natural speech.

Despite these shortcomings, MRI scans are currently the best method available for obtaining 3D vocal tract data, so will be used for all the models described throughout this chapter. The issues above should be kept in mind throughout as a possible source of error in the simulations.

5.2 Heterogeneous DWM Vocal Tract Models

As described in previous sections, mapping vocal tract shape to the domain shape in a DWM model is sufficient to produce static speech sounds. However, to produce dynamic articulations, a heterogeneous implementation is required.

The free parameter admittance in (4.7) permits the implementation of a heterogeneous DWM model. By altering the value of Y throughout the mesh, realistic heterogeneous scattering behaviour is introduced. It is also possible to produce heterogeneous FDTD models using a similar method, but in this case the parameter varied is wave speed, c. When c is varied, its maximum value c_{max} must be used in the calculation of the maximum permissible time step Δt following (4.6), which may result in significantly increased computational complexity for high values of c_{max}. Wave speed is not implemented directly in the DWM, despite being included in the definition of admittance (2.3), so the value of Δx may be calculated based on wave speed in the airway. This same mesh spacing can then be used throughout the vocal tract walls,

with a higher impedance (lower admittance) applied to the waveguides, without having to change Δx due to the change in wave speed. For this reason, the heterogeneous DWM approach does not accurately model wave propagation in the tissue surrounding the vocal tract, making it a *physically-informed*, rather than truly *physical* model. However, as this chapter will illustrate, the approach is sufficient to generate realistic vowels.

The benefit of a heterogeneous DWM model is the way that it is only necessary to alter the admittances within the mesh in order to effectively change the vocal tract shape. Therefore, the domain is set to a suitable size and shape to accommodate all possible admittance maps required for the application, and the outer boundaries of this domain do not move, preventing any errors associated with moving the domain boundaries. As demonstrated in [147], by interpolating between admittance maps over the duration of a simulation, it is possible to simulate dynamic speech sounds such as diphthongs and plosives.

5.2.1 Heterogeneous 2D DWM Vocal Tract Model

One method for producing a set of mesh admittances, known henceforth as an *admittance map*, for a 2D DWM was introduced in [147]. Like the homogeneous 2D DWM method described in Section 5.1.1, the vocal tract shape information is presented as a 1D area function. However, instead of varying the shape of the mesh to match the area function, the mesh is fixed at a size and shape sufficient to support all possible area functions: a rectangle with equivalent physical length 17.6 cm [147]. The area function value A_x, at any distance x from the glottis, is then used to define a raised cosine function across the mesh, with impedance Z_x calculated from A_x following (2.9). The definition of the raised cosine function is illustrated in Figure 5.3. A raised cosine map is produced for every x in the mesh to produce the complete impedance map illustrated in Figure 5.4 (the admittance map simply consists of reciprocal values). This map shows the central channel of low impedance along the centreline of the mesh, and how the impedance at the edges governs the effective shape of the mesh without having to alter the

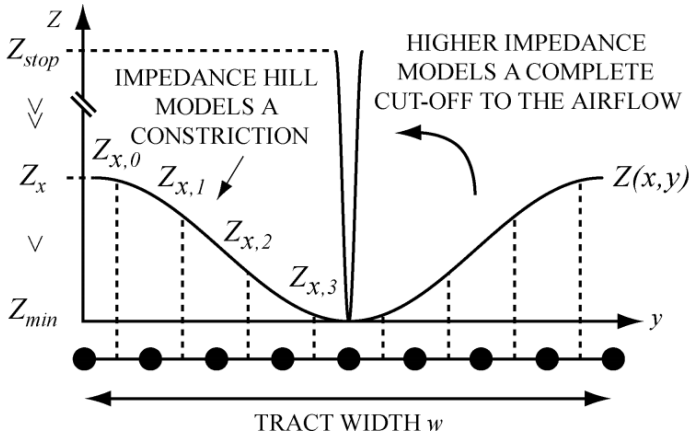


Figure 5.3: Cosine mapping procedure used in dynamic 2D DWM vocal tract model, from [147].

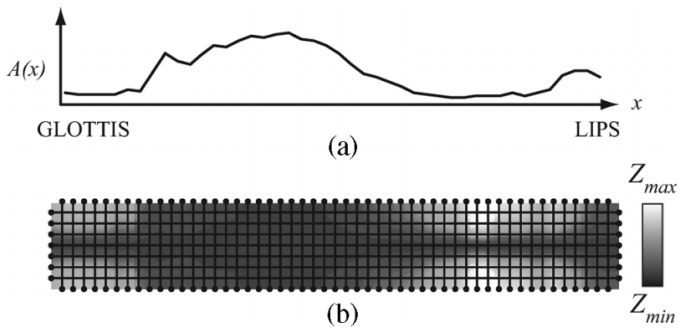


Figure 5.4: Cosine-mapped 2D DWM vocal tract model, based on the area function for vowel /i/, from [147].

domain boundaries.

The heterogeneous 2D DWM vocal tract model is clearly not a true physical model, but it does produce acceptable vowel sounds. Furthermore, as the domain does not change size, it is possible to alter the effective shape of the model by simply interpolating between admittance maps over the duration of a simulation. This process was shown in [147] to avoid the discontinuities produced when a domain changes shape, and to produce realistic spectra throughout the transition. Additionally, the model is capable of operating in real time.

The dynamic 2D DWM vocal tract model offers improved simulation accuracy over 1D models, as it permits some simulation of cross-tract modes, and has the potential to model lateral consonants, although only vowels are implemented in [147]. However, as discussed in Section 4.3, 2D models still have a number of accuracy issues that cannot be addressed except by considering the third dimension. Therefore, this thesis proposes a 3D heterogeneous DWM vocal tract model, which will be described in the following sections.

5.3 Boundaries in the DWM

For a complete DWM vocal tract model, it is also necessary to define the behaviour at the edges of the DWM domain, known as the domain *boundaries*. A number of boundary formulations exist for the DWM, and this work makes use of one of the most common, known as the locally reacting wall (LRW) method [175]. The LRW is implemented at the domain boundary, denoted Γ_D. A normalized admittance parameter, G, specifies the reflection properties of the boundary.

The LRW formulation provides an expression for a 'ghost point', lying outside the domain and neighbouring a junction lying on the boundary. By substituting this expression for the ghost point into the boundary junction's update equation, the behaviour of the boundary is modelled. For a 3D rectilinear grid, there will be three ghost points outside the corner of the domain,

two ghost points outside a domain edge, and one ghost point outside the faces of the domain. By substituting the ghost point expressions into the junction update expressions, relationships are defined for the boundary junctions.

If a boundary junction is indexed by (k, l, m), then at a face the ghost point will exist at $(k+1, l, m)$ and the pressure at the boundary junction is given by:

$$p^{n+1}_{k,l,m} = \frac{2\sqrt{3}}{3\left(\sqrt{3}+G_k\right)} p^{n}_{k-1,l,m} + \frac{\sqrt{3}}{3\left(\sqrt{3}+G_k\right)} \left(p^{n}_{k,l+1,m} + p^{n}_{k,l-1,m} + p^{n}_{k,l,m+1} + p^{n}_{k,l,m-1}\right) + \frac{\left(G_k-\sqrt{3}\right)}{\left(G_k+\sqrt{3}\right)} p^{n-1}_{k,l,m} \quad (5.2)$$

where G_k is the normalised admittance of the boundary between k and $k+1$, and n is the current time step. The equation makes use of FDTD notation, but since the FDTD and the DWM are equivalent, the boundary junction update equation (5.2) can be used in place of (4.7) where the current scattering junction is located on a boundary.

Using the same approach, with two ghost points at $k+1$ and $l+1$ for a domain edge:

$$p^{n+1}_{k,l,m} = \frac{2\sqrt{3}}{3\left(\sqrt{3}+G_k+G_l\right)} \left(p^{n}_{k-1,l,m} + p^{n}_{k,l-1,m}\right) + \frac{\sqrt{3}}{3\left(\sqrt{3}+G_k+G_l\right)} \left(p^{n}_{k,l,m+1} + p^{n}_{k,l,m-1}\right) + \frac{\left(G_k+G_l-\sqrt{3}\right)}{\left(G_k+G_l+\sqrt{3}\right)} p^{n-1}_{k,l,m} \quad (5.3)$$

Finally, at a corner, there will be three ghost points at $k+1$, $l+1$ and $m+1$:

$$p^{n+1}_{k,l,m} = \frac{2\sqrt{3}}{3\left(\sqrt{3}+G_k+G_l+G_m\right)}\left(p^n_{k-1,l,m}+p^n_{k,l-1,m}+p^n_{k,l,m-1}\right) + \frac{\left(G_k+G_l+G_m-\sqrt{3}\right)}{\left(G_k+G_l+G_m+\sqrt{3}\right)}p^{n-1}_{k,l,m} \quad (5.4)$$

5.4 Proposed Model

As noted in Section 5.2.1, a heterogeneous DWM vocal tract model is required in order to synthesise dynamic speech. The 2D DWM discussed in Section 5.2.1 approximates a 3D vocal tract using a raised-cosine function based on characteristic acoustic impedance, as a compromise to address the issue of mapping a three-dimensional shape in two dimensions. In three dimensions, however, the accurate vocal tract shape can be reproduced directly using specific acoustic impedance, eliminating the need for the raised-cosine approximation. The proposed model is therefore a heterogeneous, 3D DWM model of the vocal tract, which uses variable specific acoustic admittance throughout a cuboid-shaped domain to represent the vocal tract, surrounding tissues, and a radiation volume of air outside the mouth; the boundaries implemented using LRW as discussed in Section 5.3.

5.4.1 Data Acquisition and Pre-Processing

In order to create a DWM model of the vocal tract, it is necessary to obtain vocal tract shape information. The most accurate data for this purpose comes from MRI data of the upper airway. In this study, the MRI corpus collected in [166] is used. This corpus consists of 11 vowels and several other phonemes for five trained speakers, each of which was held for 16 s while a 3D scan procedure was completed. The images are 512 × 512 × 80 anisotropic grayscale images, resampled from 2 mm isotropic images. In order to focus machine resolution on the vocal tract, the images only extend to approximately 4 cm either side of the midsagittal plane, and hence do not

capture the subject's entire head.

In addition to the MRI data, anechoic audio recordings of the same utterances were collected immediately before and after the MRI scans. These recordings were made in MRI-like supine conditions, with MRI machine noise played back to the subject over headphones to disturb auditory feedback, recreating the vocalisation conditions within the MRI scanner. The full details of the collection process are presented in [166]. This study makes use of the vocal tract information for a single adult male subject, with the nasal tract omitted due to poor resolution in that part of the image. This is a common problem with MRI scan data, and computed tomography (CT) data is preferred for nasal tract imaging [26]; however, as CT scanning involves radiation it is very difficult to obtain approval for the collection of non-medical data. A vocal tract model without a nasal tract is sufficient for the synthesis of non-nasalised vowels, but it is acknowledged that the model must be extended to include a nasal tract if it is to be capable of synthesizing the full range of phonemes in the future.

As the proposed model will be used to synthesise the eight English diphthongs—/eɪ/ as in *day*, /aɪ/ as in *high*, /ɔɪ/ as in *boy*, /eə/ as in *fair*, /əʊ/ as in *show*, /ɪə/ as in *near*, /ʊə/ as in *jury*, and /aʊ/ as in *now*—it is necessary to use MRI data for the phonemes /e/, /a/, /ɪ/, /ɔ/, /ə/, and /ʊ/. Once collected, the MRI data must be pre-processed in order to generate a rectilinear grid of points representing the vocal tract volume. A complete overview of this process is illustrated in Figure 5.5.

The MRI scan data (Figure 5.5a) must first be analysed to obtain the vocal tract shape. The software ITK-Snap [176] is used for this purpose, which performs user-guided active contour segmentation based on image contrast, and is specifically designed for anatomical structures. However, as there is no difference in contrast between air and hard structures such as bone and teeth in MRI scan data, the segmentation volume initially includes the teeth as part of the airway[1]. MRI segmentation algorithms are also prone to leakage

[1]Methods exist to superimpose the geometry of the teeth onto the vocal tract volume [29], but for the subject under study, no dental cast was available.

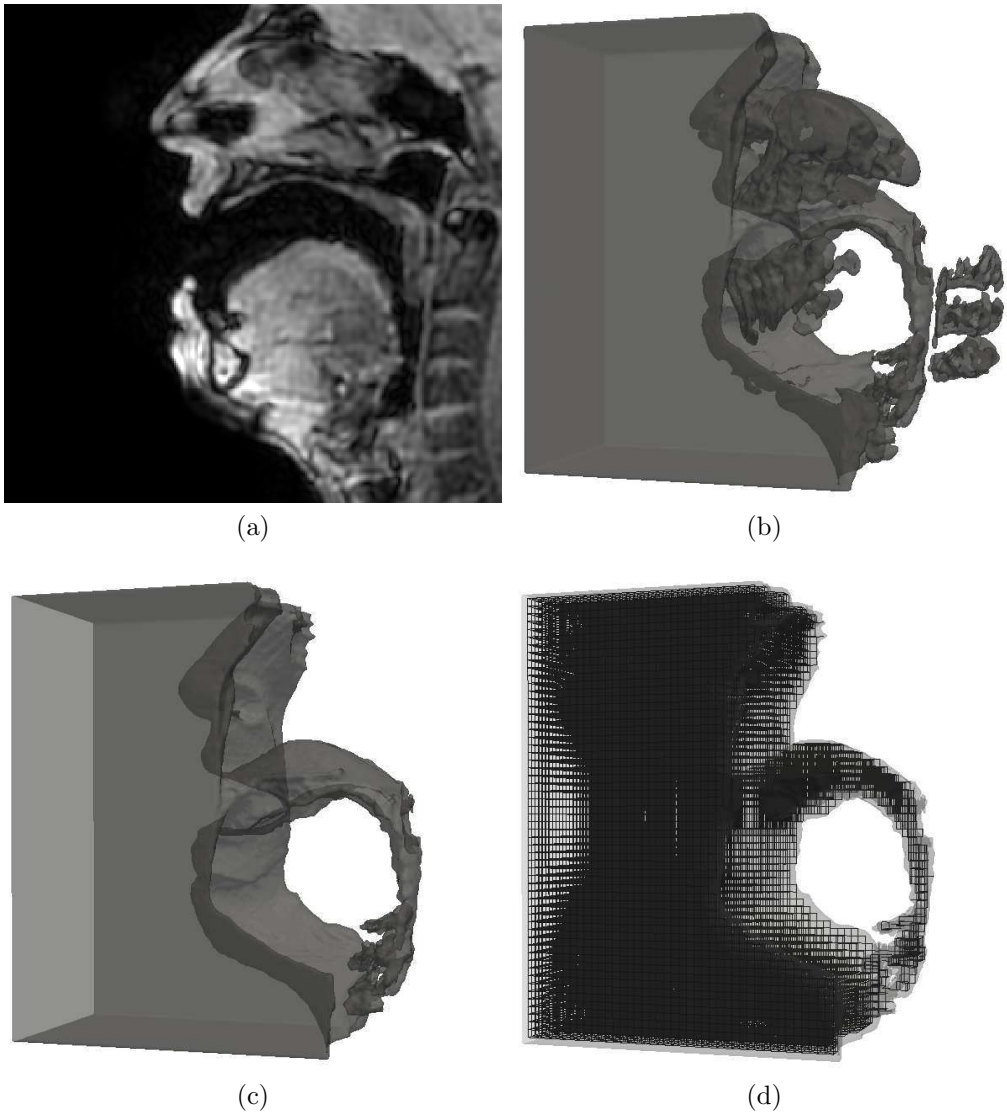


Figure 5.5: Standardization procedure for vowel /a/, from (a) MRI data (midsagittal slice shown), through (b) segmentation procedure (illustrating leakage of the segmentation volume into surrounding tissues), (c) hand-corrected segmentation data, and (d) associated rectilinear grid, calculated using a sampling frequency of 400 kHz.

into surrounding tissue areas, so the initial results of the segmentation often look similar to Figure 5.5b, with the teeth, several vertebrae, and parts of the jaw bone and nasal cavity included as part of the vocal tract airway. The resulting volume must be inspected, and any erroneous sections removed by hand, giving the final edited volume (Figure 5.5c).

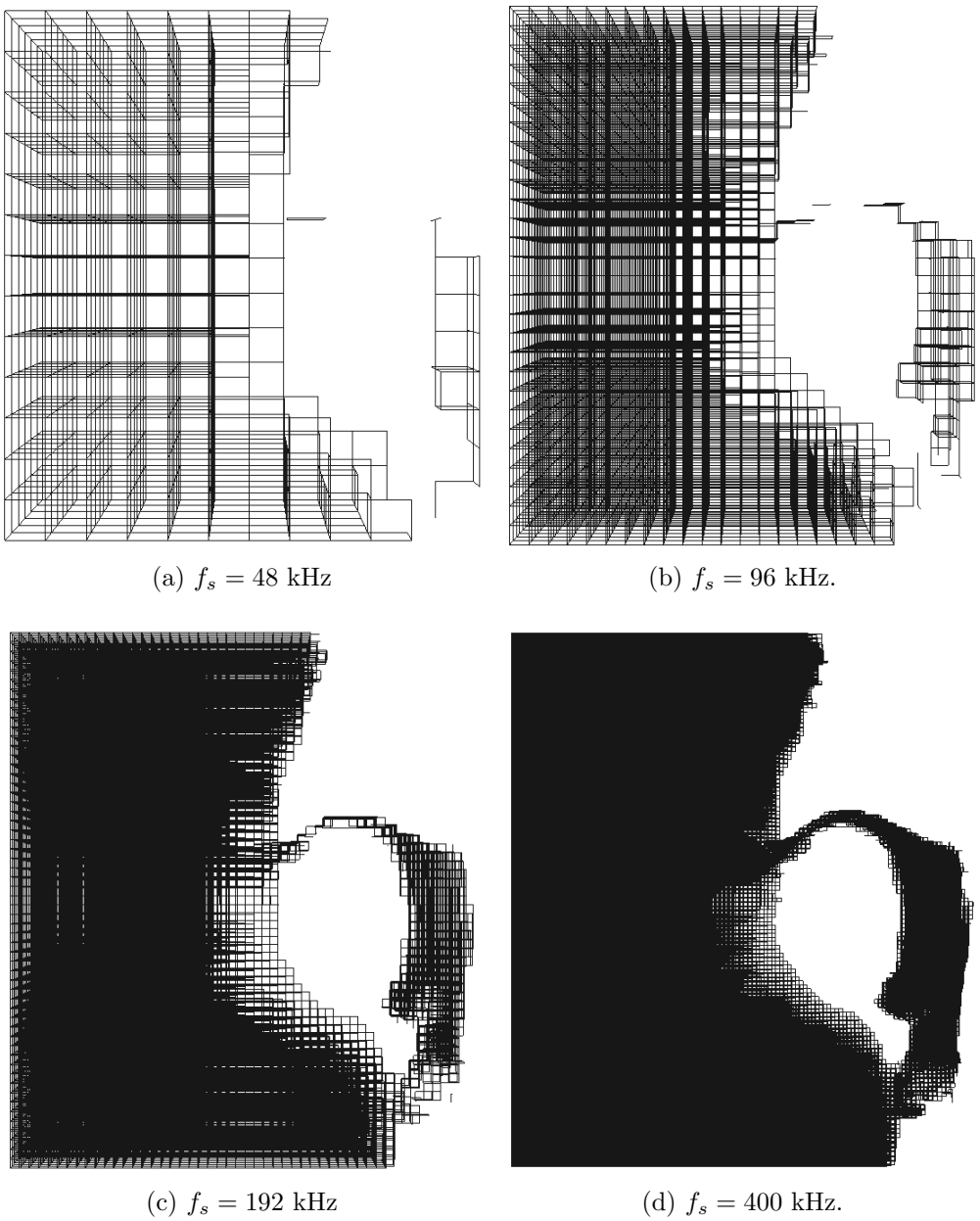
The segmentation volume is allowed to expand beyond the mouth and out to the limits of the MRI image, resulting in a roughly cuboid volume of air (visible in Figure 5.5c) coupled with the internal vocal tract airway. This air volume allows for realistic radiation behaviour at the lips. The edges of this cuboid are modelled with approximately anechoic boundaries as described in Section 5.3.

After the segmentation is complete, it must then be converted into a rectilinear 3D mesh. This step makes use of the custom code described in [166] to fit a Cartesian grid into the stencil created by the segmentation data. This process may be completed at any sampling frequency, and generates a series of points which represent the physical locations of scattering junctions in the digital waveguide mesh algorithm (see Section 4.1.4), as illustrated in Figure 5.5d.

Choosing the temporal sampling frequency for the model requires careful consideration. Since the DWM is equivalent to an FDTD model operating at the limit of the CFL condition (4.6), the equivalent physical length of unit waveguides, Δx, in the DWM is related to the sampling frequency by the following relationship:

$$\Delta x = \frac{c\sqrt{D}}{f_s} \tag{5.5}$$

where c is the speed of sound in the medium, D is the dimensionality of the system (for the proposed model, $D = 3$), and f_s is the sampling frequency. As a result, increasing f_s reduces the waveguide length Δx, providing better spatial resolution but also generating more scattering junctions in the volume. This results in increased computational complexity since calculations must be performed for every scattering junction at every time step. There is an additional lower limit on sampling frequency for vocal tract data, illus-



(a) $f_s = 48$ kHz

(b) $f_s = 96$ kHz.

(c) $f_s = 192$ kHz

(d) $f_s = 400$ kHz.

Figure 5.6: Grids for phoneme /ɪ/ with different sampling frequencies, with spatial step size given by (5.5).

trated in Figure 5.6, due to the small dimensions involved. If the sampling frequency is too low, the mesh becomes discontinuous, as seen in Figure 5.6a and 5.6b. Even at 192 kHz (Figure 5.6c), there are parts of the vocal tract represented by a single layer of scattering junction locations, equivalent to a physical depth of zero. As a result, the sampling frequency chosen for this study is 400 kHz, as illustrated in Figure 5.6d, giving a spatial resolution of approximately 1.52 mm. At this grid resolution there are at least two scattering junctions, and hence at least one waveguide, in every dimension at a constriction for all the phonemes under study. This grid spacing therefore provides an appropriate trade-off between spatial resolution and computational expense for the synthesis of diphthongs. It is also on the same order as the MRI scan resolution (2 mm), so increasing the resolution of the mesh further is expected to have limited benefit. Nevertheless, the same MRI scans were used in [166] with $f_s = 960$ kHz. It should be noted that the constrictions in the vocal tract during consonant articulation may be narrower than those for vowels, and therefore the grid resolution may indeed need to increase when the model is extended to include consonants.

5.4.2 Admittance Map Construction

After preprocessing the data, the next stage is to generate an admittance matrix, $\mathbf{Y}$, for use in the DWM algorithm (4.7)–(4.10). Unlike the raised-cosine admittance maps described in Section 5.2.1, the proposed model uses the specific acoustic admittance of the air and surrounding head tissue.

The volume matrices described in Section 5.5.1 consist of a regular arrangement of scattering junction locations, with values of either one (when the junction is located within the airway) or zero (when the junction is located within the tissue of the head). However, in a DWM it is the *waveguides*—the links *between* scattering junctions—that have a physically meaningful admittance; the scattering junctions represent the infinitesimally small points at which these waveguides meet. Therefore, in order to generate an admittance matrix it is necessary to perform a complete interrogation of the connections between scattering junctions to determine the appropriate admittances.

If two neighbouring volume matrix elements are both located within the airway, the waveguide connecting the junctions is assigned the admittance of air, $Y_{air} = 1/Z_{air}$, where the impedance of air is related to the speed of sound in air, c_{air} and the density of air ρ_{air} following (2.3). Following previous studies [21, 11], the values $c_{air} = 350\ \mathrm{m\,s^{-1}}$ and $\rho_{air} = 1.14\ \mathrm{kg\,m^{-3}}$ are used, giving $Z_{air} = 399\ \mathrm{Pa\,s\,m^{-3}}$. If two neighbouring volume matrix elements are both located within the head tissue, the connecting waveguides are assigned the admittance of the tissue forming the vocal tract wall, $Y_{walls} = 1/Z_{walls}$ where $Z_{walls} = 83666\ \mathrm{Pa\,s\,m^{-3}}$ [21, 11]. In the final case, where neighbouring volume matrix elements span the air/tissue interface Γ_w, the connecting waveguide is assigned the admittance Y_{walls}. This gives a tissue boundary location accurate to within the length of one unit waveguide, which at 400 kHz is approximately 1.52 mm according to (5.5). The MRI corpus is resampled from a 2 mm isotropic image, so this level of spatial resolution is appropriate given the data available.

The process described above is repeated for every volume matrix element and connection direction. For a 3D rectilinear mesh, this results in six admittance matrices, which for ease of conceptualization are termed $\mathbf{Y}_{north}$, $\mathbf{Y}_{south}$, $\mathbf{Y}_{east}$, $\mathbf{Y}_{west}$, $\mathbf{Y}_{front}$ and $\mathbf{Y}_{back}$. The matrices are constructed such that $Y_{north}(x, y, z)$ represents the admittance in the waveguide directly north of the junction with index (x, y, z), and $Y_{south}(x, y, z)$ represents the admittance in the waveguide directly south of this junction, such that $Y_{north}(x+1, y, z) = Y_{south}(x, y, z)$. Once these admittance maps have been populated, moving between vocal tract shapes is simply a matter of interpolating between maps over the duration of a simulation, as discussed in Section 5.7.1.

The final stage in the construction of the model is the definition of model boundaries. Figure 5.7 illustrates a midsagittal slice through an example simulation domain. The external boundary Γ_D may be further split into domain edges occurring within air, Γ_{DA}, and domain edges within head tissue, Γ_{DT}. An anechoic LRW boundary is implemented on Γ_{DA} and Γ_{DT}, as described in Section 5.3. The vocal tract wall Γ_W is not implemented as a domain boundary in the proposed model; instead the difference in admittance that occurs

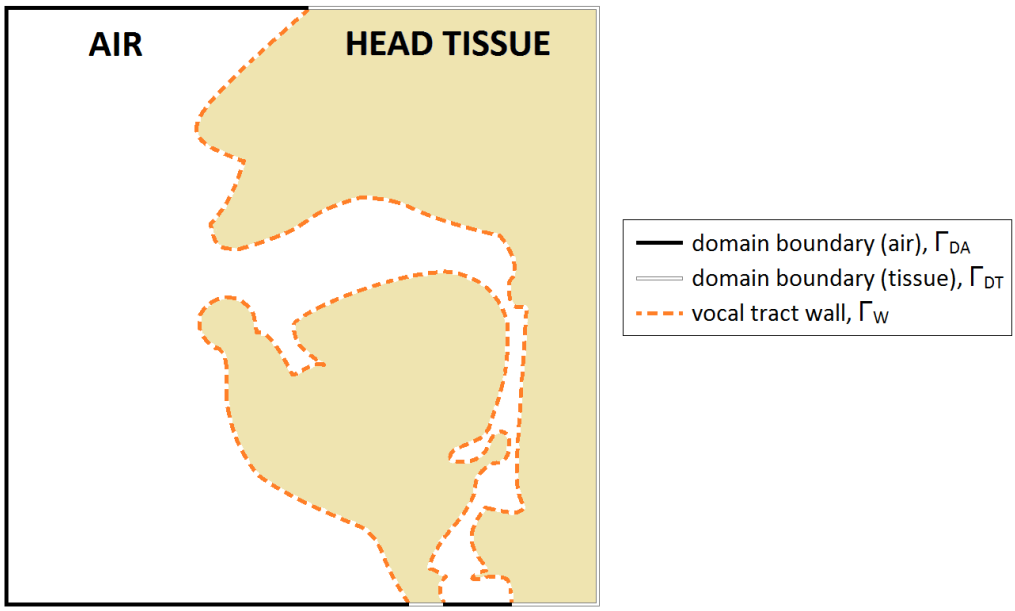


Figure 5.7: Location and definition of domain boundaries and vocal tract wall for simulations. Midsagittal slice through a 3D volume representing vowel /ɔ/ is shown.

at Γ_W causes reflection of sound waves back into the vocal tract airway. In contrast, the static 3D DWM model [166] implements domain boundaries on Γ_W and Γ_{DA}, and does not simulate wave propagation through the tissue of the head.

The proposed model makes use of LRW boundaries with an approximately anechoic condition achieved by setting the normalized admittance parameter G to one. This boundary condition is known not to perfectly reproduce ideal anechoic behaviour [175]. However, the behaviour of the boundaries is explored further in Section 5.5.1 and found to be sufficiently accurate for the current study.

Once the admittance maps and boundary implementation are complete, the propagation behaviour within the model is completely defined.

5.5 Model Refinement

The creation of an admittance map, in addition to the boundary implementation described in Section 5.3, results in a complete vocal tract model. However, a number of refinements are required in order to improve the accuracy of the model. These refinements are made based on calculation of the vocal tract transfer function (VTTF), according to (2.16). Following [30], the VTTF is calculated using a Gaussian pulse, $u_{in}(t)$, as the volume velocity source, as follows:

$$u_{in}(t) = e^{-[(\Delta t n - T)/0.29T]^2} \tag{5.6}$$

where $\Delta t = 1/f_s$, $T = 0.646/f_0$ and $f_0 = 20$ kHz, providing sufficient excitation across the entire audible frequency range of 0–20 kHz. However, for clarity, and direct comparison with other simulations such as [30] and [21], only the frequency range 0–10 kHz is displayed in the VTTF plots presented throughout this thesis. This range is sufficient to describe speech intelligibility, and much of the information relating to naturalness [79].

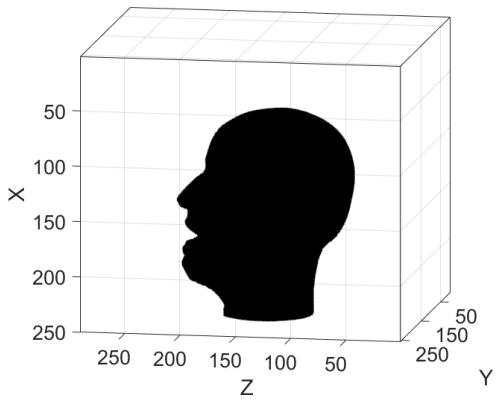
5.5.1 Mesh Alignment and Extent

The vocal tract grid data are read into MATLAB as 3D binary matrices, with ones representing vertexes within the airway, and zeros representing vertexes within the tissue of the head. These matrices—one for each of the six phonemes under study—provide a complete description of the vocal tract geometry and will henceforth be referred to as *volume matrices*. The volume matrix for each phoneme is inspected and adjusted so that fixed anatomical structures such as the hard palate and the nose are aligned across phonemes. Finally, the phoneme-specific volume matrices are combined with a volume matrix corresponding to an idealized human head, scaled and transformed to match the size and alignment of the MRI subject. This mesh is obtained from a 3D scan of a KEMAR mannequin [177] and fitted with a Cartesian grid following the same procedure as the vocal tract segmentations. This step is necessary as the MRI data only extends approximately 4 cm either

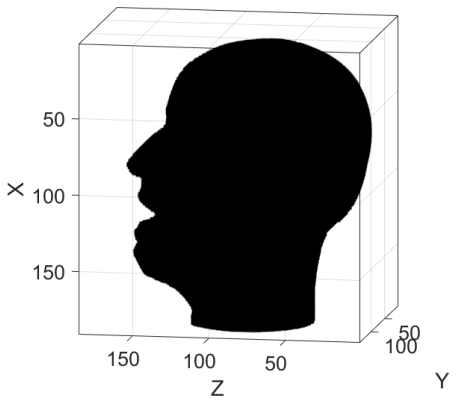
side of the midsagittal plane so the remaining head geometry is unknown. The idealised head provides an appropriate volume and shape in the area beyond the MRI scan data, while retaining subject-specific geometry such as the nose and particularly the lips, known to be essential for accurate vocal tract acoustics [21].

The *radiation volume*—the volume of air outside the lips into which speech sounds radiate—is a critical aspect of vocal tract simulations, and accurate synthesis requires this volume to be taken into account [21, 143]. However, too large a simulation domain results in high computational cost for potentially small increases in accuracy. Therefore, simulations were performed to determine how much the radiation volume can be constrained without introducing significant errors in the VTTF. Four cases were investigated, as illustrated in Figure 5.8. In every case, the bottom of the simulation domain was fixed at a location approximately 1 cm below the larynx position for the articulation of /ɔ/, which had the lowest larynx position of all six articulations studied. In Case 1, the entire head was considered, with 10 cm air surrounding it in all directions; Case 2 also features the entire head but with 1 cm air at either side and at the back, and 3 cm air in front; Case 3 is similar to Case 2 but with the back of the head removed, from 1 cm behind the back of the pharyngeal wall; and Case 4 consists of only the extent of the original MRI image, with 3 cm air in front of the face, extending approximately 4 cm either side of the midsagittal plane, and up to the top of the nose. In all cases the distances given in cm above are rounded to an integer number of waveguide lengths.

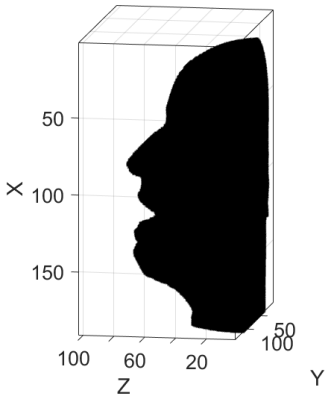
The resulting VTTFs for the vowel /a/ are presented in Figure 5.9. It is apparent from Figure 5.9 that in general the VTTFs are very similar for each of the four cases. Cases 1 and 2 in particular exhibit almost identical VTTFs, with less than 1 dB difference across the entire range 0-20 kHz. Some small errors are introduced for Cases 3 and 4, primarily affecting the depth of spectral dips. However, Case 4 introduces further errors, with a 3 dB difference in the first two formant magnitudes compared to Case 1, and a large deviation below 500 Hz. For this reason, volume matrices cropped according



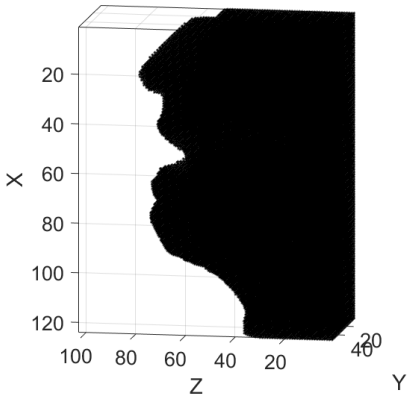
(a) Case 1: whole head with extended radiation volume



(b) Case 2: whole head with limited radiation volume



(c) Case 3: front hemisphere only



(d) Case 4: MRI extent only

Figure 5.8: Volume matrices for phoneme /a/ with different radiation volumes. Each volume matrix contains a 3D Cartesian grid of points whose extent is indicated by the surrounding boxes; black points represent scattering junction locations that exist within the head tissue.

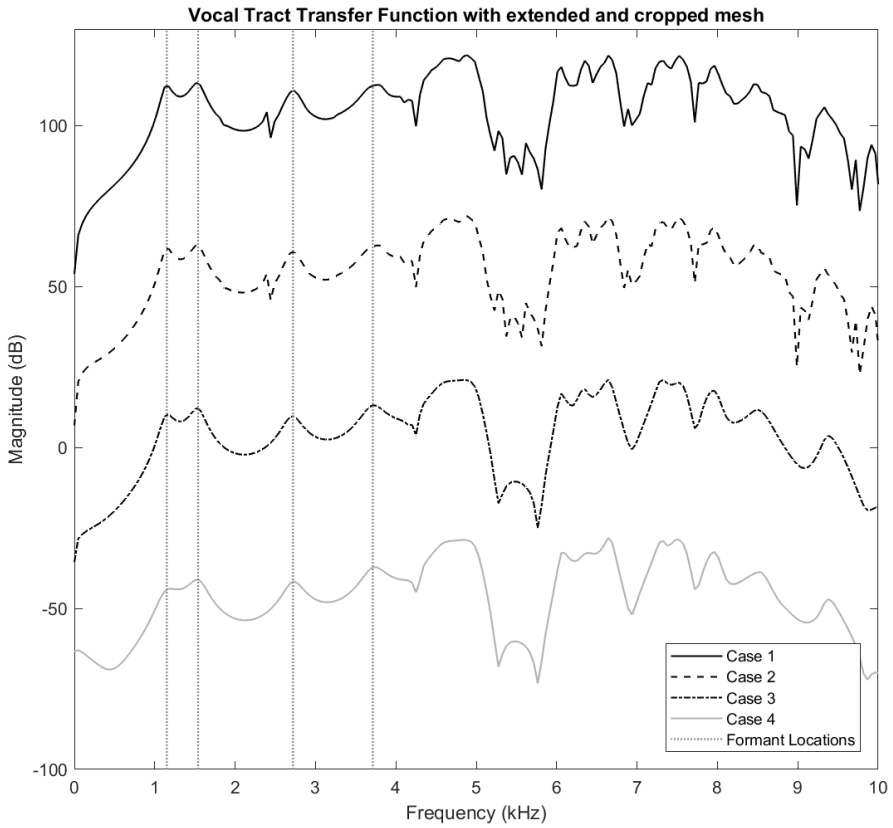


Figure 5.9: Vocal tract transfer functions for phoneme /a/ in each radiation volume case. Vertical dotted lines illustrate the positions of the first 4 formants. An artificial 50 dB offset has been added between VTTFs for clarity of illustration.

to Case 3 are used throughout the remainder of this study, as they provide an appropriate trade-off between model accuracy and simulation domain size, and hence computation time.

The similarity of Cases 1 and 2 is notable as it suggests that the LRW anechoic boundary implementation used on $\Gamma_{DA:DT}$—which is much closer to the receiver position in Case 2 than Case 1—does not have a significant effect on the simulated transfer functions. It can therefore be assumed that the LRW produces boundaries that are sufficiently close to anechoic for the purposes of the current study.

5.5.2 Source and Receiver Positions

During vowel production, a source signal is generated at the larynx, and speech is output at the lips. In simulations of the vocal tract, the aim is to replicate this behaviour. The larynx moves between vowel articulations, so an ideal simulated source would also move within the simulation domain.

Most finite element vocal tract models (e.g. [21]) apply a source signal across the entire cross-sectional area of the glottal opening, known as the glottal boundary Γ_G. As the simulation domain changes shape, so does the location of Γ_G and hence the source position. In DWM simulations, the source signal is inserted at one or more scattering junction locations. Changing the junction(s) at which a signal is input during the course of the simulation introduces audible artifacts in DWM simulations. Instead, a single scattering junction, J_{input}, is selected as the excitation point for all the phonemes under study.

Selecting an excitation point requires careful consideration, as the larynx height within the mesh varies depending on the phoneme, with /a/ having the highest larynx position, and /ɔ/ having the lowest, in the phonemes under study. To ensure that J_{input} is not located below the laryngeal opening for any of the phonemes, it is placed at a level corresponding to the highest larynx location, namely that of /a/, and is therefore also within the airway for the other five phonemes under study.

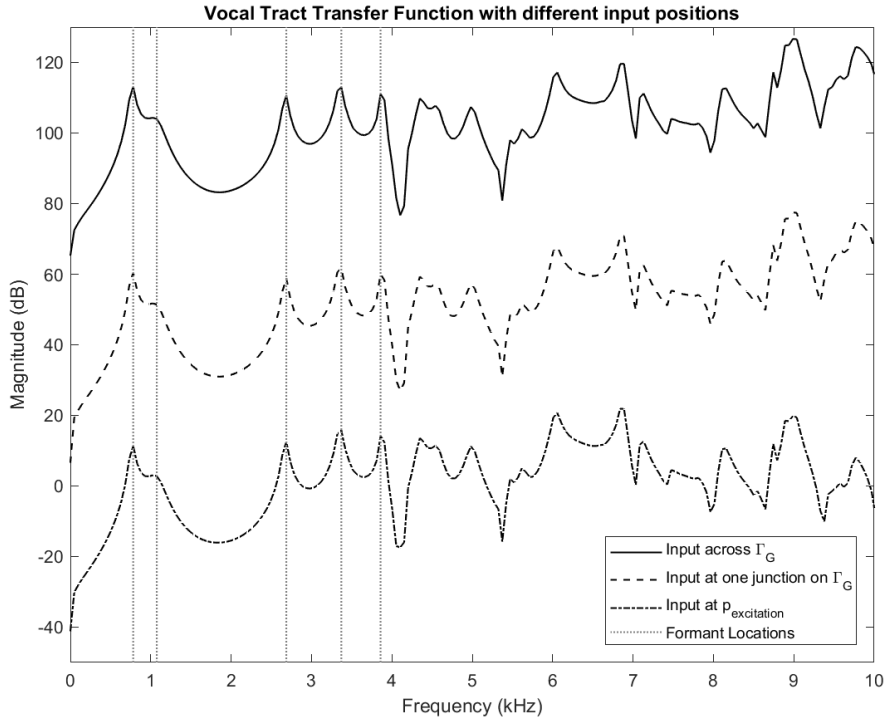


Figure 5.10: Vocal tract transfer functions for phoneme /ɔ/ with local and global source positions.Vertical dotted lines illustrate the positions of the first 5 formants. An artificial 50 dB offset has been added between VTTFs for clarity of illustration.

To investigate the error caused by locating the source at J_{input} compared to the phoneme-specific Γ_G, simulations were performed for the phoneme /ɔ/, which has the maximum difference in height between the two source positions for the phoneme under study, approximately 9 mm. The resulting VTTF when the source signal is input at J_{input} was compared to that performed with the source signal input at the scattering junctions representing Γ_G. An additional simulation was performed using a single point located on Γ_G to determine whether any of the error can be attributed to the use of a single scattering junction, as opposed to simply the difference in source height between J_{input} and Γ_G. The results of these simulations can be seen in Figure 5.10, and show that, although inputting the source signal at J_{input} does result in some error in the VTTF when the phoneme in question has a lower larynx position, the frequencies of the formants are accurately reproduced. The formant magnitudes are also within 6 dB of those generated with the input applied across Γ_G—which is assumed to be the most accurate case—in the region below 9 kHz. This is considered sufficiently accurate for the current model, given the more significant sources of error in the simulated VTTF such as the absence of a nasal cavity. As these simulations used the phoneme with the greatest difference in height between the actual larynx position and J_{input}, errors in the VTTF related to source position are assumed to be smaller for the other phonemes under study. The VTTF generated using a single point on Γ_G as the input location also results in the correct formant frequencies and less than 2 dB error across the majority of the audio bandwidth, indicating that source height rather than width is the major contributing factor to errors in the VTTF when using J_{input} as the source position.

In addition to J_{input}, a standard receiver position J_{output} is also selected, as a single point at an on-axis position level with the tip of the nose, similar to a real, close microphone position.

5.6 Monophthong Synthesis

The proposed method must be compared to recorded voice data and existing DWM synthesis techniques in order to confirm its accuracy. This section describes the procedures that were undertaken to this end, using static vowel articulations.

5.6.1 Procedure

The proposed method is compared to two existing DWM vocal tract simulation techniques: the dynamic 2D model [147] described in Section 5.2.1, henceforth referred to as the *2DD-DWM* model, and the static 3D model [166] described in Section 5.1.2, henceforth called the *3DS-DWM* model. These are compared with the proposed dynamic 3D model, labelled the *3DD-DWM* model. For the current comparison, fixed vowel articulations are used, allowing static and dynamic modelling approaches to be compared. Simulations are performed using the 3DD-DWM procedure outlined in Section 5.4 for each of the six monophthongs required to make up the English diphthongs: /e/, /a/, /ɪ/, /ɔ/, /ə/, and /ʊ/.

A 3DS-DWM simulation is performed following [166], as discussed in Section 5.1.2, based on the same volume matrix as the 3DD-DWM simulation. An approximately anechoic LRW boundary is set up on Γ_{DA}, and a reflecting boundary is set up on Γ_W with a reflection coefficient of 0.99. This value was found to provide a suitable formant bandwidth in agreement with [166].

The final simulation method used in the comparisons is the 2DD-DWM method, discussed in Section 5.2.1. This method simplifies the vocal tract by treating it as a concatenation of cylindrical tubes with cross-sectional areas obtained from the vocal tract geometry. In order to create 2DD-DWM models that are comparable to the 3DD-DWM and 3DS-DWM models, the same 3D MRI data was used. This data was converted to cross-sectional area data following the iterative bisection procedure described in [20]. The 2DD-DWM model was set up following the method in [147], as discussed in Section

5.2.1. The same sampling frequency of 400 kHz was used for the 2DD-DWM simulations, resulting in a spatial resolution of 1.24 mm according to (5.5). The 2DD-DWM model does not include provision for radiation at the lips, nor energy lost at the glottis; instead a specific reflection coefficient is set at either end of the mesh to approximate this behaviour. Following the recommendation of [146], this lip reflection coefficient is set to -0.9, the glottal reflection coefficient to 0.97, and the reflection coefficient at mesh boundaries to 0.92.

5.6.2 Results and Discussion

The results are presented across three sections. The locations of spectral peaks (formants) are determined from VTTFs, with lower formants being necessary for vowel identification and higher formants contributing to naturalness. Spectral dips or antiresonances are also obtained from VTTFs; these are caused by side branches of the vocal tract and help to validate the accuracy of the simulations, as well as contributing to speaker identification [26]. Finally, power spectral density (PSD) curves are calculated for the recorded and simulated monophthongs in order to evaluate the broader spectral characteristics of the simulations.

Formant Locations

The calculation of VTTFs is performed according to the procedure in Section 5.5. However, the true VTTFs for the human subject remain unknown. In order to determine the accuracy of formant reproduction in the simulations, the peaks in the VTTF are compared to two different formant measurements from the corresponding recorded speech sample: peaks in the spectrum produced by linear predictive coding (LPC) analysis of the recording, and formant values obtained using the automatic formant detection algorithm in the software package Praat [178]. The LPC provides an approximate spectrum but is subject to certain limitations—for instance, it considers the vocal tract to be an all-pole filter, ignoring the effect of side branches—so the Praat

		2DD-DWM						3DS-DWM						3DD-DWM					
Vowel		F1	F2	F3	F4	F5	**M.A.**	F1	F2	F3	F4	F5	**M.A.**	F1	F2	F3	F4	F5	**M.A.**
a	Hz	-167	-163	-178	100	-460		-44	273	-386	124	-485		322	8	-20	320	-9	
	%	-19.93	-10.74	-6.52	2.94	-9.64	**9.95**	-5.25	-17.98	-14.14	3.64	-10.17	**10.24**	38.42	0.53	-0.73	9.40	-0.19	**9.85**
e	Hz	-6	-222	-145	-265	-298		55	-271	-206	-387	-665		226	-15	160	126	-286	
	%	-1.01	-12.10	-5.48	-7.68	-6.55	**6.57**	9.29	-14.78	-7.78	-11.21	-14.63	**11.54**	38.18	-0.82	6.04	3.65	-6.29	**11.00**
ɪ	Hz	23	-1058	-334	-167	23		-38	-191	-175	-423	-746		169	126	325	16	-318	
	%	5.69	-52.95	-12.47	-4.86	0.51	**15.30**	-9.41	-9.56	-6.53	-12.30	-16.56	**10.87**	41.83	6.31	12.14	0.47	-7.06	**13.56**
ɔ	Hz	117	504	4	–	313		129	100	-448	–	-1115		300	394	-191	–	-395	
	%	24.95	76.83	0.14	–	7.32	**27.31**	27.51	15.24	-15.64	–	-26.07	**21.11**	63.97	60.06	-6.67	–	-9.24	**34.98**
ʊ	Hz	-12	516	-47	-28	-326		-24	-34	-633	-504	-1010		244	235	-218	21	-448	
	%	-2.66	64.26	-1.76	-0.89	-7.81	**15.48**	-5.32	-4.23	-23.69	-16.11	-24.21	**14.71**	54.10	29.27	-8.16	0.67	-10.74	**20.59**
ə	Hz	34	185	-110	–	249		22	-108	-354	–	-728		179	172	-24	–	29	
	%	6.31	15.48	-4.17	–	5.92	**7.97**	4.08	-9.04	-13.42	–	-17.30	**10.96**	33.21	14.39	-0.91	–	0.69	**12.30**
M.A.	Hz	59.8	441.3	136.3	140.0	278.2		52.0	162.8	367.0	359.5	791.5		240.0	158.3	156.3	120.8	247.5	
	%	10.09	38.73	5.09	4.09	6.29	**12.86**	10.14	11.81	13.53	10.82	18.16	**12.89**	44.95	18.56	5.77	3.55	5.70	**15.71**

Table 5.1: Error in formant frequency between simulations and LPC spectrum of recorded speech, presented in both Hz and % (M.A. is mean absolute error). Figures in bold are the mean absolute % error score for all five formants in each vowel.

		2DD-DWM						3DS-DWM						3DD-DWM					
Vowel		F1	F2	F3	F4	F5	**M.A.**	F1	F2	F3	F4	F5	**M.A.**	F1	F2	F3	F4	F5	**M.A.**
a	Hz	-209	-198	178	-67	-439		-86	-308	-386	-43	-464		280	-27	-20	153	12	
	%	-23.75	-12.75	-6.52	-1.88	-9.24	**10.83**	-9.77	-19.83	-14.14	-1.20	-9.77	**10.94**	31.82	-1.74	-0.73	4.28	0.25	**7.77**
e	Hz	-21	-278	-270	-343	-352		40	-327	-331	-465	-719		211	-71	35	48	-340	
	%	-3.46	-14.71	-9.74	-9.72	-7.65	**9.06**	6.59	-17.30	-11.94	-13.18	-15.63	**12.93**	34.76	-3.76	1.26	1.36	-7.39	**9.71**
ɪ	Hz	-12	-1097	-386	-278	-135		-73	-230	-227	-534	-904		134	87	273	-95	-476	
	%	-2.73	-53.85	-14.14	-7.83	-2.89	**16.29**	-16.63	-11.29	-8.32	-15.04	-19.38	**14.13**	30.52	4.27	10.00	-2.68	-10.21	**11.54**
ɔ	Hz	126	469	96	526	494		138	65	-356	-243	-934		309	359	-99	404	-214	
	%	27.39	67.87	3.46	17.89	12.06	**25.73**	30.00	9.41	-12.84	-8.26	-22.80	**16.66**	67.17	51.95	-3.57	13.74	-5.22	**28.33**
ʊ	Hz	-21	502	-63	-92	-83		-33	-48	-649	-568	-767		235	221	-234	-43	-205	
	%	-4.57	61.44	-2.34	-2.88	-2.11	**14.67**	-7.17	-5.88	-24.14	-17.79	-19.52	**14.90**	51.09	27.05	-8.71	-1.35	-5.22	**18.68**
ə	Hz	-2	114	-189	46	177		-14	-179	-433	-467	-800		143	101	-103	83	-43	
	%	-0.35	9.00	-6.96	1.37	4.14	**4.36**	-2.43	-14.14	-15.94	-13.95	-18.70	**13.03**	24.87	7.98	-3.79	2.48	-1.00	**8.02**
M.A.	Hz	65.2	443.0	197.0	225.3	280.0		64.0	192.8	397.0	386.7	764.7		218.7	144.3	127.3	137.7	215.0	
	%	10.37	36.61	7.19	6.93	6.35	**13.49**	12.10	12.97	14.55	11.57	17.63	**13.77**	40.04	16.12	4.68	4.31	4.88	**14.01**

Table 5.2: Error in formant frequency between simulations and formants obtained from Praat for recorded speech, presented in both Hz and % (M.A. is mean absolute error). Figures in bold are the mean absolute % error score for all five formants in each vowel.

formant values are used to cross-check the results.

The errors of the first five formants relative to corresponding recorded speech are presented for each simulation method, in both absolute frequency value and percent, in Tables 5.1 and 5.2. The calculated vocal tract transfer functions (VTTFs) are presented in Figures 5.11 to 5.16. Tables 5.1 and 5.2 give similar error values for each simulation method, with LPC analysis missing the F4 values for /ɔ/ and /ə/ that were obtained with Praat. Since the analysis from Praat includes F1–F5 for all 6 vowels under study, these values (from Table 5.2) will be used throughout the following discussion; nevertheless the values obtained from LPC analysis are included as the LPC spectrum is more easily visualised and is therefore incorporated in Figures 5.11–5.16.

It is apparent from both tables that simulation accuracy varies between vowels for all simulation methods. In particular, simulation of the vowel /ɔ/ results in large errors for every method, suggesting that the subject may not have been consistent in their articulation between the MRI scan and audio recording. The simulation errors vary even for the other five vowels, raising the important point that the accuracy of 3D vocal tract models may be vowel-specific. In addition to the possibility that the subject's articulation may have differed between the MRI and audio data collection procedures, which may affect different vowels to differing degrees, the segmentation and voxelisation procedures may also be a source of vowel-specific errors in vocal tract geometry.

It is worth noting that very few 3D vocal tract models in the literature compare their output to recorded speech, perhaps due to the lack of suitable speech data from the same subject used for MRI scans. The only available comparisons appear to be from [164] (FDTD), [166] (DWM) and [179] (FEM). Simulations in [164] produce a mean absolute error of 6.07%, for the vowel /a/ only, using a 3D FDTD vocal tract model, although the origin of the speech data used for comparison is not clear. This is of comparable magnitude to the mean error of 7.77% obtained for /a/ using the proposed 3DD-DWM method, while the 2DD-DWM and 3DS-DWM simulations have higher mean average errors of 10.83% and 10.94%, respectively, for the same

simulation. The 3DS-DWM method in [166] provides formant errors in Hz rather than percentages, but with the exception of F1 which will be discussed shortly, these are comparable to and sometimes larger than the formant errors observed for the 3DS-DWM simulation in the present study. Finally, [179] indicates highly vowel-dependent results, but across the four vowels studied the mean absolute error is 12.3%. This compares with a mean average error of 14.01% across all 6 vowels for the proposed model, but if the results for /ɔ/ are omitted for the reasons described above, this error reduces to 11.14%. It can therefore be said that the 3DD-DWM model is at least comparable to, if not an improvement upon, several existing 3D vocal tract models.

The value of the first formant—except in the case of /ɔ/ noted above—is generally underestimated by the 2DD-DWM and 3DS-DWM methods, whereas in the proposed 3DD-DWM method, F1 is generally overestimated by a significant margin. The frequency of the first formant is known to increase when yielding walls are taken into account [12], which may help to explain why the 3DD-DWM model, where sound is allowed to propagate through the vocal tract walls, results in a higher F1 value than the other simulations, which feature effectively hard walls with simple losses. The values of F1 for the proposed model remain higher than for the recordings, indicating that the value chosen for the wall impedance in Section 5.4.2 may be too low, despite being a common choice in comparable simulation methods [21, 11]. In general, however, the values of the higher formants F2–F5 are more accurately reproduced in the proposed model than the 2DD-DWM and 3DS-DWM models. This result suggests that frequency-dependent impedances may be a necessary addition to the proposed model in order to more accurately model F1 without reducing the accuracy of of F2–F5 reproduction. This finding is in agreement with previous studies [180]. Future versions of the model will incorporate filters to approximate frequency-dependent behaviour at the vocal tract walls; in the meantime, audition indicates that the high F1 values do not affect vowel identification.

With the previous explanations in mind, the mean absolute errors for the higher formants F2-F5 using the proposed model, averaged across the 6

Mean Absolute Formant Error (%)	2DD-DWM	3DS-DWM	3DD-DWM
F1–F5	13.49	13.77	14.01
F2–F5	14.27	14.18	7.50
F2–F5 excluding outlier /ɔ/	12.06	14.35	5.28

Table 5.3: Mean absolute error values for formants 1–5 in each simulation method compared with recordings, across all diphthongs. When the first formant F1 is excluded, the proposed 3DD-DWM model exhibits the lowest error. Removing the outlier results for monophthong /ɔ/ reduces the error of the proposed model further.

monophthongs, are compared to those of the 2DD-DWM and 3DS-DWM in Table 5.3. Additionally, results are provided with /ɔ/ omitted from the averages for the reasons described aboce. It is clear from these results that the proposed model offers a significant increase in accuracy over the comparison models for formants F2–F5. It is also important to note that formants above F3 contribute to the perception of naturalness [12], suggesting that the proposed method offers the most natural-sounding output in terms of formant locations. The comparison simulations show consistently larger errors in these higher formants and the 2DD-DWM method in particular shows errors of greater than 50% for F2 for some vowels, which may affect intelligibility. The improvement of the 3DD-DWM model over the 3DS-DWM model is especially interesting, as the two models use identical geometry. It must therefore be concluded that the improvements in higher formant reproduction occur, at least in part, due to the vocal tract walls as having some depth in the proposed model through which sound can propagate, as in the real vocal tract. The 3DS-DWM, by comparison, simply imposes a loss factor at the vocal tract walls, but the domain does not continue outside this boundary, giving the vocal tract walls an equivalent physical depth of zero. However, this apparent depth in the vocal tract walls of the proposed model also appears to be the main cause of the upward shift in F1 noted for the proposed model.

The simulated VTTFs, illustrated in Figures 5.11–5.16, provide more detail about the three simulation methods. It is immediately apparent that,

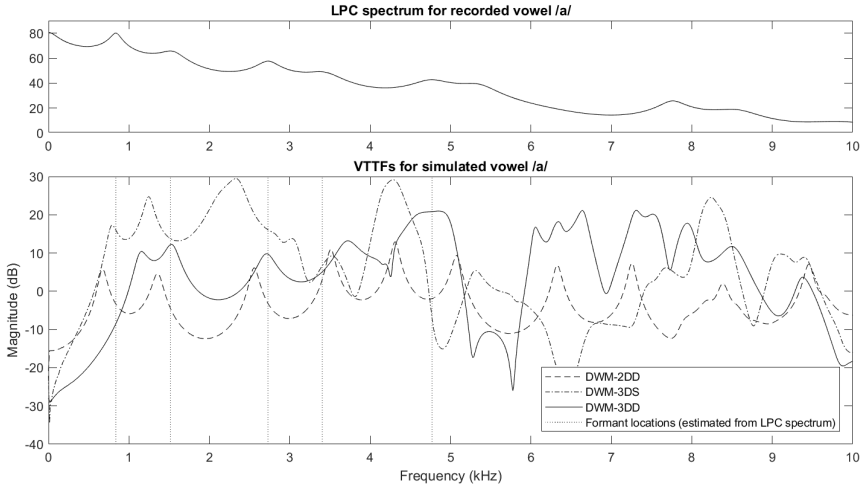


Figure 5.11: LPC spectrum of recording (top) and simulated vocal tract transfer functions (bottom) for phoneme /a/.

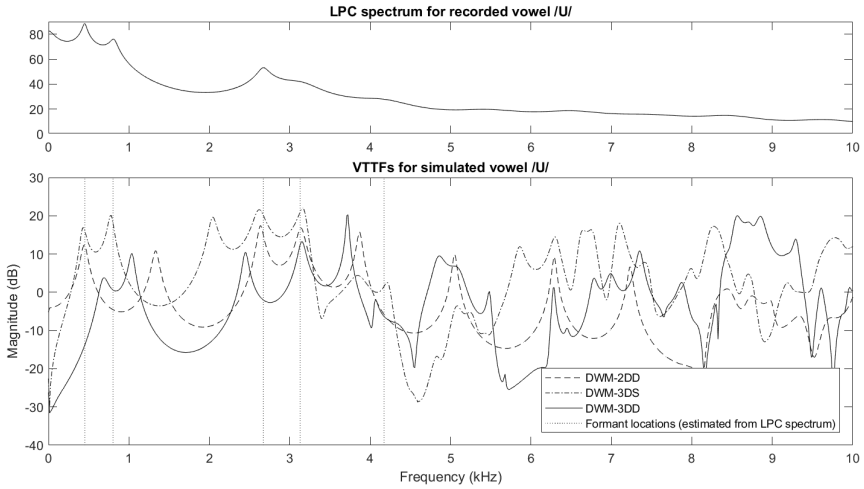


Figure 5.12: LPC spectrum of recording (top) and simulated vocal tract transfer functions (bottom) for phoneme /ʊ/.

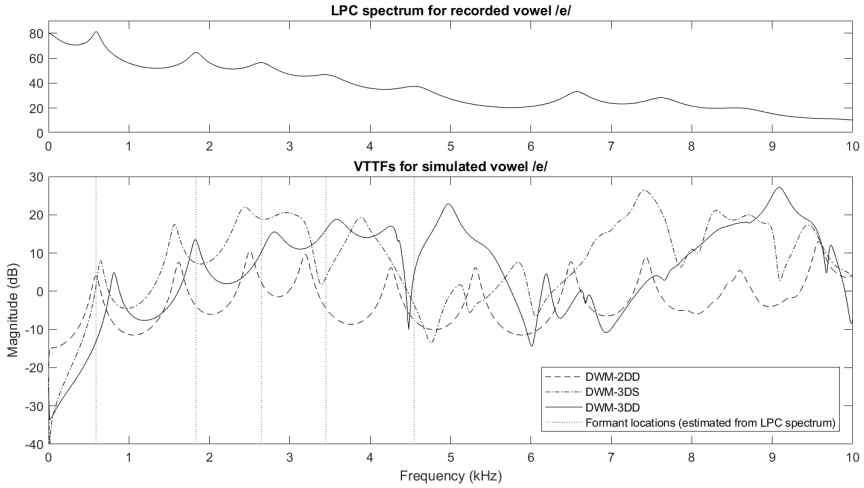


Figure 5.13: LPC spectrum of recording (top) and simulated vocal tract transfer functions (bottom) for phoneme /e/.

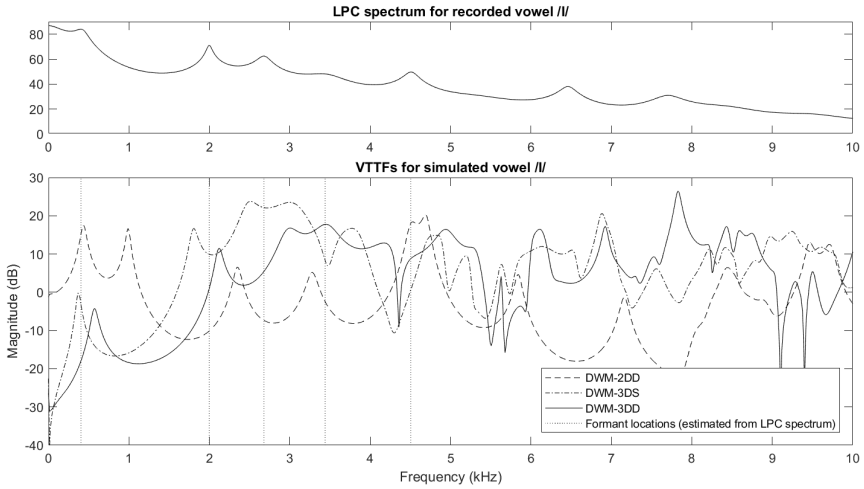


Figure 5.14: LPC spectrum of recording (top) and simulated vocal tract transfer functions (bottom) for phoneme /ɪ/.

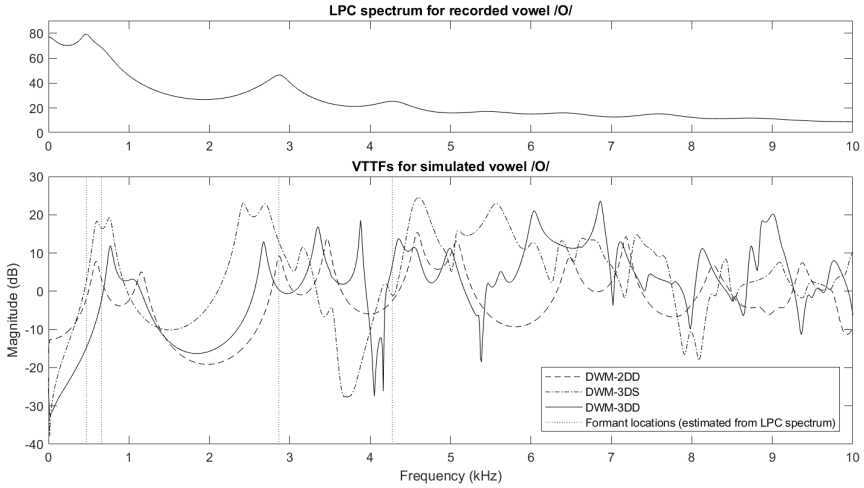


Figure 5.15: LPC spectrum of recording (top) and simulated vocal tract transfer functions (bottom) for phoneme /ɔ/.

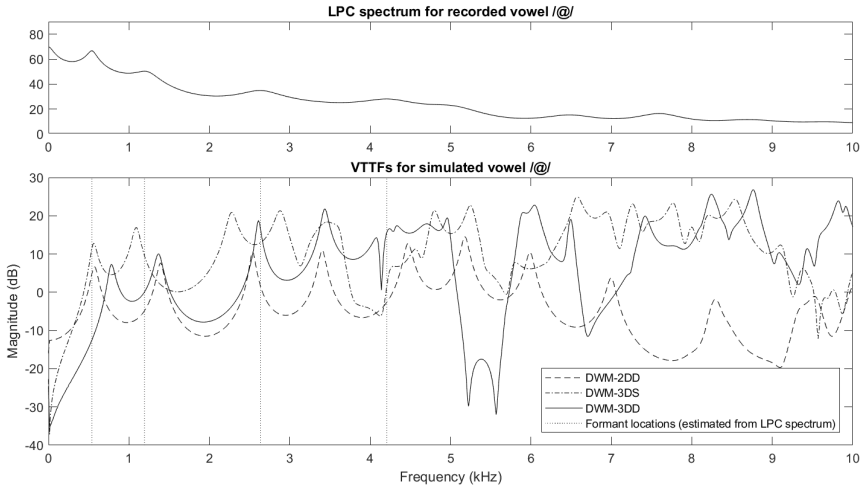


Figure 5.16: LPC spectrum of recording (top) and simulated vocal tract transfer functions (bottom) for phoneme /ə/.

in addition to the differences in formant frequencies described in Tables 5.1 and 5.2, the relative formant magnitudes, and often the formant bandwidths, differ with simulation method. Relative formant magnitudes are important in the perception of naturalness [21], but without a VTTF of the real vocal tract for comparison—planned for future work—the accuracy of formant magnitudes are difficult to assess.

Antiresonance Locations

One clear feature of the 3D VTTFs (see Figures 5.11–5.16) are large spectral dips, occurring at different frequencies depending on the vowel. An example can be seen in Figure 5.16 at around 5.4 kHz in the 3DD-DWM simulation, and a shallower dip at around 4 kHz for the 3DS-DWM simulation. By systematic occlusion of the vocal tract side branches, individually and in combination, the spectral dips are identified as the contribution of the piriform fossae and epiglottic valleculae. Although [30] found the acoustic effects of the epiglottic valleculae to be small, for this subject they are found to contribute significantly to the VTTF, both individually and in combination with the piriform fossae. As the dips are associated with vocal tract side branches, which are not modelled in the 2DD-DWM simulation, they are not present in the VTTFs for the 2D simulation method.

Using the vowel /ə/, which has a range of spectral dips visible in the VTTF, as an example, the contribution of the different side branches to the 3DD-DWM VTTF can be seen. The results of this analysis are presented in Figure 5.17. The piriform fossae appear to introduce a small spectral dip at 4.1 kHz, while the epiglottic valleculae appear to be responsible for the spectral dip at 6.7 kHz. It is the piriform fossae and epiglottic valleculae interacting with one another that produces the large spectral minimum at 5.4 kHz; neither set of side branches completely account for this dip on their own. Note also the shift in F2–F4 caused by occlusion of the side branches, indicating that they have an effect across the whole spectrum.

The 3DS-DWM simulation features spectral dips for the same reasons, but their frequencies are typically shifted lower than those of the 3DD-DWM

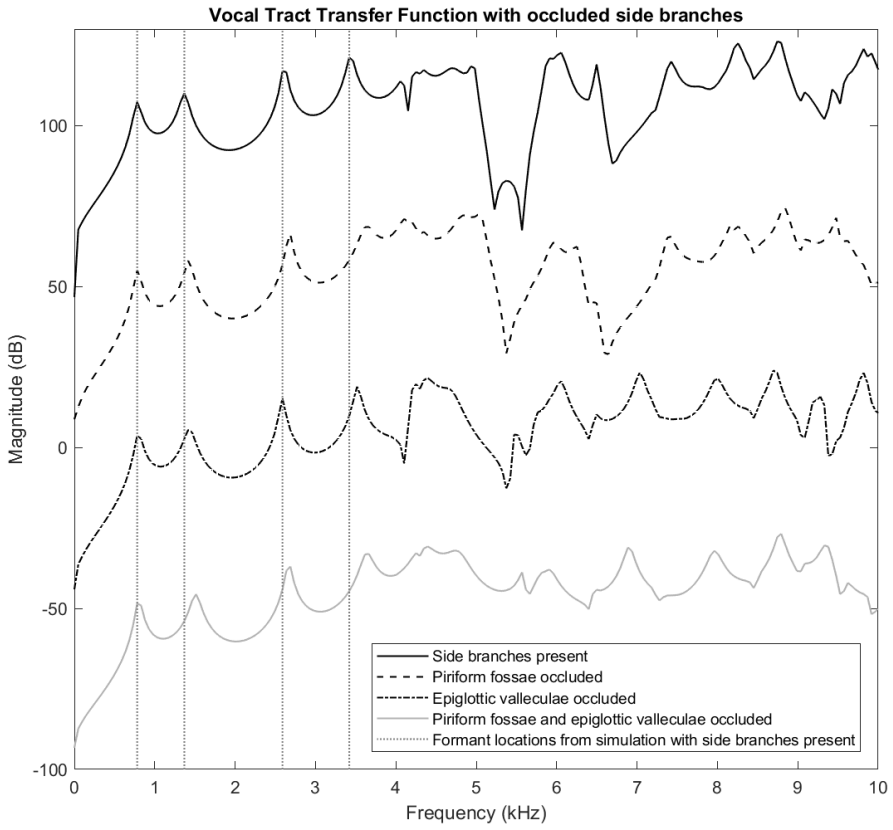


Figure 5.17: Vocal tract transfer functions for phoneme /ə/ with different combinations of side branch occlusion. Vertical dotted lines illustrate the positions of the first 4 formants in the unoccluded simulation. An artificial 50 dB offset has been added between VTTFs for clarity of illustration.

VTTFs. This difference is a result of the boundary implementation on Γ_W in the 3DS-DWM simulation; in contrast, in the 3DD-DWM model, a waveguide with the admittance of tissue spans this interface, effectively reducing the size of the vocal tract by up to one waveguide length in each direction. It is difficult to determine which of the simulated VTTFs is correct without access to a measured VTTF of the subject, although the 3DD-DWM simulation produces piriform fossae dips closer to the expected range of 4–5 kHz [40]. This disparity illustrates an inherent problem in systems using Cartesian meshes, which require a constant spatial sampling interval across the entire domain, making the fine detail of a vocal tract boundary difficult to model. Such problems may be alleviated using an increased sampling frequency, or interpolation of non-Cartesian locations such as in the immersed boundary method [150], but both of these techniques result in increased computational expense.

Spectral Shape

The VTTFs offer insight into the behaviour of the models given an idealised input, but are not directly comparable to recorded speech. In order to produce synthesised speech for comparison, an electrolaryngograph (Lx) signal is used as input to the models. The Lx signal is a measure of vocal fold conductivity, and is inverted to provide a signal approximating the real glottal flow during an utterance [181]. The Lx data was recorded simultaneously with the benchmark audio recordings, and is used as the simulation input to provide the correct pitch contour and amplitude envelope associated with the recorded audio. The use of the recorded Lx source facilitates direct comparison between the simulations and the recordings. An example of an Lx signal used as an input to the models is shown in Figure 5.18, where the slight fluctuations of pitch and amplitude associated with a natural voice source can be seen.

The PSDs of the resulting simulated vowels are illustrated in Figures 5.19–5.24. The PSDs have been smoothed with a 10-point moving average filter, using the MATLAB function `smooth`, to remove the fine harmonic detail that

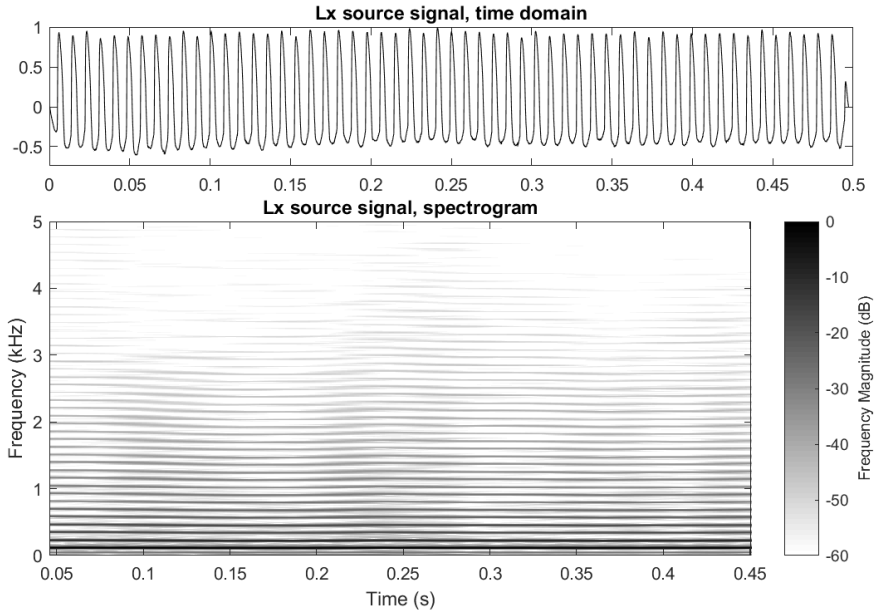


Figure 5.18: Time-domain plot (top) and spectrogram (bottom) of an example Lx source used as input to simulations.

obscures the spectral shape. The accuracy of formant locations has been discussed above, so this section will focus upon broader spectral characteristics. In general, all of the simulated vowels exhibit a roll-off characteristic associated with the Lx source. However, in all cases the natural speech has significantly more attenuation at high frequencies than any of the simulated vowels. This is believed to be at least partly due to the low-pass filter characteristic associated with viscous and thermal losses in the air and at the vocal tract boundaries, which are known to affect the speech spectrum [41], but are not accounted for in any of the simulations.

The PSDs show that the 2DD-DWM simulations result in a series of spectral peaks that continue to occur with relatively even spacing up to 10 kHz. By contrast, the 3DS-DWM and 3DD-DWM simulations have more complicated spectral shapes, due to the influence and interaction of vocal tract side branches as discussed above, and also potentially due to cross-tract modes occurring in the non-uniform cross-sections [30]. This more complicated spectral shape is close to that of the natural speech recording. In many cases,

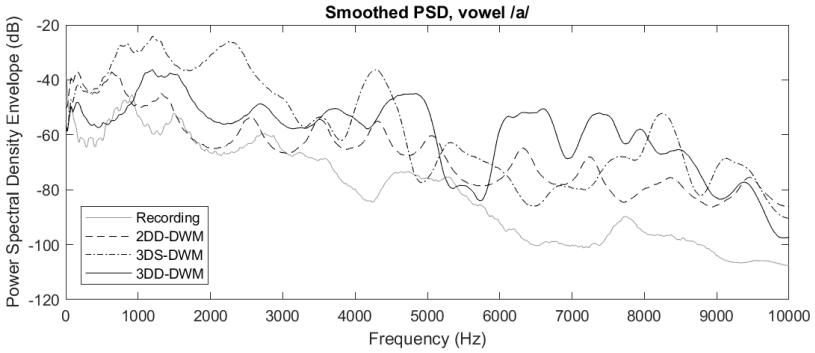


Figure 5.19: PSD of recorded and simulated vowel /a/, smoothed with a 10-point moving-average filter.

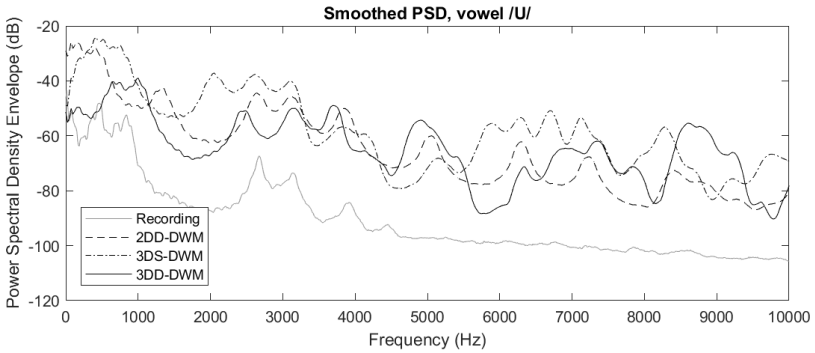


Figure 5.20: PSD of recorded and simulated vowel /ʊ/, smoothed with a 10-point moving-average filter.

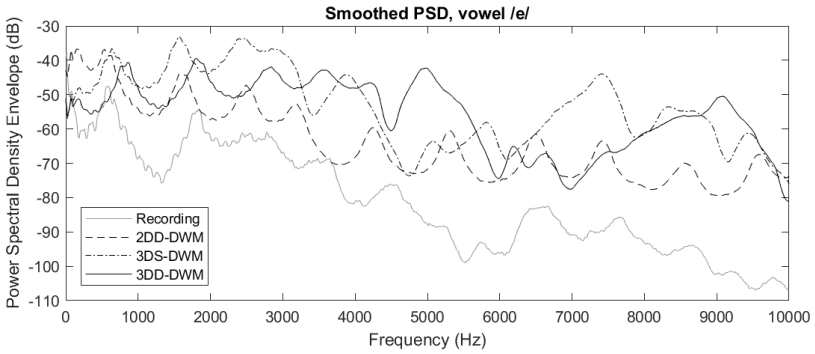


Figure 5.21: PSD of recorded and simulated vowel /e/, smoothed with a 10-point moving-average filter.

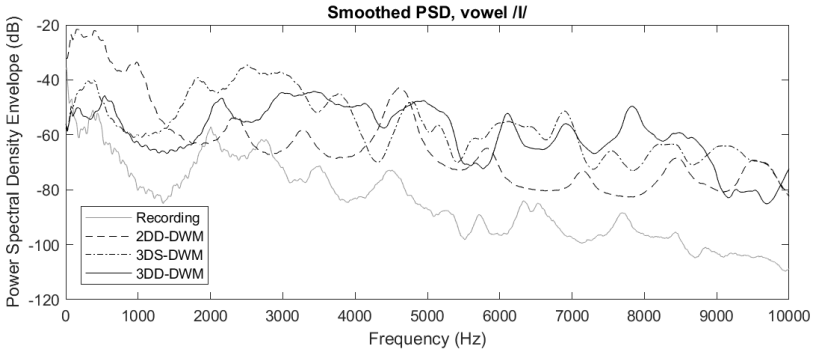


Figure 5.22: PSD of recorded and simulated vowel /ɪ/, smoothed with a 10-point moving-average filter.

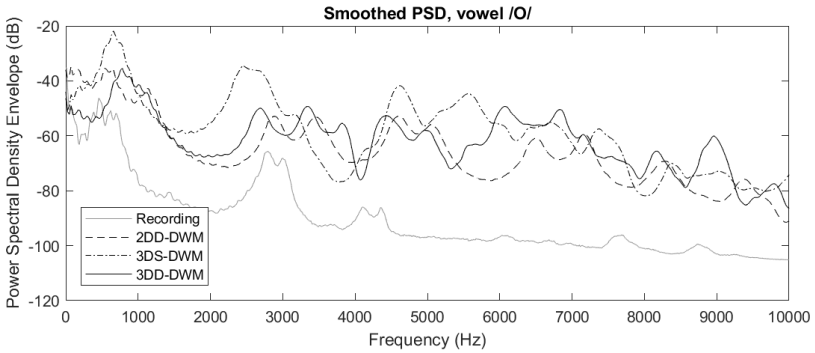


Figure 5.23: PSD of recorded and simulated vowel /ɔ/, smoothed with a 10-point moving-average filter.

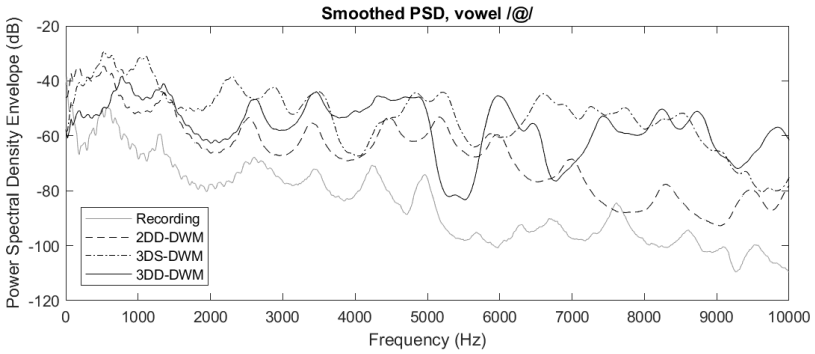


Figure 5.24: PSD of recorded and simulated vowel /ə/, smoothed with a 10-point moving-average filter.

although the specific shapes differ, the general trend of the 3DS-DWM and 3DD-DWM PSD curves are similar, as might be expected for two models based on identical geometries. The differences in individual shape between the two models, which appear to be due to the walls in the 3DD-DWM having a physical depth as discussed above, are larger than expected and suggest that the influence of the yielding walls is greater than anticipated. None of the models appear to approximate the recording particularly well at frequencies above 5 kHz, so any improvement of the 3DD-DWM model over the 2DD-DWM and 3DS-DWM models appears to be mostly in terms of formant accuracy.

Audio Examples

Audition of the synthesised vowels (presented with this work and indexed in Appendix A) permits further insight into the simulation accuracy. In general, the 2DD-DWM simulations sound intelligible but buzzy, with a metallic ringing sound that may be caused by the evenly-spaced spectral peaks at higher frequencies noted above. As might be predicted from the large errors in F2, the simulated /ɪ/ sounds more like /ʊ/, and /ɔ/ sounds more like /ʌ/. The 3DS-DWM and 3DD-DWM simulations present a definite improvement over 2DD-DWM, although neither might be considered as sounding natural. Each of the 3D simulations has a different character, but both present intelligible vowel sounds. As expected from the PSDs, there is more high frequency energy audible in the 3D simulations, but it does not have the same ringing character as the 2DD-DWM simulation.

It is important to note that the results presented in this section may be specific to the MRI subject used, and further studies must consider additional participants, with a range of ages and sexes, before general conclusions can be drawn.

5.7 Diphthong Synthesis

Dynamic models have a further advantage in that they are capable of moving between vocal tract shapes and hence simulating dynamic speech. In this section, results of a comparison between the dynamic 2D model, dynamic 3D model, and recorded diphthongs are presented.

5.7.1 Procedure

The procedure for the diphthong simulations is similar to that of the monophthongs in Section 5.6.1, but the 3DS-DWM model is excluded from the comparison as it is not capable of producing dynamic sounds. In both the 2DD-DWM and 3DD-DWM simulations, admittance maps are generated representing the start and end points of the diphthongs—for example, /a/ and /ɪ/ for the diphthong /aɪ/—and the admittance map used in the simulation is interpolated between the two over the duration of the simulation, following a half-wave sinusoid trajectory (the shape of a sine wave from $-\pi/2$ to $\pi/2$). This trajectory has been found to be more suitable for general diphthong synthesis than a linear interpolation between admittance maps, but it is acknowledged that using the same trajectory for every transition may affect the perceived naturalness of the synthetic diphthongs.

As the simulations in this comparison are dynamic, a source signal is injected directly into the simulation domain at the source position detailed in Section 5.5.2. As above, an Lx signal recorded simultaneously with the benchmark audio is used for this purpose.

5.7.2 Results and Discussion

The spectrograms of the synthesised and recorded diphthongs are presented in Figures 5.25–5.32. As is common for speech research, a pre-emphasis FIR filter, with coefficients [1 -0.97], has been applied to the recorded and synthesized speech data to more clearly illustrate the high frequency components.

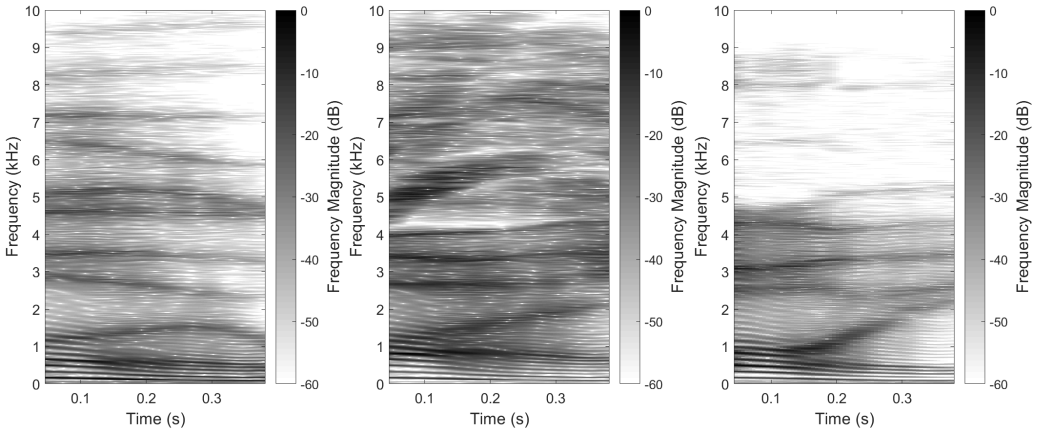


Figure 5.25: Spectrograms for diphthong /ɔɪ/: (a) 2DD-DWM simulation, (b) 3DD-DWM simulation, and (c) recording.

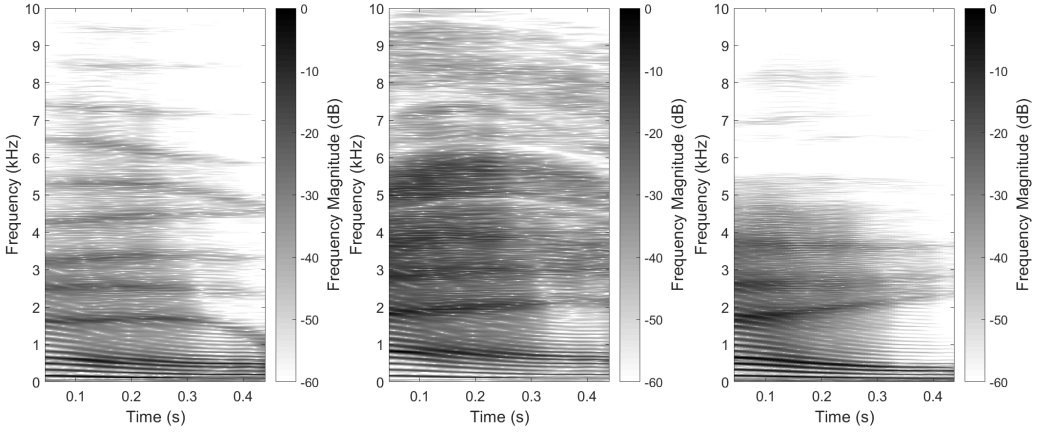


Figure 5.26: Spectrograms for diphthong /eɪ/: (a) 2DD-DWM simulation, (b) 3DD-DWM simulation, and (c) recording.

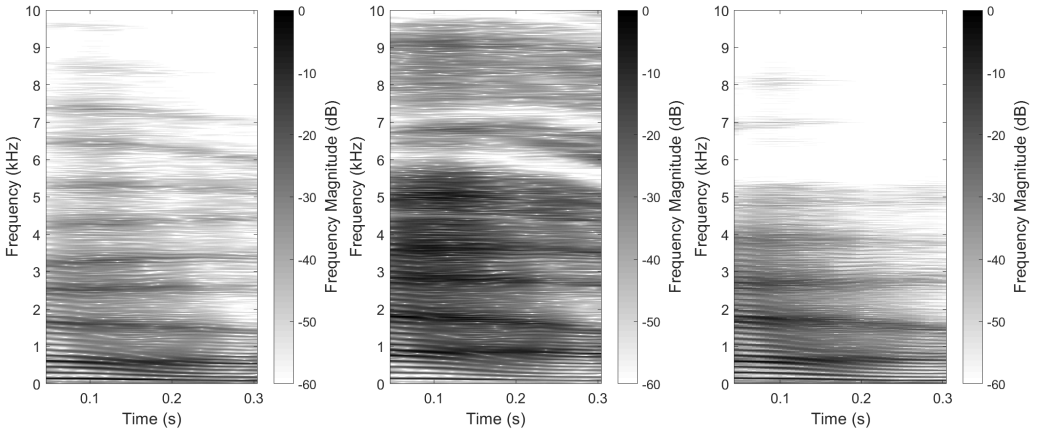


Figure 5.27: Spectrograms for diphthong /eə/: (a) 2DD-DWM simulation, (b) 3DD-DWM simulation, and (c) recording.

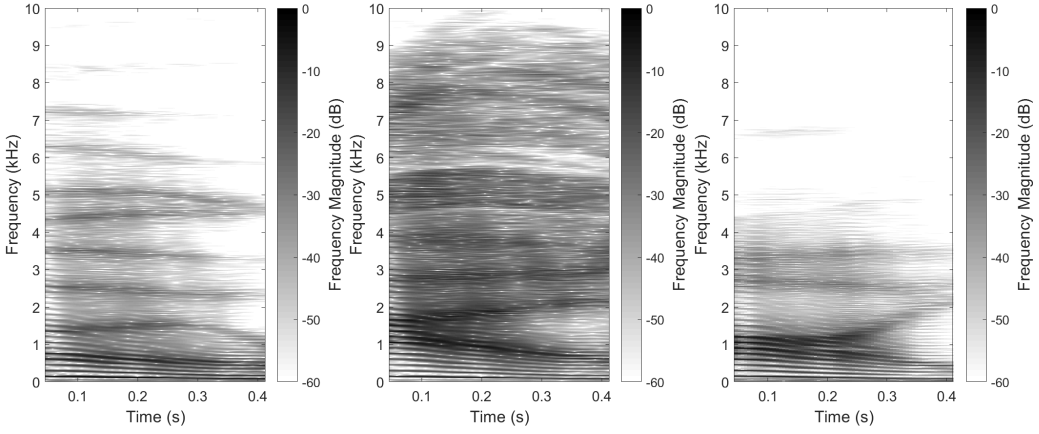


Figure 5.28: Spectrograms for diphthong /aɪ/: (a) 2DD-DWM simulation, (b) 3DD-DWM simulation, and (c) recording.

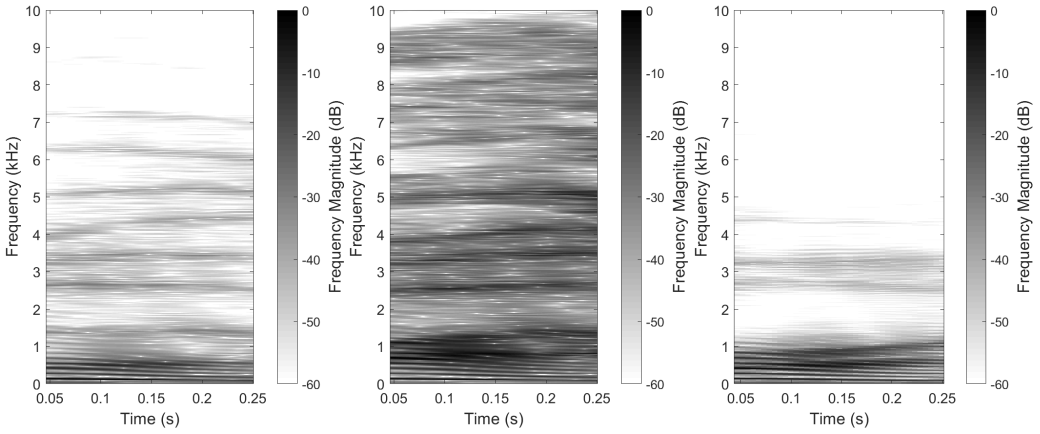


Figure 5.29: Spectrograms for diphthong /ʊə/: (a) 2DD-DWM simulation, (b) 3DD-DWM simulation, and (c) recording.

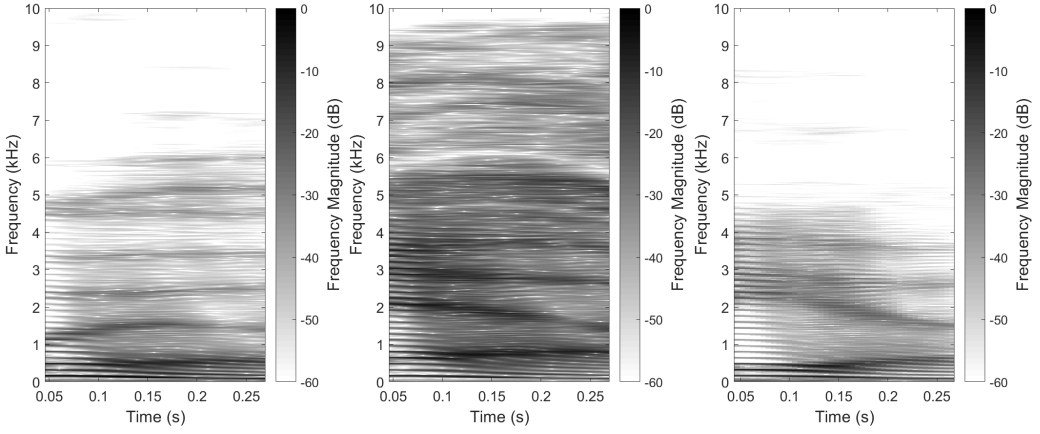


Figure 5.30: Spectrograms for diphthong /ɪə/: (a) 2DD-DWM simulation, (b) 3DD-DWM simulation, and (c) recording.

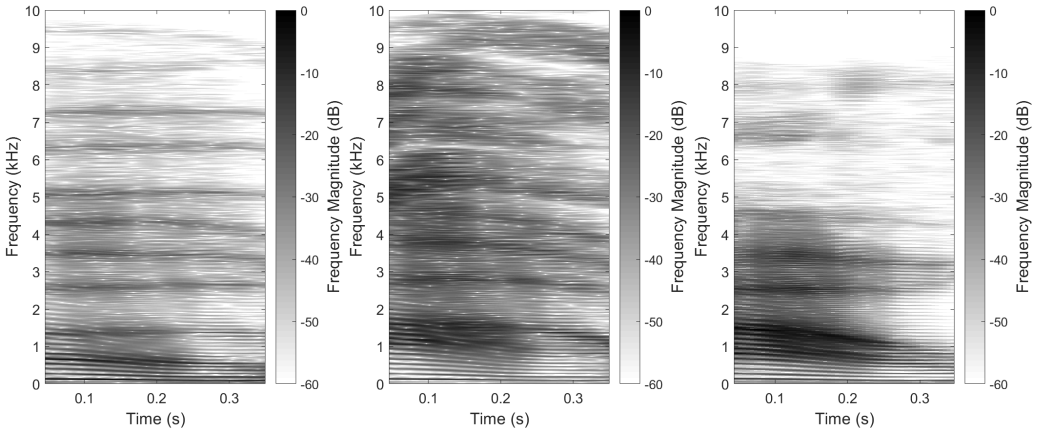


Figure 5.31: Spectrograms for diphthong /aʊ/: (a) 2DD-DWM simulation, (b) 3DD-DWM simulation, and (c) recording.

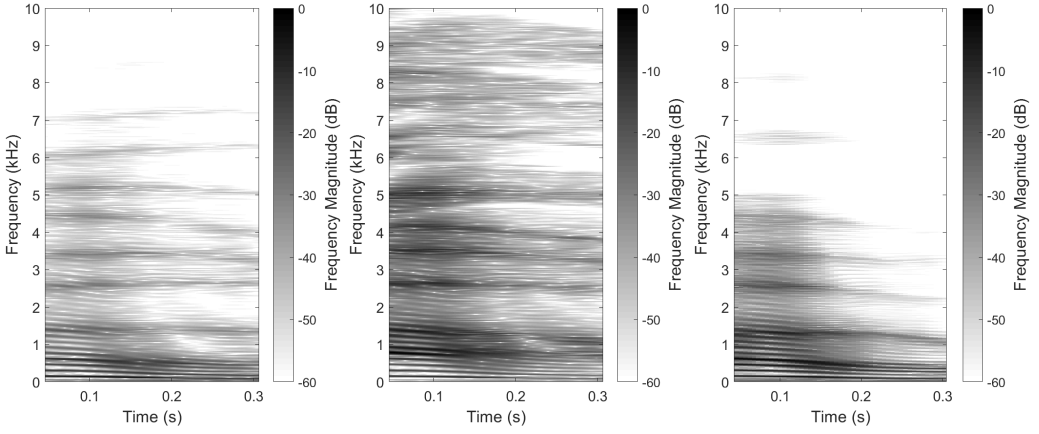


Figure 5.32: Spectrograms for diphthong /əʊ/: (a) 2DD-DWM simulation, (b) 3DD-DWM simulation, and (c) recording.

It can be seen that although the 2DD-DWM model reproduces the lower formants with relative accuracy—apart from /ɪ/ which was noted in the previous section to have an artificially low F2 value—above approximately 4 kHz additional spurious, closely-spaced formants are introduced due to simplification of the vocal tract geometry in the 2DD-DWM simulation. This is consistent with the findings of the previous section. The 3DD-DWM simulation more closely reproduces the number and frequencies of higher formants in the recorded data, due largely to the improved geometrical accuracy of detailed 3D simulations such as the inclusion of side branches.

The results in Figures 5.25–5.32 illustrate one of the main limitations of the proposed system in its current form. Both figures show how the output of the dynamic 3D simulation contains notably more high frequency energy than recorded speech data. This is consistent with the findings from [166] and the previous section. The model currently contains no mechanism to reproduce the frequency-dependent damping within the vocal tract, which causes the reduction in high frequency energy visible in the spectrograms of recorded speech. This occurs due to viscous and thermal losses and other absorption phenomena, and may be approximated by the addition of a filter to the model. It should be noted, however, that similar high frequency energy is present in diphthongs simulated using the FEM method [11], indicating that the 3DD-DWM method produces comparable results; this will be explored in more detail later in the chapter.

Another issue that must be addressed in the dynamic 3D model is the matter of phoneme-specific transitions. As F2 in Figure 5.25(a) illustrates, simulated formant transitions may not follow the same half-wavelength sinusoid shape used in the simulation. Furthermore, every articulation of a given diphthong will be different, even when uttered by the same speaker. It is also clear from the formant traces in Figure 5.25(c) that articulations may be held fixed before and after a transition occurs. Clearly, a simple model such as a sinusoidal interpolation is insufficient for controlling the vocal tract model. Much work has been done on the subject of phoneme transitions in the context of transmission-line articulatory models (see, for example, [129] and

references therein). The process of translating the control parameters of a highly simplified 1D vocal tract model into parameters suitable for control of a detailed 3D geometry presents a significant engineering challenge, but one which is essential to the generation of a suitable control system for the proposed model.

Comparison with Dynamic 3D FEM Vocal Tract Model

Recently, a 3D FEM model capable of dynamic vowel simulations has been introduced [11]. An audio example of the diphthong /aɪ/ produced using this method has been made available online [168], and appears to have been generated according to the 3D FEM method outlined in [11]. This provides a basis for comparison with the proposed model.

The 3D FEM example was generated using a Rosenberg-type input signal [37] rather than an Lx recording, so the first step was to reverse engineer an equivalent source signal for use in the DWM models. Spectral analysis on the FEM sample allowed the pitch curve to be determined, and jitter and shimmer were also added to match the method described in [45], which appears to be the same source signal used in [11] for the example diphthong. The time- and frequency-domain plots of the resulting Rosenberg input signal are presented in Figure 5.33. Note the considerable lack of energy in the 3–5 kHz range compared with the Lx signal in Figure 5.18.

Spectrograms of the resulting 2DD-DWM and 3DD-DWM simulations are presented in Figure 5.34, in addition to that of the comparison 3D FEM simulation obtained from [168]. In the absence of a recorded speech signal for comparison, factors discussed earlier in this chapter provide some indication of simulation quality. For example, the 2DD-DWM simulation again shows multiple, regularly-spaced formants occurring above 5 kHz, which are not seen in the spectrograms of natural speech shown earlier in the chapter. These higher formants occur despite the reduced high-frequency energy of the Rosenberg source compared to the Lx signals.

The 3D FEM simulation appears to contain even more higher frequency

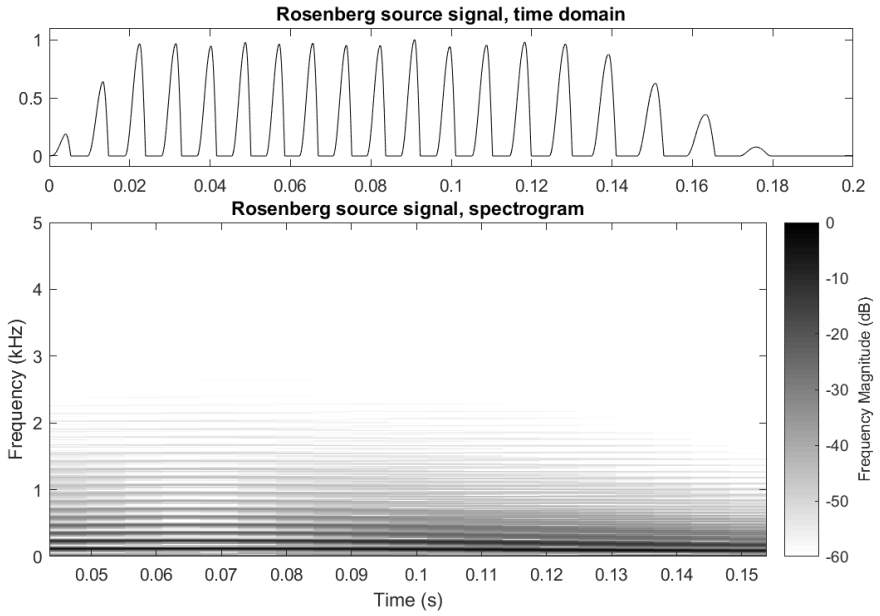


Figure 5.33: Time-domain plot (top) and spectrogram (bottom) of Rosenberg source signal used as input to DWM simulations for comparison with FEM simulations.

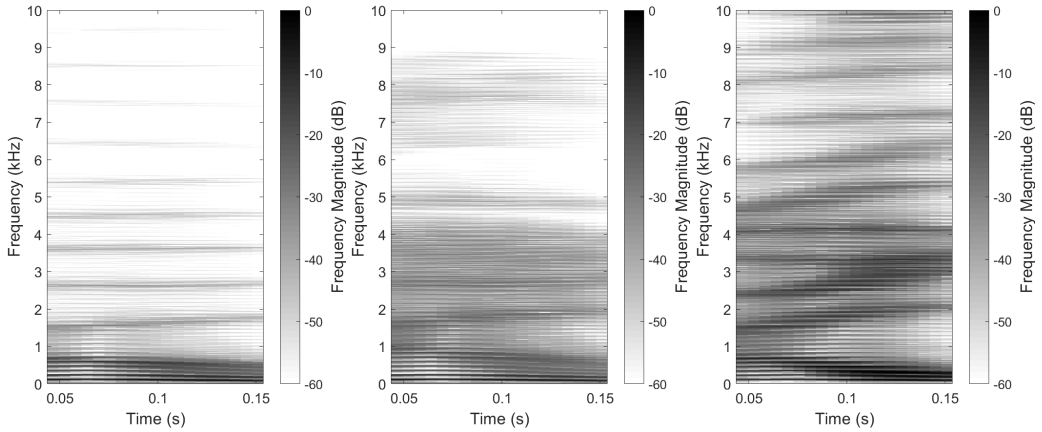


Figure 5.34: Spectrograms for vowel combination /ɑi/: (a) 2DD-DWM simulation, (b) 3DD-DWM simulation, and (c) 3DD-FEM simulation.

energy and clearly defined formants above 5 kHz than the 2DD-DWM simulation. Without an associated recording it is difficult to determine the accuracy of the simulation, but it is anticipated that, like in the 2DD-DWM model, this additional high frequency energy will have a negative impact on the perceived naturalness of the simulation.

The 3DD-DWM simulation shows a distribution of energy throughout the spectrum that is more similar to the natural speech examples in Figures 5.25–5.32 than either the 2DD-DWM or the 3D FEM simulations. However, formant bandwidths in the 3DD-DWM simulation appear to be larger than those of the comparison simulations, which may affect the perceived naturalness results.

The formant values and trajectories in the DWM models appear to be similar to those of the FEM model, which is interesting as MRI data for a different subject was used in the creation of the FEM model. This may explain some of the discrepancies between the two 3D modelling approaches. It is clear, however, that without a natural speech recording to compare against, the only sure way to assess the simulation quality would be a perceptual test. Such tests are described in the next chapter. Future work is planned to perform 3D FEM and 3DD-DWM simulations under identical conditions, using the same MRI geometry and comparison audio, to better quantify the accuracy of both models.

Audio Examples

Audio data is also provided for the dynamic simulations presented in this section. As in Section 5.6.2, the errors in F2 in the 2DD-DWM simulation lead to issues with vowel identification, which impacts several of the simulated diphthongs. Comparing the 3DD-DWM and 2DD-DWM simulations, the 3DD-DWM simulations sound significantly more natural than the 2DD-DWM simulations. As before, this may be attributed to the considerably improved geometry of the 3D model over the 2D model. However, as noted in the spectrograms, the additional high frequency energy in the spectrum of the 3DD-DWM simulations mean that the results are still not comparable to

the recorded speech in terms of naturalness. Comparison with the 3D FEM simulation also reveals that similar levels of unnatural high frequency energy are present in the FEM simulation.

Although the various limitations described throughout this paper imply that the 3DD-DWM model does not yet reproduce completely natural voice sounds, the results presented in this section indicate a significant increase in accuracy, in terms of the reproduction of higher formants that are crucial for naturalness, compared to previously available DWM models.

5.8 Implementation

The complexity of the proposed algorithm is presented in Table 5.4. The primary computational expense is the large number of divisions required. The static 3D model explored in Section 5.6 requires significantly fewer operations, as the assumption of a homogeneous mesh eliminates divisions from step 2 and removes step 6 entirely, resulting in an overall requirement of $9KLM + 1$ additions and $7KLM$ multiplications per time step for a mesh of size $K \times L \times M$. In addition, the static model is a stationary system, and as such its impulse response is sufficient to describe its behaviour: once this has been calculated, it may be convolved with any source. The dynamic model, however, cannot be completely defined in this way, and as such the simulation must be performed over the duration of a given input, resulting in much higher run times overall. A serial MATLAB implementation of the system requires processing times on the order of 13 hours to generate a second of output sound, which is still significantly faster than the 70–80 hours required to generate 20 ms of output using the FEM method in [143]. This speed can be further improved using parallel architectures and/or faster programming languages: for example, a parallel MATLAB implementation using an NVIDIA 1070 GPU (graphics processing unit) reduces the processing time to approximately 3 hours per second of output. While the 2D dynamic model of [147] is capable of running in real time, this is only possible due to the highly simplified geometry and low sampling rate used in the study. In the

Table 5.4: Algorithm complexity of dynamic 3D DWM simulation in terms of addition, multiplication and division operations

For mesh size $K \times L \times M$, for time $n = 1, 2, \ldots$			
	$+$	$\times$	$\div$
1. add input values to p_{input}	1	–	–
2. calculate junction pressure p_J	$10KLM$	$7KLM$	KLM
3. calculate node outputs $p^-_{J,J_{nei}}$	$6KLM$	–	–
4. update $p^+_{J,J_{nei}}$ values	–	–	–
5. extract output sample	–	–	–
6. update admittance maps	$6KLM$	–	–
Total	$22KLM + 1$	$7KLM$	KLM

case of the dynamic 3D model, the complexity is necessary in order to obtain both dynamic movement and improved naturalness in the output.

5.9 Conclusion

This chapter has presented a novel 3D heterogeneous dynamic DWM vocal tract model, in addition to several comparison simulation methods based on existing DWM approaches to vocal tract modelling. The proposed model consistently produces first formant values that are 30–40% too high, which appears to be due to the implementation of yielding walls. However, for the higher formants F2–F5, the proposed model outperforms the 2DD-DWM and 3DS-DWM in terms of formant frequency accuracy. The improvement of the proposed model over a 2DD-DWM is due to the improved detail in the vocal tract geometry that can be captured by a 3D model. Somewhat less expected was the improvement in higher formant frequencies using the proposed model compared to a 3DS-DWM vocal tract model. This improvement also appears to be an effect of the implementation of the vocal tract walls in the 3DD-DWM model, where they are simply a lower-admittance part of the domain rather than an explicit domain boundary. In all cases, the simulations were found to have significantly more energy in the high frequencies than recorded

speech, which is believed to be due to a lack of suitable frequency-dependent loss mechanisms in the vocal tract models.

A consequence of the implementation of the proposed model is that it is capable of dynamic simulations, and it appears to show improvements in diphthong simulation compared to the previously available 2D DWM method. Direct comparison with a state-of-the-art FEM model of the vocal tract also indicates that the proposed model produces comparable output, with significantly reduced computational cost.

This chapter has used some of the objective methods available to determine the accuracy of vowel simulations. However, this can only go some of the way towards describing the *naturalness* of the results. In order to truly determine the naturalness of the simulated vowels, it is necessary to perform perceptual tests using human listeners. The next chapter describes the design and implementation of such perceptual tests.

Chapter 6

Perceptual Testing

The previous chapter provided objective assessments of physical vocal tract modelling techniques. However, as described in Chapter 3, similarity to a particular recording offers no description of synthesis quality in general. Furthermore, the hypothesis that informs this thesis requires an assessment of naturalness, which is an ill-defined concept without known relationships to objective parameters. In order to address these concerns, perceptual tests are required upon the speech samples synthesised as an output from this research.

A perceptual test makes use of a panel of listeners and asks them to rate audio samples in some way. The first part of this chapter describes different perceptual test methodologies and design choices suitable for testing the present hypothesis. Two pilot tests are then described: one comparing the naturalness of a highly simplified 3D DWM vocal tract model against lower-dimensionality DWM approaches, and one comparing sampling frequencies in the 3DD-DWM model to inform a trade-off between model accuracy and computational expense. Finally, a concluding perceptual test is described which tests the hypothesis by comparing the 3DD-DWM model to the 2DD-DWM model, comparable natural voice recordings and—where possible—the 3DD-FEM model [11], forming a perceptual counterpart to the objective results in the previous chapter.

No assumption is made about the distribution of listening test scores in this chapter—such as whether they follow a normal distribution—as non-parametric statistical measures are used throughout. In all following sections, p-values are calculated using the Kolmogorov-Smirnov test [182], and effect sizes are calculated using the A-measure [183]. More details about non-parametric statistics can be found in [182].

6.1 Subjective Testing Methods

There are a number of different testing methods available for the rating of audio samples. This section reviews some of the methods most relevant in the assessment of naturalness. In discussing each of these methods, it is important to bear in mind that naturalness is a relative concept. Each of the tests described is assumed to include synthesised samples from the system under study *as well as* those from alternative systems in order to determine their *relative* naturalness rankings.

6.1.1 Test Format

Paired Comparison

A paired comparison test presents a subject with two samples—in this case, audio clips—and asks them to choose one on the basis of a certain criterion; for example, which sounds most natural, or which sounds most similar to a reference. This test is sometimes called an 'AB' test, referring comparison samples 'A' and 'B', or when a reference is included, an 'ABX' test. Occasionally a third answer option is included for 'don't know', or alternatively a *forced choice* method may be used, in which participants must choose one of the options (if they don't know they must guess, which theoretically produces results at the chance level). Paired comparison tests are very simple and do not require complicated rating tasks, making them suitable for participants with any amount of experience. Listener reliability can be assessed by presenting the same comparison in reverse (e.g. 'AB' and then 'BA')

within the random order of questions in the test and post-screening those whose answers do not agree. The results of a paired comparison test take the form of binomial count data—the number of times sample 'A' was chosen compared to sample 'B'—providing a basis for statistical analysis. As paired comparison tests are simple, they are quick to run, but do not provide any estimate of perceptual distance between the two samples.

Mean Opinion Scores

Perhaps one of the best-known test methodologies for assessing quality—not just for audio—uses a scaling system called the mean opinion score (MOS). Listeners are presented with audio samples, and then asked questions and given five-point rating scales on which to submit their answers. This method was standardised for speech research in ITU-T recommendation P.85 [184], and an example question from this standard is "How do you rate the quality of the sound of what you have just heard?", with the available answers being *excellent*, *good*, *fair*, *poor*, and *bad*. Depending on the test formulation, the quantities being measured may include overall impression, listening effort, comprehension problems, articulation, pronunciation, speaking rate, voice pleasantness, and acceptance. As speech synthesis has improved and naturalness has become more important, improvements have been suggested and the modified MOS scale proposed in [103] has been widely used, which adds the quantities naturalness, ease of listening and audio flow, as well as several modifications to the answer scales. While the entire P.85 questionnaire may be unsuitable for comparisons of short vowel sounds such as those presently under study, the use of individual MOS rating scales may be suitable.

MUSHRA

While [103] defends the five-point rating scale used by ITU-T P.85 on the basis that relatively few, clearly labelled options give the test subject a clear idea of what each question is asking for, many studies have found that this does not give sufficient resolution (which may lead to poor inter-listener agree-

ment [185]), and if one is interested in estimating actual perceptual distance between two samples rather than obtaining a simple ranking, clearly such a discrete scale is insufficient. For these reasons, another standard is also regularly used in perceptual tests: ITU-R recommendation BS.1534-3 [186], also known as the MUSHRA (MUlti Stimulus test with Hidden Reference and Anchor) test. In a MUSHRA test, listeners are presented with several audio samples to compare against a reference, and asked to rate them on a sliding scale. One of the samples presented for comparison will be the reference recording, providing a baseline and a means of post-screening unreliable test subjects. There will also be two hidden anchors, which are low-pass-filtered versions of the reference signal. These anchors are appropriate when testing the quality of an audio system, but when rating the naturalness of synthetic speech, the filtered versions of the reference signal retain much of the naturalness of the original. Therefore, the hidden anchors are commonly omitted when using MUSHRA to assess speech naturalness (e.g. in [187]) to prevent confounding effects on the results.

Effects of Reference and Scale Resolution

In [185], a number of listening test configurations were compared to determine the causes of inter-listener disagreement in a speech quality assessment task. The underlying assumption appears to be that the more closely listeners agree in their answers, the better the task is understood, and therefore the more reliable the test results. The results indicated that the most important factors influencing agreement between listeners were scale resolution (with a continuous scale being preferred to a discrete scale) and the use of a reference stimulus (rather than an inconsistent "internal reference" in each subject's mind). On the other hand, [188] noted that "the inclusion of an extreme version such as natural speech may affect not only the mean ratings of the other versions but also the differences between them". The study found that the distance between ratings of synthetic speech systems were much smaller when a reference signal of natural speech was included. This leads to poorer resolution in the area of interest: the comparison between synthetic

speech systems. It also seems likely that when a reference recording is used as a comparison point, listeners will rate samples based on similarity to that recording, rather than the quantity under study (such as naturalness). Therefore, the decision of whether to include a reference must be carefully weighed against the desired test outcomes.

Alternative Test Methods

Alternative approaches to the paired comparison, MOS and MUSHRA methods for measuring naturalness have been suggested. Some of these tests are very simple and therefore may be more suitable for naïve test subjects, including yes/no decisions ("was this sound made by a human or a computer?" [109]) or choices ("which sounds more natural, sample A or sample B?"). The subjects' answers—and occasionally other data such as reaction times—are recorded. These results may then be processed using a suitable statistical technique to obtain numerical results.

A number of listening tests have begun to use reaction time as a measure of cognitive load [112, 113]. This approach assumes that an increase in cognitive load, indicated by an increased reaction time, is a reliable indicator of speech naturalness, since the human brain has evolved to process natural human speech quickly and therefore unnatural speech will require greater processing times. Experiments measuring reaction times must be delivered in a highly controlled environment, using a system with known latency. This makes reaction time experiments unusable over the internet, due to unknown transmission times and the possibility that participant distraction may confound the results.

6.1.2 Test Delivery

The delivery of the test is an important consideration and, as noted above, may affect the test design. Two main options exist: to perform the experiment in a laboratory setting under controlled listening conditions, or to present the experiment over the internet to allow subjects to take the test

from anywhere. The former option has the benefit of greater experimental control, improving the validity of the results; however, the latter option is likely to receive more participants, improving the statistical significance, and studies have shown that the two situations can give comparable results with sufficient participants [189].

The problem of statistical significance is very important, and [103] highlights some of the problems with studies that use too few participants to generalise the results into solid conclusions. Therefore, the use of internet-based experimenting is very attractive as it allows the test to be accessed by many more participants than one requiring attendance at a laboratory. However, there are a number of important concerns if internet-based experiments are to be used. In 2001, [190] presented 16 standards for internet-based experiments which are still relevant today, including the collection of demographic information at the start of the survey, techniques to reduce participant dropout, and means of obtaining a wider participant population.

In addition to the above concerns about internet-based experiments, for audio-based perceptual tests, the listening conditions will vary between subjects and may be poor, for example due to poor-quality headphones and/or noisy environments [191]. This must be considered when analysing the results, although it is generally assumed that the anticipated greater number of participants will help to average out any particularly poor results. Problems that may occur due to the playback of audio over the internet—such as skipping or buffering—can be somewhat mitigated by using short samples, allowing participants to listen to the samples as many times as required, and asking participants to tick a box on questions where they experienced audio difficulties, so that those responses may be removed from the data.

6.1.3 Test Materials

The samples presented to listeners during a listening test must be carefully selected in order to prevent any confounding variables from affecting the results. For example, simulations based on different physical vocal tract models

should use the same source signal and comparable model parameters, so that differences between samples are only due to the model type. Furthermore, if the samples are to be compared to recorded natural speech, a source signal as close to the natural source as possible should be used. This is achieved using the Lx source signal as described in Section 5.6.1.

In order to ensure that the playback method does not affect the results, the audio signals should also have envelopes applied to prevent clipping due to truncation at the start and end of the signal - a linear amplitude ramp of 5 ms at either end of the sample is sufficient. Depending on the test delivery method and the original sample rate, it may be necessary to downsample the audio files to standard audio rate (44.1 kHz) for playback over the internet.

6.1.4 Summary

In this section, available approaches and important design decisions for perceptual tests have been discussed. Since, in the present study, relative naturalness of the proposed model is an essential measurement, the MUSHRA methodology is considered a good fit to provide not just rank orders of naturalness but an estimate of the size of the difference. While the strengths of internet-based testing are acknowledged, the studies in this section take place under controlled listening conditions to avoid confounding factors. However, the tests are designed using the online software Qualtrics [192], making the future implementation of internet-based tests simpler. As described in the previous chapter, Lx recordings are used as the source signal for all simulations that are compared to recordings, and the simulation parameters are made as similar as possible, to ensure a fair comparison between samples that addresses the research question. The next sections discuss the implementation of each perceptual test.

6.2 Pilot Test: Dimensionality Increase

The initial pilot study was performed to determine the perceptual effect of increasing the dimensionality of the simulation method, from two to three dimensions, while retaining the simplifications of lower-dimensionality models such as a straightened vocal tract with a circular cross-section. In addition, the test presented a first opportunity to investigate the use of the MUSHRA test methodology.

6.2.1 Method

This study makes use of a simplified dynamic 3D DWM vocal tract model, as opposed to the 3DD-DWM model introduced in the previous chapter which uses the detailed 3D vocal tract geometry. This simplified model, henceforth referred to as the S3D-DWM model for brevity, is constructed in a similar way to the 2DD-DWM as described in Section 5.2.1. However, in place of a single raised-cosine function at every point along the vocal tract, the S3D-DWM model uses a two-dimensional function based on multiplying two cosine functions to produce a 2D array of values as illustrated in Figure 6.1. As with the 2DD-DWM, the maximum value of impedance, Z_x, at any distance x from the glottis is calculated from the cross sectional area, A_x, at the same location, following (2.9). The resulting 3D admittance map is illustrated in Figure 6.1 and is essentially a direct extension of the 2DD-DWM model into three dimensions. As such, it retains the assumptions that the vocal tract is a straight tube with circular cross-sectional area.

Using raised-cosine functions based on one-dimensional area function data does not permit the production of lateral consonants, but the S3D-DWM model offers plenty of opportunities for this problem to be overcome, for example by using two periods of a cosine wave across the tube at suitable values of x, or by eliminating cosine mapping altogether and using a specific acoustic admittance based technique similar to the 3DD-DWM but without a bend in the tract. Since this pilot study only requires the production of vowels, the issue with lateral articulations is not relevant here, and the

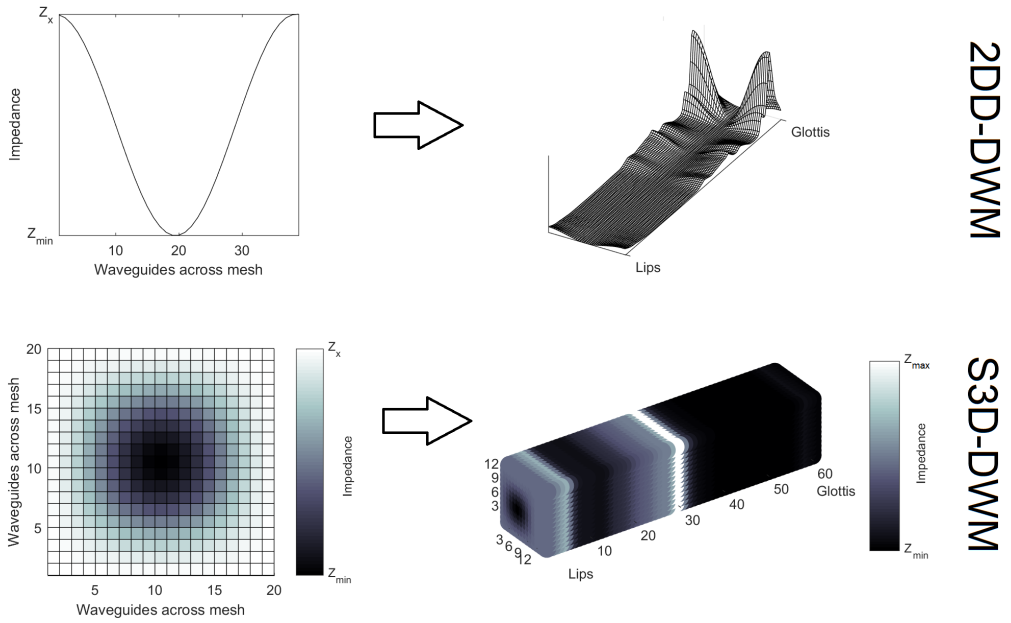


Figure 6.1: Construction of the 2DD-DWM model (top row) and simplified 3D dynamic DWM (S3D-DWM) model (bottom row) of the vocal tract.

cosine-based approach is used.

As the S3D-DWM retains the majority of simplifications of the 2DD-DWM model, there is not expected to be a large improvement in naturalness scores for the former over the latter. However, some improvement in naturalness may be achieved simply by increasing the dimensionality [149], as the difficult problem of mapping an area to a two-dimensional surface is avoided. Furthermore, in both simulation methods the source signal is input across all scattering junctions at the glottis end of the mesh; while neither option replicates the true vocal tract behaviour, the 3D behaviour is expected to be slightly more natural as the source is imposed over an *area* rather than a line.

The perceptual test was performed using a MUSHRA-type methodology, but without hidden anchors as previously discussed. Participants were provided with a description of what the diphthong was supposed to sound like; for example, "each of the clips below represent the vowel sound in the word

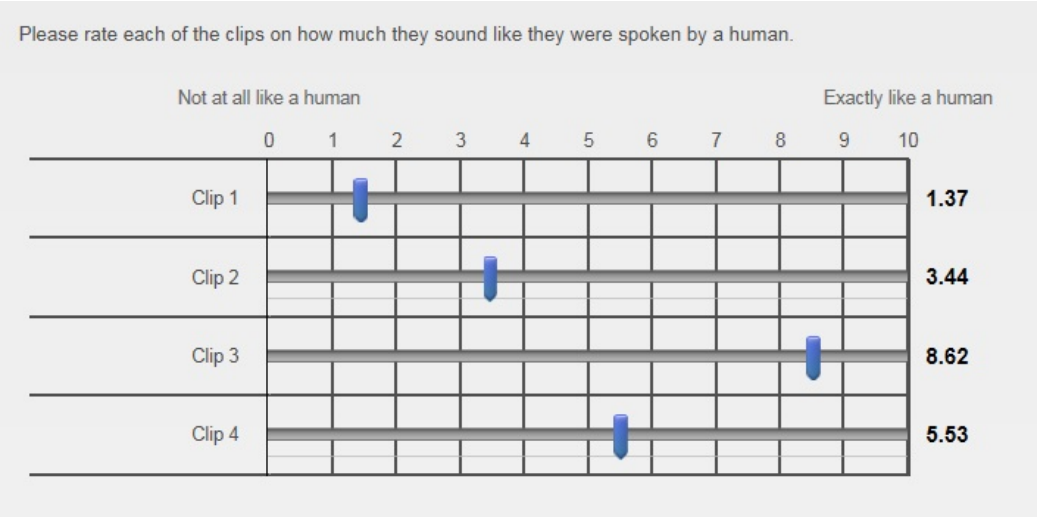


Figure 6.2: Example of slider-type naturalness questions used in pilot perceptual test. Four audio clips of a single diphthong, including a 1D, 2D and 3D simulation and a recording, are presented at the top of the page in a random order, and the sliders are used to assign a naturalness score to each.

'bay'". They were then presented with 4 clips: diphthongs synthesised using the 1D Kelly-Lochbaum and 2DD-DWM and S3D-DWM methods, and a recording of corresponding natural speech. The original sampling frequencies were 192 kHz for the synthesised examples, and 96 kHz for the recordings, and all audio examples were low-pass filtered with a cut-off frequency of 10 kHz and downsampled to 48 kHz. All examples had a bit depth of 16 bits and each was around 0.5 s long.

Participants were first asked to rate the similarity of the examples to the vowel described, and then asked to rate the naturalness of each sample in a separate question on the same page. The separation of similarity and naturalness ratings was done to encourage participants to think about naturalness as a separate quality, which they could rate even if the intended synthesis targets were not achieved. Both questions used a sliding scale from 0-10, with the ends of the naturalness scale marked "not at all like a human" (0) and "exactly like a human" (10). An example of the question format can be seen in Figure 6.2. Participants were also given the option to comment on each diphthong comparison. The listening test took place under controlled listen-

	Median Normalised Naturalness Rating					
	1D K-L	2DD-DWM	S3D-DWM	Recording	h-value	p-value
eɪ	0.22	0.49	0.46	1.00	0	0.962
aɪ	0.15	0.29	0.38	1.00	0	0.085
ɔɪ	0.20	0.45	0.40	1.00	0	0.825
ɪə	0.12	0.43	0.42	1.00	0	0.422
eə	0.27	0.55	0.54	1.00	0	0.422
ʊə	0.09	0.14	0.26	1.00	1	0.000
əʊ	0.10	0.30	0.46	1.00	0	0.085
aʊ	0.13	0.21	0.31	1.00	0	0.422
Combined	0.14	0.35	0.39	1.00	1	0.005

Table 6.1: Median normalised naturalness scores for the eight English diphthongs with 1D Kelly-Lochbaum (K-L), 2DD-DWM and S3D-DWM simulations and recorded natural speech. A significant difference between the 2DD-DWM and S3D-DWM is indicated by a h-value of 1, and the associated significance level (p-value) is given. The bottom row contains the results obtained when all diphthong comparisons are combined to produce a single score.

ing conditions (quiet room, Beyerdynamic DT990 Pro headphones set to a suitable level by the investigator) using the online survey tool Qualtrics [192]. There were 34 test participants, with ages ranging from 20 to 60, of which 22 were male, 11 were female and 1 other. Participants were asked whether they had any experience with phonetics or synthetic speech; 15 reported that they had experience, and 19 reported that they did not.

6.2.2 Results and Discussion

Prior to analysis, naturalness scores were normalised so that the maximum score a participant had given at any point in the test became 1, and the minimum score given throughout the whole test became 0. In this way, it is possible to equate the scores given by different participants, without losing information about the relative naturalness of different diphthongs and synthesis techniques. The medians of the resulting normalised scores are presented in Table 6.1, and a box plot representing the scores when all diphthongs are combined is given in Figure 6.3.

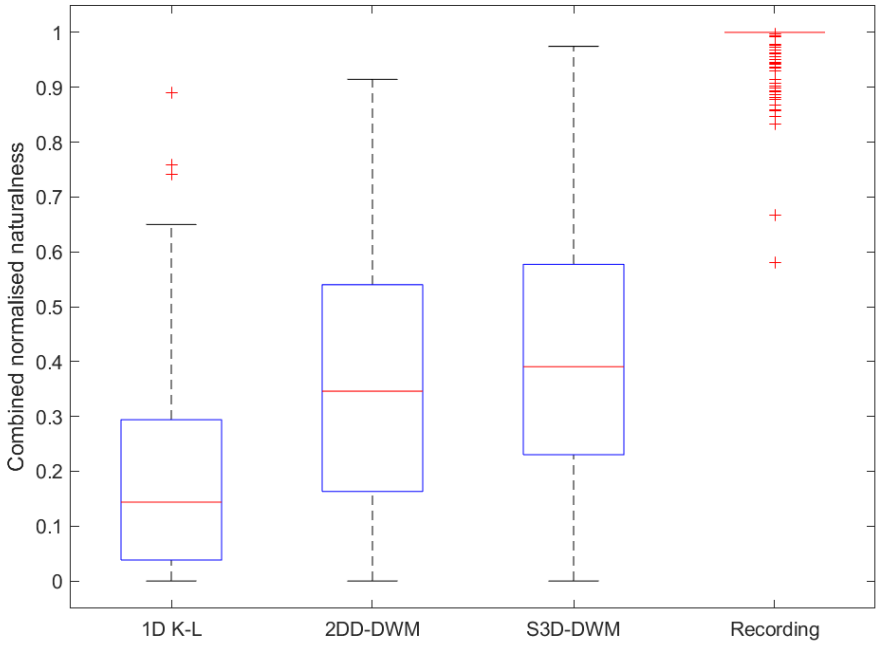


Figure 6.3: Combined normalised naturalness scores for 1D K-L, 2DD-DWM and S3D-DWM simulations and recordings across the eight diphthongs studied.

Considering at first the combined scores, given in the bottom row of Table 6.1 and with distributions illustrated by Figure 6.3, it can be seen that the S3D-DWM model was rated as significantly more natural than the 2DD-DWM model, with $p < 0.005$, and the effect size of this difference calculated using the A measure [183] is 0.563 which is considered a small, but non-trivial, effect. This supports the initial hypothesis that increasing the model dimensionality would offer a small increase in naturalness. The 3DS-DWM model was also rated as more natural than the 1D Kelly-Lochbaum vocal tract model with $p < 0.001$ and $A = 0.783$, which is a large effect size.

Table 6.1 illustrates the differences in score that were obtained across all diphthongs. It is apparent that naturalness ratings differ significantly across diphthongs. In /aɪ/, /ʊə/, /əʊ/ and /aʊ/ the S3D-DWM was rated as more natural than the 2DD-DWM, while in the four other cases the 2DD-DWM was considered most natural. However, none of these differences were significant except in the case of /ʊə/. Furthermore, there were differences between diphthongs in the naturalness scores given, with /eə/ obtaining the highest naturalness scores for 1D, 2D and 3D simulations, while /ʊə/ obtained the lowest scores in all three cases. This may suggest that certain diphthongs are harder to synthesise realistically than others, although /ʊə/ was a special case that will be discussed in more detail below.

Finally, it should be noted that the naturalness scores given to each of the synthesised diphthongs remain well below those assigned to natural speech. Given the simplifications of the models and the spurious high-frequency energy known to occur in the synthesised diphthongs (see Section 5.7.2), this is to be expected. Comments made by the listeners during the test can shed more light upon the factors involved in making decisions about naturalness.

Listener Comments

The listener comments associated with this test can be found in Appendix B. The comments largely fall into two categories: comments on the high-frequency artifacts in the synthesised audio examples, and comments on the difference between the audio examples and the example word given.

The high-frequency artifacts in the simulations are described variously by participants as “metallic”, “buzz”, “ringing”, “breathiness” and occasionally as an “aliasing-style effect”. As discussed in the previous chapter, the vowels simulated using the DWM are known to include more high-frequency energy than recorded speech, as frequency-dependent losses are not implemented, which causes the high-frequency artifacts or ‘buzzing’ that listeners found so disturbing. The comments about aliasing are not as troublesome as they first appear: during the simulated diphthongs in question, higher formants which are not audible in recorded speech are seen to increase in frequency during the transition even as the pitch drops. This causes an audible effect similar to aliasing. The audio examples were low-pass filtered with a cut-off of 10 kHz before downsampling from the simulation frequency of 192 kHz to 44.1 kHz (for playback over the online interface), to prevent any actual aliasing from occurring.

Listener comments also note that the synthesised diphthongs—and even sometimes the recordings—do not always match with the expected word given in the question text. This is most obvious in the comments on /ʊə/, highlighting an issue with the use of the example word ‘boar’, which is more likely to be pronounced /bɔə/ than the /bʊə/ intended by the question. The choice of ‘boar’ as an example word was motivated by the desire to give the same context for every diphthong to avoid confounding the results with different contexts, and the context /b-[diphthong]/ was chosen as English words were available in most cases; ‘boar’ was selected as the closest available pronunciation for /ʊə/. With hindsight, enforcing such context restrictions actually may have affected the results more than not doing so, and indeed the naturalness scores assigned for /ʊə/ are the lowest for any of the diphthongs, as illustrated in Table 6.1.

The 1D Kelly-Lochbaum simulation is reliably referred to in the comments as being unnatural, and occasionally perceived as having phasing or flanging effects to it. However, there are certain circumstances where it is referred to as sounding natural, and notably for the diphthong /aɪ/ one participant refers to the 1D simulation as “the least natural” while another describes it

as "most human". It is clear from these comments that the perception of naturalness varies greatly from person to person, and suggests that in future tests as many subjects as possible should be recruited in order to reduce the effect of these differences.

6.2.3 Summary

The aim of this test was to determine whether a simple dimensionality increase from a 2D vocal tract simulation to a 3D version—while retaining all the other simplifications of a 2D model—resulted in a significant increase in the perceived naturalness of the output. Analysis of the results of this listening test indicates that listeners do find the simple 3D simulation significantly more natural than lower-dimensionality approaches, although the effect size is small when compared to a 2D DWM approach. This is sufficient to justify perceptual testing and investigation of models making use of the detailed 3D vocal tract geometry which are anticipated to significantly improve upon simplified models. The MUSHRA methodology was found to be a suitable technique for assessing naturalness, however the inclusion of a description of the intended sound appeared to confound the results somewhat, with some participants appearing to base their scores on what they thought the samples were 'meant' to sound like rather than their naturalness. For this reason, future MUSHRA-based tests will exclude the description, and simply present the samples for comparison with a direction to rate them in terms of naturalness and no other context information.

6.3 Pilot Test: Sampling Frequency

Following the development of the 3DD-DWM as described in Chapter 5, one important issue that remains is the sampling frequency, f_s required by the model. The 3DS-DWM model proposed in [166] uses $f_s = 960$ kHz, and hence a spatial resolution of 0.63 mm following (5.5). A 3D DWM model at such a high sampling frequency—especially the proposed 3DD-

DWM model which also models the head tissue—is very computationally expensive. This pilot test therefore compares the output of a 3DD-DWM model with spatial resolution of 0.63 mm, corresponding to $f_s = 960$ kHz, with one that has a spatial resolution of 1.58 mm, corresponding to $f_s = 384$ kHz. This lower sampling frequency was found to be the lowest value for which the meshes of the vowels under study retain an equivalent physical depth at a constriction, as described in Section 5.5.1. While the model with a higher sampling frequency is expected to produce output that is objectively more similar to recorded audio samples, this test aims to determine whether these differences are perceptually relevant, or whether the two simulation methods are judged to be similar enough that the much larger computational expense of the 960 kHz simulation can be avoided.

6.3.1 Method

The aim of this pilot test was to determine whether there is a significant difference in the perceived similarity to natural speech between simulated speech created using a 3DD-DWM model with the lowest viable sampling frequency for the vocal tract shapes produced during vowels /a/,/e/, /ɪ/, /ɔ/, /ʊ/, and /ə/ (384 kHz), compared to a model using a much higher sampling frequency previously found to provide accurate simulations (960 kHz) [166]. It was expected that if a perceptual difference was apparent between the two modelling approaches, it would be more evident for monophthongs than diphthongs, because any errors in the simulation are likely to be more noticeable when a vowel is held static for a longer time.

Unlike the previous study, this test required direct pairwise comparison between samples. Therefore, a paired comparison methodology was selected. A two-alternative forced-choice system was used, and listeners were asked to judge which of the two available options sounded most similar to an example, which was always a recording of the diphthong or monophthong being spoken by a human. Every possible pair of options—384 kHz and 960 kHz simulations, 384 kHz simulation and recording, and 960 kHz simulation and recording—was presented for each of the 14 test cases (6 monophthongs and

8 diphthongs). The audio was presented as a 16-bit wav file, downsampled to 48 kHz and with linear amplitude ramps applied to the first and last 5 ms to prevent any audible clicks. Each sample was around 0.5 s in duration. The order in which questions were presented, as well as the order of options within questions, was randomised to minimize the effect of presentation order on the results. The listening test was presented in the form of an online survey using Qualtrics [192], however for the present study all participants took part under controlled listening conditions as previously described. A total of 22 participants, aged between 21 and 61, took part in the pilot study, of which 14 were male, 7 female, and 1 other. 21 of the 22 participants reported that they had experience with critical listening and/or synthetic speech.

6.3.2 Results

The binomial count data resulting from the listening test described in the previous section can be seen in Table 6.2. The top row of data in the table is of most interest, as it contains the results of direct comparisons between the simulations. The results across all 14 test cases are presented first, and it can be seen that the 384 kHz simulation was chosen as sounding most similar to the recording in 181 out of 294, or 61.6%, of comparisons, while the 960 kHz simulation was chosen in 113 out of 294, that is 38.4%, of comparisons. As the data is dichotomous, and the number of samples exceeds the limit for calculation of the exact binomial distribution, the z-score for this difference is calculated based on the normal approximation to the binomial distribution following [193]. The calculated z-score is 3.79, indicating that the difference is significant with $p < 0.0001$; that is, the 384 kHz simulation is rated as significantly more similar to the recording than the 960 kHz simulation. Furthermore, the effect size of this difference using the A-measure [183] is 0.80, which is considered to be a large effect. The difference between simulation sampling frequencies is even more distinct when only the dynamic simulations—the 8 diphthong test cases—are considered. In this case the difference is also significant at $p < 0.0001$ with a large effect size. Finally, the difference between simulation methods for the 6 monophthong test cases

	All Phonemes		Diphthongs		Monophthongs	
Paired comparison	d1	d2	d1	d2	d1	d2
384 kHz – 960 kHz	181 (61.6%)	113 (38.4%)	116 (69.0%)	52 (31.1%)	65 (51.6%)	61 (48.4%)
384 kHz – Recording	0 (0%)	294 (100%)	0 (0%)	168 (100%)	0 (0%)	126 (100%)
960 kHz – Recording	0 (0%)	294 (100%)	0 (0%)	168 (100%)	0 (0%)	126 (100%)

Table 6.2: Paired comparison results: raw frequency counts and counts as a percentage of total responses. Frequency counts d1 and d2 indicate how many times participants chose the first or the second method, respectively, from the pairs described in the first column, as most similar to a recorded reference signal.

is not significant, indicating that listeners cannot reliably choose one model type as more similar to the recording than another for the monophthong examples tested.

Given the differences in model selection between the static and dynamic simulation cases, a chi-squared test for independence was also used to determine if there is any significant program dependence [193]. The results indicate that there is a significant difference in how participants responded to the static test cases compared to the dynamic test cases: they were more likely to select the 384 kHz simulation as sounding most similar to the recording when the stimulus was a diphthong, with $p < 0.01$. Similar tests within groups indicated no significant difference in rating behaviour between diphthongs, but significant differences between monophthongs with $p < 0.0005$; that is, the particular monophthong under study had a significant effect on which model output was selected as most similar to the recording. As an example, the 384 kHz simulation was chosen as most similar to the recording for the vowel /ɔ/ by 20 out of 21 test participants, while the 960 kHz simulation was preferred to varying degrees of significance for /a/, /ʊ/, /e/ and /ə/.

The final two rows of Table 6.2 show the results when the simulations were compared with recorded speech. The scores indicate that the listeners are reliable in identifying the recording as most similar to itself in every case. The lack of ambiguity here does illustrate that the simulations cannot yet be considered natural, as they were never mistaken for the real human speech.

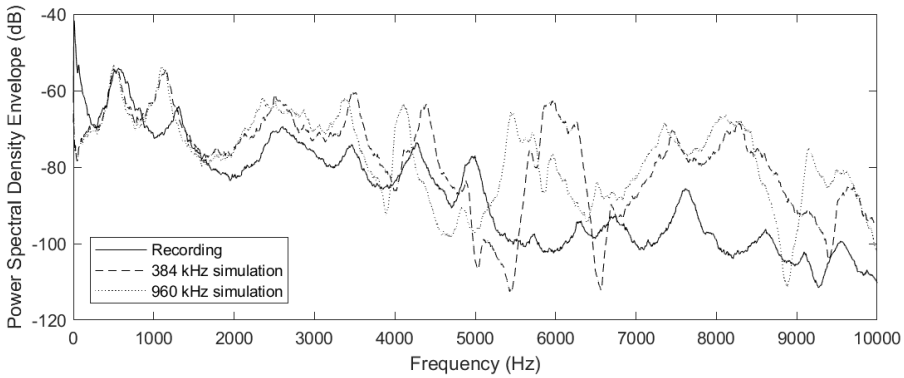


Figure 6.4: Power spectral density (PSD) plots for vowel /ə/ in recording, 384 kHz simulation and 960 kHz simulation. PSDs have been smoothed using a 40-point moving-average filter.

6.3.3 Discussion

The results presented in the previous section show some interesting trends, particularly in light of the original hypothesis that if any differences could be identified, the 960 kHz simulations would be considered more similar to the recordings than 384 kHz simulations. This hypothesis was based on the observation that the 960 kHz simulations more accurately reproduced the behaviour of natural speech, particularly with regards to formant frequencies, than the 384 kHz simulation. This can be seen in Figure 6.4, for the vowel /ə/, which was the monophthong for which the 960 kHz simulation was rated most natural. While both simulations exhibit significant deviation from the spectrum of the recording, the 960 kHz simulation more closely approximates the formant frequencies of the recording, particularly in the 3–6 kHz region.

The results appear to show that the simulations with a lower sampling frequency actually receive a higher rating from listeners when the stimulus is dynamic, and that listeners did not prefer one model type over the other when the stimulus is static. However, this is a small pilot study and care must be taken not to over-interpret the results, as $N = 21$ is too small a sample size to draw completely robust conclusions. Consider Figure 6.5, for example, which shows the spectrograms for the two synthesized versions of the diphthong /eɪ/, for which 76.2% of listeners rated the 384 kHz simulation

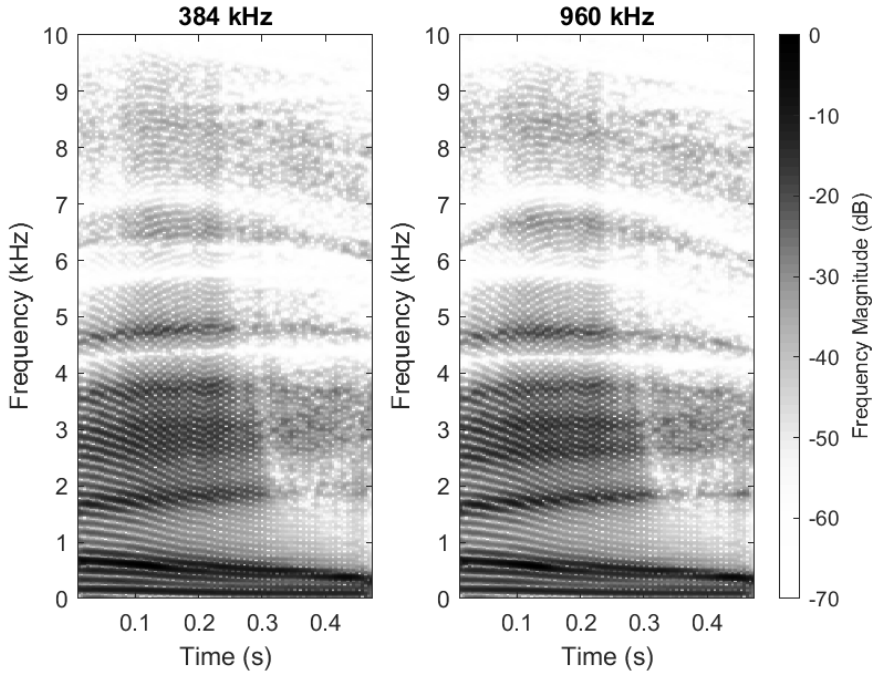


Figure 6.5: Spectrograms of diphthong /eɪ/ synthesized using 384 kHz simulation (left) and 960 kHz simulation (right).

as most similar to the recording. Looking at the spectrograms, it is difficult to identify the source of such a perceptual difference between the two model types. However, the data certainly suggest that the perceptual difference between models with $f_s = 384$ kHz and $f_s = 960$ kHz is not as large as might be expected given the large difference in sampling frequencies.

The purpose of this test was to determine whether simulations run at 960 kHz, i.e. 2.5 times the lowest viable sampling frequency of 384 kHz, were worth the large increase in computational expense, given that it requires roughly $2.5^4 \approx 39$ times as many calculations per second. The examples synthesised for this test, each ∼0.5 s, required ∼20 min computation time at 384 kHz and ∼9 h computation time at 960 kHz when synthesized in MATLAB using built-in GPU array functionality on a 1152-core GPU. On the basis of the results, it can be said that $f_s = 384$ kHz is sufficient for the synthesis of diphthongs using the 3DD-DWM vocal tract synthesis technique, particularly when the output consists of dynamic elements of speech, saving signif-

icant computational expense. Given that this model is intended for future use in generating running speech, which is of course dynamic, this result is encouraging.

It is important to note that none of the synthesised examples presented to test subjects were mistaken for the recording of natural speech. This result is not surprising given the significant deviations from the desired spectrum evident in Figure 6.4. As seen in the previous chapter, the synthesised samples contain more high-frequency energy than the recordings, helping to identify them. Listener comments from this pilot test (found in Appendix B) lend further support to this assumption, with remarks suggesting that the two simulated samples were perceptually very similar to one another, but had "bad high frequency artifacts" which identified them as unnatural, and were still "very different" from the recorded diphthongs.

6.3.4 Summary

The results of this study presented in this section indicate that for a 3DD-DWM vocal tract model, increasing the sampling frequency does not necessarily result in a corresponding increase in perceptual similarity of the synthetic output to recorded human speech. These data hint at a potential perceptual limit on sampling frequency in such models, and suggest that using a sampling frequency as high as 960 kHz for dynamic speech synthesis is not justified given the additional computational cost it entails. This has important implications for algorithm efficiency and the eventual generation of natural sounding synthetic speech in real time.

6.4 Perceived Naturalness of 3DD-DWM Vocal Tract Model

The final set of perceptual tests aim to provide an assessment of the relative naturalness of the proposed 3DD-DWM model compared to the 2DD-DWM

model and recorded speech. As explored in Chapter 5, one example of a diphthong generated using a dynamic 3D FEM model [11] is also available online [168], so this study also provides an initial assessment of the relative naturalness of the 3DD-DWM technique compared to the FEM model.

6.4.1 Method

The aim of this study is to determine whether speech synthesized using the proposed 3DD-DWM technique is perceived as more natural than the established 2DD-DWM technique [147], and in addition, whether the naturalness of the 3DD-DWM model is comparable to that of a dynamic 3D FEM model [11], referred to as the 3DD-FEM model for consistency. Based on objective comparison of Section 5.7.2, and audition of the synthesised samples (audio examples of which are available in the accompanying files, see Appendix A), it is hypothesised that the output from the 3DD-DWM method will be considered more natural than that of the 2DD-DWM technique, and comparable to the output from 3DD-FEM synthesis.

A perceptual test was designed, based on the MUSHRA methodology used in Section 6.2. Participants were asked to rate the naturalness of each audio file on a sliding scale from 0 to 100. The zero end of the scale was annotated "not at all like a human" and the 100 end described as "exactly like a human". Eight comparisons were made—one for each English diphthong—between recorded speech and 2DD-DWM and 3DD-DWM synthesis, using Lx recordings as the source as described in Chapter 5. References and example words were not used, except the recording which serves as a hidden reference among the options, to prevent listeners from using similarity to the recording as their criterion instead of naturalness. The sliding scales provide both a rank ordering of naturalness and an estimate of perceptual distance between the samples.

As in Chapter 5, a further comparison was made between 2DD-DWM and 3DD-DWM synthesis and 3D dynamic FEM synthesis [11] for the diphthong /ɑi/, based on example output from the FEM system available online [168].

In this case, a Rosenberg-type glottal pulse signal [37] of 200 ms duration was used as input to the DWM model, with an amplitude envelope, pitch curve, jitter and shimmer applied to match the source signal used in [168, 45], following the process described in Section 5.7.2. A set of four comparisons were made for this diphthong, with low-pass filters applied at 16 kHz, 12 kHz, 8 kHz and 5 kHz respectively, to determine the effect of high-frequency spectral components on the perceived naturalness. As both DWM and FEM vocal tract models are known to be affected by spurious high-frequency energy, comparing low-pass filtered versions of the simulations makes it possible to infer some information about the relative quality of the synthesis once this energy is removed.

The audio was presented as 16-bit wav files, downsampled to 44.1 kHz and with linear amplitude ramps applied to the first and last 5 ms to prevent any audible clicks. Samples were approximately 0.5 s long, except those made using the Rosenberg source which were 0.2 s long to match the FEM comparison signal. All samples were low-pass filtered with a cut-off of 16 kHz [168]. The order in which questions were presented, as well as the order of options within questions, was randomized to minimize the effect of presentation order on the results. Participants were given two example questions to accustom them to the test format and audio, and were offered the chance to make comments after every comparison. A total of 34 participants took part in the study, with ages 19–62, of which 25 were male, 7 female and 2 other. 30 participants reported experience with critial listening, 27 with synthetic speech, and 25 were native English speakers. The study was presented using Qualtrics [192] under controlled listening conditions as described in Section 6.2.

6.4.2 Comparison with Recordings

The first stage in processing the results is normalisation, as described in Section 6.2, which accounts for differences in use of the scale by participants while preserving relative distances between scores, and allows direct comparison between different questions and participants. Although analysis is

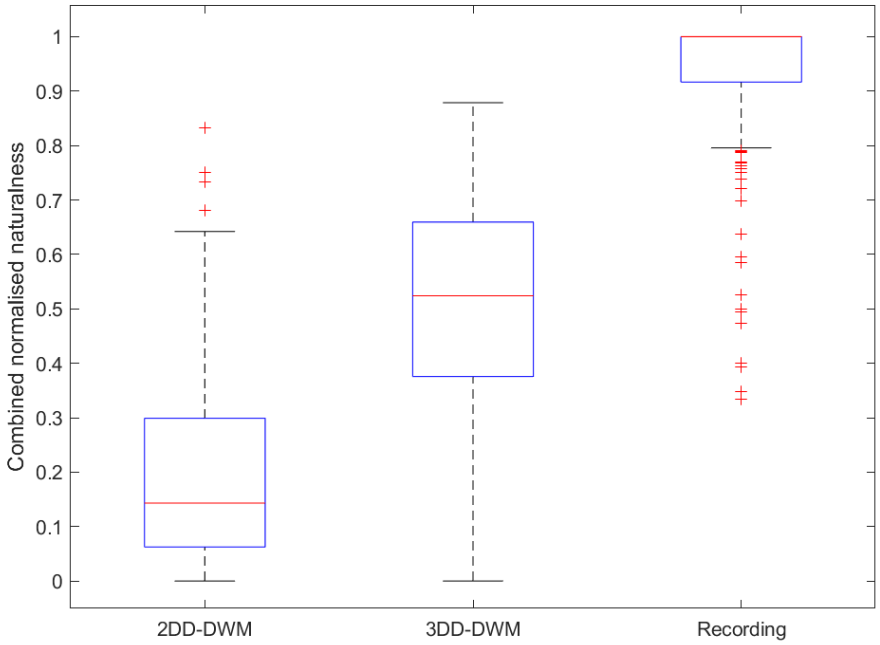


Figure 6.6: Combined normalised naturalness scores for 2DD-DWM and 3DD-DWM simulations and recordings across the eight diphthongs studied.

presented separately, the 2D-3D and DWM-FEM comparisons were presented randomly to participants in a single test block.

Figure 6.6 illustrates the results of perceptual tests comparing the 2DD-DWM, 3DD-DWM, and recordings of natural speech. These are the normalized naturalness values combined across all eight diphthongs. Table 6.3 presents the median normalized naturalness scores separately for each of the diphthongs, as well as a combined score based on all test conditions. The differences between each pair of conditions—2DD-DWM, 3DD-DWM and recording—are statistically significant at $p < 0.005$, with a large effect size, for every diphthong studied.

The results clearly indicate that for every diphthong tested, the 3DD-DWM modelling technique is considered more natural-sounding than the 2DD-DWM method. This is to be expected given the increased geometrical accuracy of the 3DD-DWM model, and supports the results in Section 5.6.2,

	Median Normalised Naturalness Rating			2DD-DWM vs. 3DD-DWM	
	2DD-DWM	3DD-DWM	Recording	h-value	p-value
eɪ	0.16	0.56	0.98	1	<0.000
aɪ	0.15	0.58	1.00	1	<0.000
ɔɪ	0.16	0.58	1.00	1	<0.000
ɪə	0.11	0.45	1.00	1	<0.000
eə	0.24	0.52	0.92	1	0.002
ʊə	0.11	0.33	0.90	1	<0.000
əʊ	0.13	0.55	1.00	1	<0.000
aʊ	0.09	0.60	1.00	1	<0.000
Combined	0.14	0.52	1.00	1	<0.000

Table 6.3: Median normalised naturalness scores for the eight English diphthongs with 2DD-DWM and 3DD-DWM simulations and recorded natural speech. A significant difference between the 2DD-DWM and 3DD-DWM is indicated by a h-value of 1, and the associated significance level (p-value) is given. The bottom row shows the results obtained when all diphthong comparisons are combined to produce a single score.

which found multiple, closely-spaced formants in the 2DD-DWM model output, leading to audible higher resonances and a metallic sound in the resulting simulations. This unnatural behaviour is not observed in the 3DD-DWM model, resulting in higher naturalness ratings.

6.4.3 Comparison with 3D FEM Model

Figure 6.7 illustrates the results of naturalness tests comparing the 2DD-DWM, 3DD-DWM, and 3DD-FEM methods for each of the four cut-off frequency conditions. Table 6.4 provides the median normalised naturalness scores for each of the four filtering conditions. In every case, 2DD-DWM simulations were rated as less natural than 3DD-DWM simulations with $p < 0.05$ and a medium effect size, supporting the findings of the previous section despite the use of a Rosenberg source model rather than the recorded Lx data. The 2DD-DWM and 3DD-FEM simulations were rated differently with $p < 0.01$ and a large effect size. However, in all cases but the 5 kHz filtering condition, 3D DWM and 3D FEM were not rated as significantly different at the 5% significance level (at 5 kHz, this difference is significant

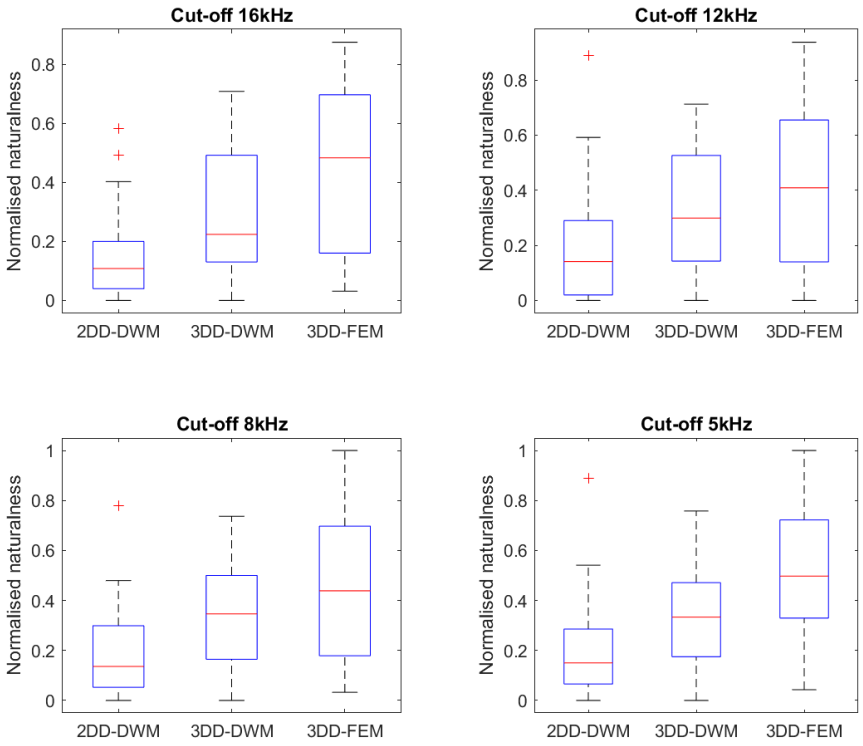


Figure 6.7: Normalised naturalness scores for 2DD-DWM, 3DD-DWM and 3DD-FEM simulations of vowel combination /ɑi/ with different filter cut-offs.

at $p < 0.01$ with a medium effect size).

The results indicate that although the median naturalness scores for the 3DD-FEM simulations were higher than those given to the 3DD-DWM simulations, the difference is not significant where filters with cut-off frequencies 16 kHz, 12 kHz and 8 kHz are used. This indicates that 3DD-DWM and 3DD-FEM simulations are comparable in most normal speech applications. Where the filter cut-off is set to 5 kHz, the 3DD-FEM model was rated as significantly more natural; this is expected, as FEM simulation is known to be more accurate than DWM simulation [30], and the 5 kHz cut-off filter removes the confounding high frequency artifacts that might otherwise affect the naturalness rating.

	Median Normalised Naturalness Rating			2DD-DWM vs. 3DD-DWM		3DD-DWM vs. 3DD-FEM	
Cut-off	2DD-DWM	3DD-DWM	3DD-FEM	h-value	p-value	h-value	p-value
16 kHz	0.11	0.22	0.48	1	0.044	0	0.155
12 kHz	0.14	0.30	0.41	1	0.021	0	0.422
8 kHz	0.14	0.35	0.44	1	0.009	0	0.085
5 kHz	0.15	0.33	0.50	1	0.009	1	0.004

Table 6.4: Median normalised naturalness scores for the diphthong /ɑi/ with 2DD-DWM, 3DD-DWM and 3DD-FEM simulations and different cut-off frequencies. 2DD-DWM and 3DD-DWM simulations, and 3DD-DWM and 3DD-FEM simulations, are compared, with a significant difference indicated by a h-value of 1. The associated significance level (p-value) is also given.

As Figure 6.7 illustrates, there was a large overlap in the scores assigned to each of the two simulation techniques when energy above 5 kHz was included, and in every case, the normalised scores assigned to the FEM technique covered almost the full range from zero to one. This large variability reflects the differences in the assessment of naturalness among test participants, and highlights the importance of perceptual testing for any synthetic speech system, using as many participants as possible to reduce the impact of outliers.

The naturalness ratings of the 3DD-DWM system obtained during the comparison with the 3DD-FEM technique were significantly lower ($p < 0.001$, large effect size) than those obtained in the comparison with recordings in Section 6.4.2. These effects may be due to the use of a Rosenberg-type source in the FEM comparison, as described in Section 5.7.2. Although this artificial source signal does contain a pitch curve and some jitter and shimmer to make it more natural-sounding, it still results in synthesised output that sounds obviously less natural than when using an Lx input signal, and the differences between the two signals are quite clear when Figures 5.18 and 5.33 are compared.

The results presented in this section indicate that in the case of the diphthong /ɑi/, the naturalness ratings of 3DD-DWM and 3DD-FEM vocal tract modelling methods are comparable for most scenarios. Note that Section 6.4.2 showed that naturalness scores varied across diphthongs, so in order to draw more general conclusions, more FEM simulations would be required. Future studies are planned in which 3DD-DWM and 3DD-FEM simulations

will be performed using identical vocal tract segmentations and source signals, allowing a more rigorous comparison to be made.

6.4.4 Listener Comments

During the test, listeners were given the option to comment upon each comparison and to provide overall comments at the end. These comments offer some insight into the reasoning behind some of the scores, and are provided in Appendix B.

As may be anticipated from the previous chapter and from the comments on the perceptual test presented in Section 6.2, spurious high frequency energy featured heavily in the participant comments. One participant referred to a simulation as sounding like "it had quite a bit of white noise in it" while another remarked on "the sound of resonances, almost like going overboard on a finite Q-factor boost at certain frequencies". These two comments point to two different types of high-frequency error: one broadband, and one featuring audible resonances at high frequencies. Based on Figure 5.34, it seems likely that the 3DD-DWM simulations fall into the former category, while the 3DD-FEM simulations fall into the latter.

In the fourth comparison case, where all synthesised samples were low-pass filtered with a 5 kHz cut-off, one participant stated that "these are starting to sound more human", indicating that the high frequencies are indeed a major contributor to the perceived unnaturalness. It is clear that, at present, the absorption of higher-frequency energy by the vocal tract is not well-modelled by either simulation technique.

6.4.5 Demographic Effects

It is useful to consider whether the general pattern of results differs for participants who are not native English speakers, as the synthesis method will eventually be applicable to every language. Furthermore, it is of interest to determine whether the amount of exposure a participant has had to synthetic

speech affects ratings of naturalness, and in what direction. Participants were asked questions about each of these factors at the start of the test, in order to permit the necessary analysis.

When considering the comparison between the 2DD-DWM and 3DD-DWM simulations and recorded speech, there was no significant difference found in the scores given across all diphthongs between participants who indicated that they were native English speakers and those that did not. This suggests that, with a recording to compare against, providing a naturalness rating is a relatively well-defined task, although more participants and different language examples would be required before any firm conclusions can be made.

Native English proficiency did appear to interact with naturalness ratings in the comparison between DWM and FEM synthesis methods, with the ratings given to 3DD-DWM and 3DD-FEM simulations significantly different between native and non-native speakers ($p = 0.010$ and $p = 0.035$ respectively). The 3DD-DWM model was generally rated higher, and the 3DD-FEM model rated lower, by native English speakers than by non-native speakers. There were more native speakers taking part in the test (25 out of 34 participants), so the small number of non-native speakers may be insufficient to draw general conclusions; however, these results suggest that such differences should be taken into account in the assessment of synthetic speech, even for isolated vowel sounds. The vowel combination /ɑi/ does not act as a diphthong in English, which may have affected the results as it may sound less natural to native English ears, but this was the only diphthong available using the 3DD-FEM model. More rigorous comparisons between the two methods are expected to be possible in the future, and this is described in more detail in Section 8.4.

Experience with synthetic speech is another aspect that may affect naturalness ratings. For example, it is well known that exposure to synthetic speech leads to increased acceptance [194]. For this reason, participants were also asked about their experience with synthetic speech. The following analysis groups all those who had even limited experience with synthetic speech to-

gether as experienced users, as there were insufficient participants to divide into smaller sub-groups; however, it may be the case that a greater level of detail is required.

In the opposite pattern to the results considering native English ability above, experience with synthetic speech was found not to affect the comparisons between DWM and FEM simulation methods. However, when comparing DWM synthesis to recordings, experienced participants rated the 2DD-DWM as significantly less natural than inexperienced participants, with $p = 0.007$. Ratings for the 3DD-DWM were not significantly affected by experience. This is an interesting result as it suggests that participants with experience of synthetic speech find the 2DD-DWM even less acceptable than the general population. However, as participants for the study were recruited primarily from audio undergraduate and research courses at the University of York, 27 out of 34 participants described themselves as having experience with synthetic speech, so the size of the two groups is not balanced. Larger tests are required to draw any solid conclusions on the effect of participant experience on naturalness rating.

6.4.6 Summary

Perceptual tests indicate that simulations produced using the 3DD-DWM method sound more natural than those produced using the established 2DD-DWM method. Furthermore, the 3DD-DWM method appears to be comparable in naturalness to established 3DD-FEM techniques requiring much longer simulation times. However, neither the 3DD-DWM nor the 3DD-FEM technique achieve comparable naturalness to recordings of real speech, indicating room for improvement in the models, particularly in the treatment of high frequency energy. This supports the findings of the previous chapter.

6.5 Conclusion

This chapter has presented the results of several perceptual tests designed to determine crucial aspects of the simulation method and measure the naturalness of the resulting simulations. In the first pilot study, a simple dimensionality increase from 2DD-DWM to S3D-DWM was implemented in order to determine the effect of such an action on naturalness, and the results indicated that further perceptual tests were required on models using the detailed 3D vocal tract geometry. The second pilot study considered whether there was a perceptual justification for using very high sampling frequencies in the model, and found that there was no significant difference in naturalness ratings between a simulation run at 384 kHz and one at 960 kHz; in fact, sometimes the former was preferred. This justifies the use of the lower sampling frequency in the proposed 3DD-DWM model, saving significant computational expense.

The final perceptual test considered the naturalness of the proposed model when compared to the established 2DD-DWM model and recordings, and the proposed model was found to be significantly more natural in every case; however, the 3DD-DWM model remains significantly less natural than recorded speech samples. The hypothesis of this thesis is that *a detailed, three-dimensional dynamic digital waveguide mesh model of the vocal tract can produce more natural synthetic diphthongs than a two-dimensional dynamic digital waveguide mesh model*, and the results of the final perceptual study support this hypothesis. However, listener comments throughout the pilot and final tests indicate that high-frequency energy in the simulations is still an important cause of unnaturalness, which supports the objective findings in the previous chapter. Addressing this must be a priority for future work.

A second part of the final perceptual test compared simulations made using the proposed model with those generated using a state-of-the-art FEM vocal tract model, and the results suggest that participants do not consider the two methods to be significantly different in terms of naturalness, except where

high-frequency error in both simulations is removed using a low-pass filter with a 5 kHz cut-off. This is particularly encouraging as the FEM models require much longer simulation run-times (on the order of 1000 hours per second of output) than the 3DD-DWM model (around 3 hours per second of output), suggesting that the proposed model may be more suitable for use in complete speech synthesis applications than the FEM model with no loss of perceived naturalness.

As discussed in Section 3.5, naturalness in synthetic speech applications takes two forms: linguistic and acoustic naturalness. The results in this and the previous chapter have considered the acoustic naturalness improvements offered by the proposed 3DD-DWM vocal tract model. However, in order for the DWM vocal tract model to be used for synthetic speech applications, it must be combined with a linguistic model such as a text-to-speech front end that can produce suitable instructions to control the simulation. The next chapter describes an initial study undertaken to assess the feasibility of combining a DWM vocal tract model with well-established statistical parametric techniques for text-to-speech synthesis in order to achieve this goal.

Chapter 7

Combining DWM and Statistical Approaches

The physical models of the vocal tract described in the previous chapters are controlled entirely by MRI-derived vocal tract geometries and, if dynamic, an arbitrary transition trajectory. Such models are of interest in the study of speech acoustics, but of limited use for speech synthesis applications. As discussed in Chapter 3, a control model linking the acoustic model to some input signal, such as articulator movements or written text, is necessary if physical vocal tract models are to become a viable method of speech synthesis.

This chapter describes a proof-of-concept study for a new technique combining a DWM vocal tract model with deep neural network (DNN) based statistical parametric speech synthesis (SPSS) approaches, permitting the synthesis of speech from written text using a DWM. As described in Chapter 5, detailed 3D vocal tract models incorporate hundreds of thousands of degrees of freedom, making them unsuitable, at present, for use with SPSS approaches, which optimise every parameter individually over multiple iterations. However, the 2D heterogeneous DWM model [147] requires far fewer parameters, and is sufficiently similar in implementation to the proposed 3D heterogeneous DWM model that it can be used for a proof-of-concept study investigating the potential for control of physical vocal tract models with SPSS techniques.

This chapter first presents more detail about the DNN-based TTS synthesisers upon which the combined DNN-DWM model is based. Important differences in the implementation of the 2D DWM in the combined model, compared to the 2DD-DWM explored in previous chapters, are then introduced. The experimental method and results of the pilot study are presented next, followed by an evaluation of the method and priorities for future investigation. The work in this chapter took place in collaboration with members of the Tokuda and Nankaku Laboratory at Nagoya Institute of Technology (NITech) in Japan, funded by the Japan Society for the Promotion of Science Summer Programme Fellowship. The method and results of the pilot study were presented at INTERSPEECH 2017 with the title "Articulatory text-to-speech synthesis using the digital waveguide mesh driven by a deep neural network" [91]. The specific contribution of the author to this work was the reformulation of the DWM to make it suitable for use with DNNs, development of the DWM-specific code to be implemented within NITech's existing DNN architectures, and analysis of the results.

7.1 The DNN-driven TTS Synthesiser

The DNN-driven TTS synthesiser was introduced in Section 3.2.3 as a state-of-the-art technique for SPSS. This section provides more detail on the implementation of DNN-based TTS synthesis and highlights important design aspects.

An artificial neural network (ANN) is a machine learning paradigm based loosely upon the operation of the human brain, comprising of layers of interconnected *neurons* that generate an output signal based on a nonlinear relationship to an input signal, producing complex emergent behaviour. The ANN is used as a 'black box' to model some unknown relationship between the input and output of a system—such as handwritten text and the ASCII character it corresponds to—and using training data and optimisation, the ANN learns the transformation between the two. Once trained in this way, the model can be applied to new data to produce an appropriate output.

Great care is required at the training stage to ensure that the model a) converges upon a minimum-error solution, and b) does not *overfit* to the training data by "learn[ing] the 'data' and not the underlying function" [195].

Each neuron in an ANN is connected to those in other layers, and computes a nonlinear *activation function*, such as a simple threshold operation or a sigmoid function [196], on the weighted sum of its inputs, producing an output value which passes to the next layer. The *weights* applied to each input, and the *biases* of the neurons (where the bias is "a measure of how easy it is to get the [neuron] to fire" [196]), are what determine the behaviour of the ANN [195], and these are the parameters that are optimised during the training stage. The weights and biases may be initialised randomly, or knowledge about the system may be incorporated to speed up training and promote known relationships between input and output.

The architectures of ANNs vary in the number of layers, number of neurons per layer, and how the neurons are interconnected; a *deep* neural network (DNN) is simply one that contains multiple 'hidden' layers in between the input and output layers, as illustrated in Figure 7.1. In most applications relevant to the current study, neuron outputs are only allowed to pass towards the network output rather than forming feedback loops; this is known as a *feedforward* architecture.

The input and output layers are constructed according to the requirements of the system. Following an example given in [196], an ANN designed to recognise handwritten numbers might take 784 input features, each representing a pixel in a 28×28 image segmented from some larger block of writing, and each consisting of a greyscale value between 0 (white) and 1 (black). In this case, the output layer might consist of 10 features, representing the digits 0–9, and the highest value in the output layer would indicate which digit the ANN 'thinks' corresponds to the handwritten number. Alternatively, a *softmax* layer might be used at the output, which ensures that the values at all neurons in the output layer sum to 1, effectively providing a probability distribution [196].

In the case of TTS synthesis, the use of DNNs has led to improved speech

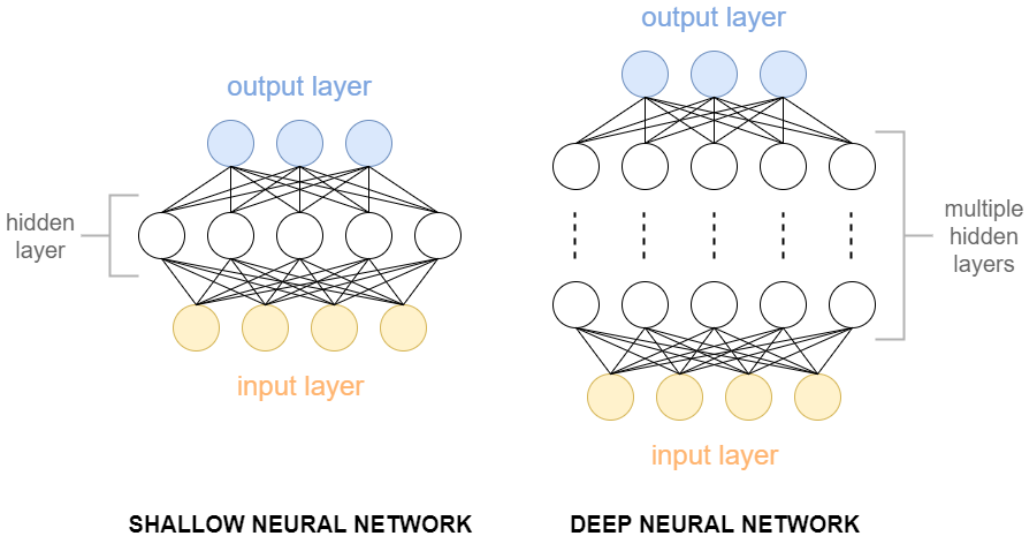


Figure 7.1: Comparison of shallow and deep neural network architectures.

synthesis quality compared to earlier, HMM-based techniques [75]. In this application, the input to the DNN is typically a linguistic feature vector, which is a set of several hundred features describing the current phoneme and its context that is obtained automatically from written text [197]. Example input features for a DNN system are given in [63]:

> "The input features include binary answers to questions about linguistic contexts (e.g. is-current-phoneme-aa?) and numeric values (e.g. the number of words in the phrase, the relative position of the current frame in the current phoneme, and durations of the current phoneme)."

By incorporating this much context information, the DNN-based synthesiser is able to produce varied, natural-sounding speech output. The output vector consists of a set of synthesiser parameters, typically MFCCs and voicing information such as pitch, as well dynamic information about the parameters such as the difference in their values between frames, calculated using methods similar to the finite difference approximation and known as *delta* (first difference) and *delta-delta* (second difference) parameters [63]. These

parameters are then used to drive a vocoder in order to produce synthetic speech.

The sheer size of the input vector (371 features in [63] and 355 in [197]) means that a realistic training dataset could never contain every possible combination of features, so the model must be able to generalise to unseen contexts, which means overfitting must be avoided. This is a common concern for ANNs and a number of approaches exist to address it. One of the best known is *dropout* [198], which randomly removes neurons and their connections from the ANN during different phases of the training, preventing neurons from 'co-adapting' and building redundancy into the network. Regularisation may also be used [197], which promotes lower values for the model weights, making them less sensitive to small changes in the input [196]. Training a DNN for speech synthesis applications requires many iterations and training sentences, and takes several weeks even using a parallel implementation on a GPU.

Although the above describes a general SPSS approach using the DNN, recent models have moved beyond a frame-based approach and have begun to directly model the speech waveform, reducing the error associated with the additional step of resynthesising speech using a vocoder. Google's WaveNet [78] uses a convolutional neural network to model the speech waveform on a sample-by-sample basis, using thousands of previous audio samples at the input layer in addition to linguistic features. If linguistic features are omitted from the training, the network learns to produce eerily speech-like sounds that lack any linguistic content [199]. The brute-force approach of WaveNet leads to highly natural-sounding synthetic speech [199], but provides very little insight into the workings of the vocal system and has an extremely high computational cost.

In [77] and [200], an ANN is used to learn the transform from linguistic feature vector to cepstral parameters as previously described, but instead of driving a vocoder, these are used to calculate an impulse response which is then convolved with a voiced or unvoiced source (a pulse train and noise respectively). The network is then trained using the error between the filter

output and the reference speech signal. In this way the network learns to produce a linear time-invariant (LTI) filter, approximately representing the vocal tract, for each speech frame. This method can be adapted for use with the DWM model, which also acts as an LTI filter, but has the potential to encode spatial information such as vocal tract size and shape in a human-interpretable manner. The combination of these two techniques forms the basis for the DNN-DWM model proposed in this chapter. The next section will return to the DWM and consider some of the changes required if it is to be incorporated into a DNN system.

7.2 Reformulating the 2D DWM

The heterogeneous 2D DWM algorithm was introduced in Section 5.2.1. However, due to the vast number of iterations required in the training of a DNN-based TTS system, any increases in computation speed will have a significant effect on the overall time required to train the model, and may make the difference between a viable and an unviable system. This section describes a number of changes to the previously-introduced method in order to reduce its computational expense.

7.2.1 The K-DWM

The DWM algorithm as previously presented is known as a *wave-DWM* (W-DWM), as it uses travelling-wave variables. This formulation requires a pressure update for every scattering junction, but also for the two travelling-wave variables in every waveguide, and there are 4 waveguides per scattering junction in a 2D rectilinear DWM and 6 per junction in a 3D rectilinear DWM system. There is an alternative, equivalent form of the DWM known as the *Kirchhoff-DWM* (K-DWM) [201] which uses Kirchhoff-type physical variables in place of travelling-wave variables. In the K-DWM, only the pressure values at the scattering junctions require updating every iteration, saving significant computational expense. The K-DWM formulation is directly equiv-

alent to an FDTD simulation in the band-limited case [201] but as before, is calculated using impedance as a free parameter, permitting the modelling of heterogeneous media in the same way as the DWM models explored previously.

The K-DWM has not been used in the previous chapters as the W-DWM is known to be more robust to numerical error [201], considered an essential factor in the development of large systems with many hundreds of thousands of scattering junctions such as the 3DD-DWM model. Furthermore, the use of travelling-wave variables means that the W-DWM can be directly integrated with wave digital filters, planned as a means of implementing frequency-dependent losses in the model in the future. However, in the present case use of the K-DWM was found to be necessary in order to reduce the training time of the DNN system to manageable levels. The K-DWM formulation replaces the update steps of the W-DWM ((4.7), (4.9) and (4.10)) with a single expression to be completed for every time step, n, at every scattering junction location, (x, y, z), within the mesh:

$$p_J(n) = \frac{2\sum_{i=1}^{N} Y_{J,J_{nei}} p_{J_{nei}}(n-1)}{\sum_{i=1}^{N} Y_{J,J_{nei}}} - p_J(n-2) \tag{7.1}$$

Note, by comparison with (4.7), that Kirchhoff variables such as $p_{J_{nei}}(n-1)$ are used in place of travelling-wave variables such as $p^{+}_{J,J_{nei}}(n-1)$, and the pressure at the current junction is based on the values of neighbouring junctions at the previous time step and its own value two time steps ago. This makes explicit a property of both K- and W-DWMs: not every junction pressure is required for every n; in fact every *other* junction pressure is, so the mesh effectively consists of two interleaved grids [202]; to put it another way, the mesh is effectively undersampled by a factor of two. This has the effect of reducing the valid frequency range to $f_s/4$, or half the Nyquist frequency. This will become relevant in Section 7.3.

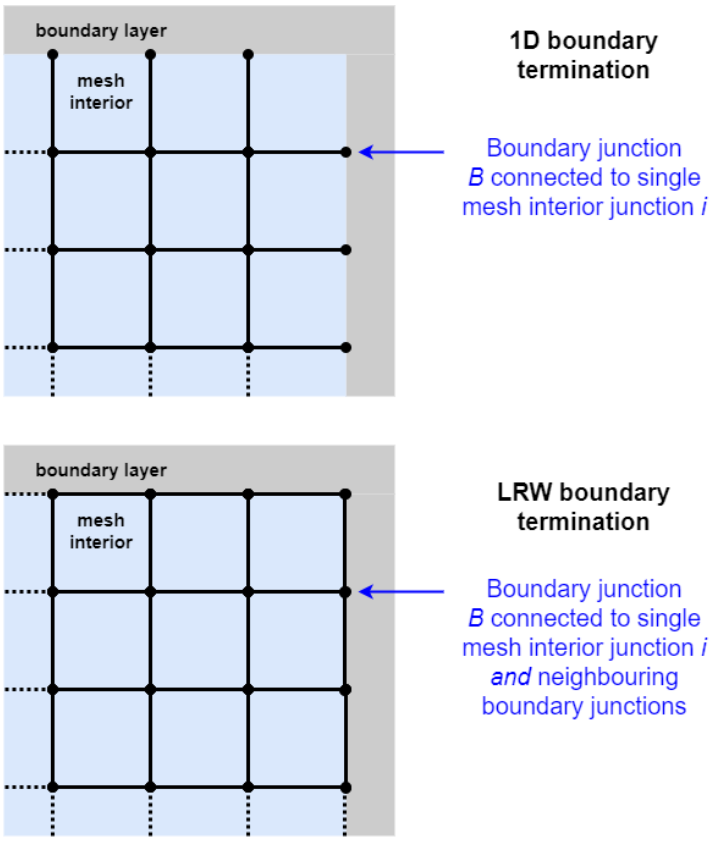


Figure 7.2: Comparison of 1D boundary terminations in the DWM (top), and LRW boundary terminations as described in Section 5.3 (bottom).

7.2.2 Simple DWM Boundaries

A further means to reduce the computational expense of the DWM model is to simplify the behaviour at the edge of the domain. The models previously introduced make use of a locally reacting wall (LRW) implementation [175], as described in Section 5.3, which models scattering junctions beyond a boundary as multi-dimensional junctions, connected to their neighbours much like a mesh-interior junction. This requires the implementation of complex expressions (5.2), (5.3) and (5.4). A much simpler model of domain boundaries exists, which treats the scattering junctions at a boundary as simple 1D terminations, connected to a single mesh-interior junction. Figure 7.2 illustrates the difference between the two methods. In place of the complex expressions for LRW boundary updates, a simple 1D DWM boundary is updated as follows [203]:

$$p_B(n) = (1 + r)p_{J_{nei}}(n - 1) - Rp_B(n - 2) \tag{7.2}$$

where p_B is the pressure at the boundary node, $p_{J_{nei}}$ is the pressure at the one connected interior waveguide (see Figure 7.2) and R is the reflection coefficient describing how much energy is reflected at the boundary; for passive boundaries, $|R| \leq 1$. The 1D boundary model is known to introduce error into the simulation [175], particularly at high angles of incidence, as multidimensional wave propagation is converted into 1D propagation. Furthermore, following (5.5), the equivalent waveguide length will differ in the 1D parts of the domain, altering the location of the boundary. However, the 1D boundary was used in the original 2DD-DWM model [147] and did not significantly affect the production of identifiable phonemes so is considered sufficiently accurate for the current purpose.

7.3 The DNN-driven 2D DWM Synthesiser

The requirement for a control model capable of driving the DWM synthesiser based on input text motivates the combination of DWM and DNN-based syn-

thesis approaches. However, there are also a number of advantages to the approach from the point of view of DNN synthesis. For example, while speaker characteristics such as speaking style and emotion can be implemented in DNN-based synthesis by transforming the model parameters [204, 205], the relationship between these parameters and actual vocal tract shapes is unclear. This makes it difficult to synthesise speech with different speaker characteristics such as age and gender. Although the transformation can be estimated using data-driven techniques, reliable estimation requires a large amount of speech data containing the desired characteristics. A more flexible approach is offered by physical modelling synthesis, where changes to speaker characteristics may be implemented simply by changing the size, shape, and other parameters of the modelled vocal tract. It is also hoped that the combination of the two methods will eventually lead to a reliable automatic estimation of vocal tract shape, eliminating the need for MRI data which is difficult to obtain (see Section 5.1.3). A similar approach was used in [206] to obtain vocal tract shape data from speech recordings using a genetic algorithm, but this model did not link to a TTS front end, and was concerned only with static vowel sounds.

The combined DWM-DNN model presented here is based on the simple concept of replacing the vocoder parameters typically estimated by DNN-based TTS systems [63] with the set of admittances required to describe a 2D DWM vocal tract model. The impulse response of the resulting DWM is then calculated, and used to filter a voiced (pulse train) or unvoiced (white noise) excitation signal, producing an output signal which can be compared to recorded speech during the training phase. The DNN weights are then updated on the basis of the error between the synthesised and recorded speech frames. A similar approach using a spectral filter was introduced in [77, 200], but to the author's knowledge this is the first attempt to estimate a physical model, which provides information about the vocal tract shape as well as a filtering operation, using such a technique.

7.3.1 Mesh Construction

As described in Section 7.2, a K-DWM implementation with simplified boundaries is required if the training time of the DNN system is to be kept within reasonable limits. A further consideration is the sampling frequency, f_s of the mesh, which affects spatial resolution, Δx, following 5.5. Initial attempts using the DNN-DWM model used $f_s = 48$ kHz, giving $\Delta x \approx 1.03$ cm. This was used to construct a mesh with 9 junctions across its width and 18 junctions along its length, giving an equivalent physical size of roughly 17.51×8.24 cm, corresponding to the dimensions of a male vocal tract [20]. However, this resulted in 247 admittance parameters to be estimated, and up to 46 further parameters representing reflection coefficients (depending on whether they were to be grouped into, for example, 'vocal tract wall' reflection coefficients, or estimated individually). This was found to be too many parameters for the model to learn in any reasonable amount of time. The relatively high sample rate (the waveform-estimation DNN system [77] uses $f_s = 16$ kHz) also means more samples of impulse response are required, increasing computation time and hence training time further.

As a trade-off, a sampling frequency of 24 kHz was used in the final version of the model. This provides a spatial resolution of 2.06 cm, and a 9×5 grid of scattering junctions is used, making the equivalent physical size approximately 16.48×8.24 cm. The spatial resolution is very coarse given the fine structure of the vocal tract; however, the number of admittance parameters that must be estimated is reduced to 52. Additionally, the values of the boundary reflection coefficients were fixed, following the values given in [146]: $r_{glottis} = 0.97$, $r_{lips} = -0.9$ and $r_{sides} = 0.92$. These values were used in [147] to produce acceptable vowel sounds, so were considered sufficient for this proof-of-concept study. With 52 parameters to estimate and impulse responses calculated at 24 kHz, approximately 3 weeks were still required to train the DNN-DWM system using a parallel GPU implementation.

The reduction in sampling frequency introduces several important considerations for the output of the DNN-DWM model. The DWM output is valid only to $f_s/4$, as described in Section 7.2.1. Additionally, *dispersion error* is

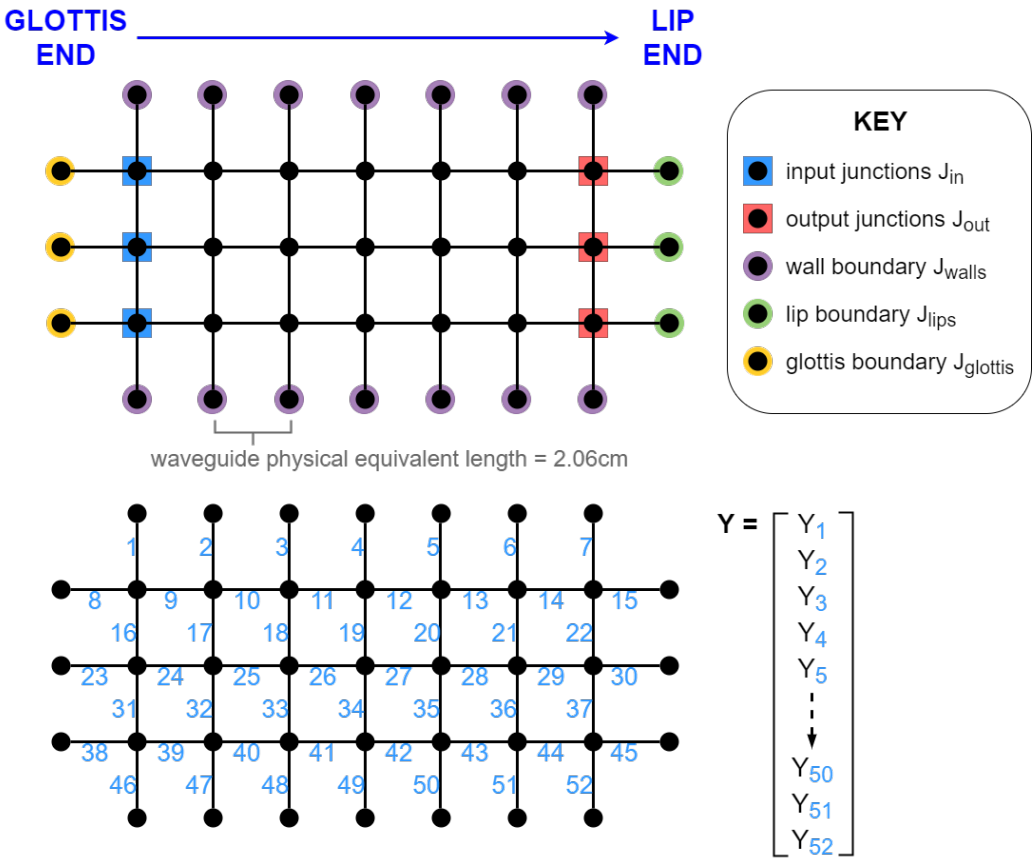
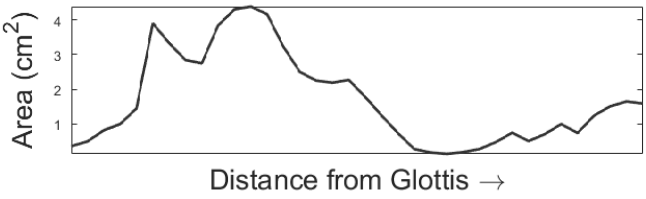


Figure 7.3: The 2D DWM formulation used in the combined DNN-DWM model, illustrating 1D boundary connections and input and output locations (top), and the construction of admittance vector $\mathbf{Y}$ (bottom).

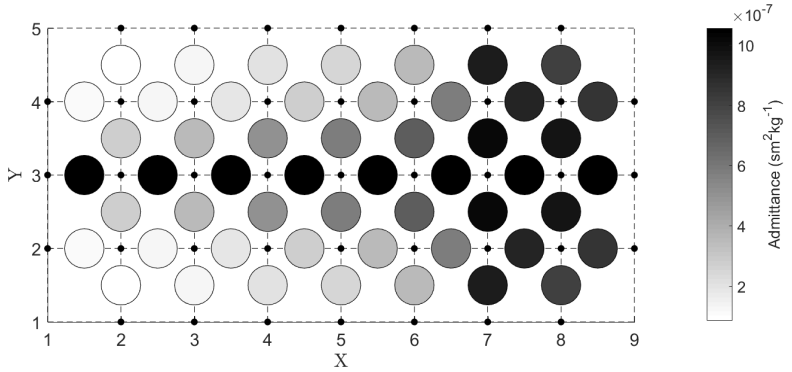
introduced by the Cartesian grids used in FDTD and DWM models. Von Neumann error analysis on such models [207] indicates that wave propagation speed is different along the axial directions compared to the diagonals. This produces a frequency-dependent error that can lead to mistuned resonances in the model [207]. The effect of dispersion error becomes perceptually significant above about 15% of the sampling frequency [124]. These concerns have not been relevant until now, with the minimum sampling frequency used in previous chapters being 384 kHz, placing $f_s/4$ and even $0.15 \times f_s$ well outside the range of human hearing. However, the DNN-DWM model uses $f_s = 24$ kHz, so the output is only valid up to 6 kHz—sufficient for intelligibility, but not naturalness [79]—and perceptual errors may be audible above 3.6 kHz. The effect of these limits must be considered when evaluating the output.

An illustration of the DWM structure is given in Figure 7.3. To produce a vector of admittances, $\mathbf{Y}$, for estimation by the DNN, an arbitrary row-wise assignment is made as illustrated. The scattering junctions used as input and output points are also highlighted. Boundary junctions are not used as input or output locations due to their non-physical 1D coupling with the mesh interior (see Section 7.2.2), so the next-nearest set of junctions to either end of the mesh are used.

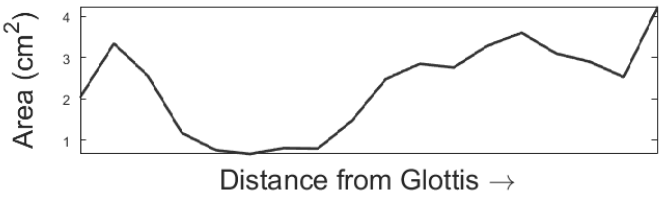
Example admittance maps generated using the 2DD-DWM raised-cosine mapping procedure on the 24 kHz model are illustrated in Figure 7.4. The area functions are interpolated to match the length of the vocal tract model, and then a raised-cosine admittance function is applied across the mesh, following the procedure detailed in 5.2.1. The dark areas in Figure 7.4(c) and (d) represent areas of high admittance, or low impedance, as previously seen in Figure 5.4. These admittance maps are provided by way of illustration only; in the combined DNN-DWM system, the cosine-mapping constraint used in [147] is removed, taking advantage of the 2D structure by permitting the generation of asymmetrical admittance maps. The figures illustrate that the admittances in the DWM exist *between* scattering junctions, as admittance is a property of the waveguides rather than the junctions.



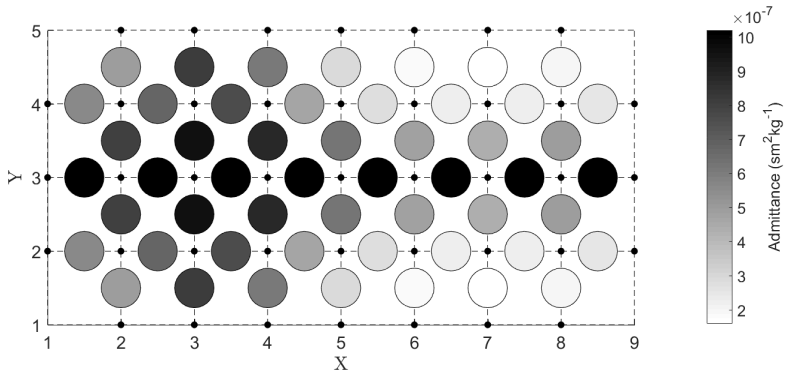
(a) /i/ area function



(b) /i/ admittance map



(c) /ɑ/ area function



(d) /ɑ/ admittance map

Figure 7.4: Construction of 2D raised-cosine admittance maps (b) and (d) from vocal tract area functions (a) and (c), for English vowels /i/ and /ɑ/. The underlying DWM grid and scattering junctions are represented by the smaller circles and grid.

7.3.2 Optimisation Method

Once a vector of admittances, $\mathbf{Y}$, has been defined as above, the DNN may be trained. Frames of speech are used, and frame number is denoted by τ to distinguish from continuous time, t, and discrete time, n, used elsewhere in this thesis. The input to the DNN at frame τ is the linguistic feature vector $\mathbf{l}_\tau$, and therefore $\mathbf{Y}_\tau$ is calculated by:

$$\mathbf{Y}_\tau = \mathcal{H}(\mathbf{l}_\tau) \tag{7.3}$$

where $\mathcal{H}$ denotes the nonlinear function represented by a DNN.

Training the DNN requires the definition of an objective function to be maximised. Following [77], the log likelihood function is used for this purpose. The speech is approximated with a zero-mean Gaussian distribution:

$$P(\mathbf{s}_\tau \,|\, \mathbf{Y}_\tau) = \mathcal{N}(\mathbf{s}_\tau; \mathbf{0}, \Sigma_{\mathbf{Y}_\tau}) \tag{7.4}$$

where $\mathbf{s}_\tau \in \mathbb{R}^M$ is a discrete-time widowed speech signal based on a zero-mean stationary Gaussian process [208], $\mathbf{0} \in \mathbb{R}^M$ is the zero vector, and $\Sigma_{\mathbf{Y}_\tau} \in \mathbb{R}^{M\times M}$ is the covariance matrix that can be decomposed as follows:

$$\Sigma_{\mathbf{Y}_\tau} = \mathbf{H}_{\mathbf{Y}_\tau}^{\mathsf{T}} \mathbf{H}_{\mathbf{Y}_\tau} \tag{7.5}$$

where

$$\mathbf{H}_{\mathbf{Y}_\tau} = \begin{bmatrix} h_\tau(0) & & & 0 \\ \vdots & h_\tau(0) & & \\ h_\tau(N-1) & \vdots & \ddots & \\ & h_\tau(N-1) & \vdots & h_\tau(0) \\ & & \ddots & \vdots \\ 0 & & & h_\tau(N-1) \end{bmatrix} \tag{7.6}$$

and $\mathbf{h}_\tau$ is the impulse response of the 2D DWM, and N denotes the impulse response length. The impulse response $\mathbf{h}_\tau$ is calculated in a recursive manner, based upon (7.1), by taking the average of pressures across the output

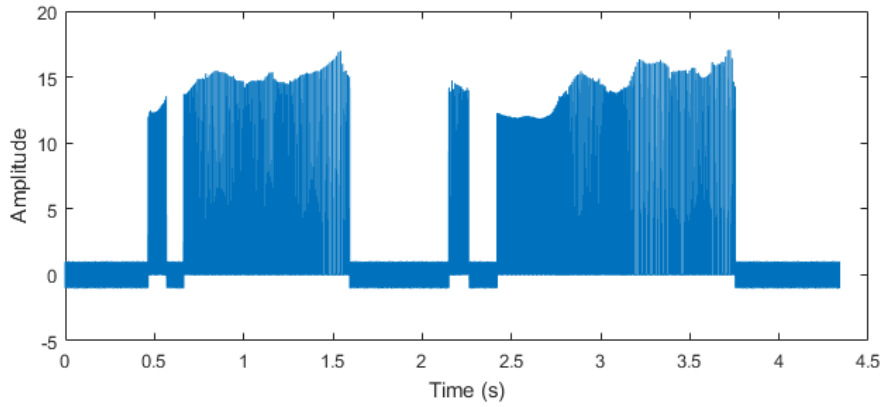


Figure 7.5: Example excitation signal used by the DNN system in the resynthesis of an utterance. The signal consists of a pulse train, with low-amplitude noise during unvoiced portions of the synthetic speech, and frequency and amplitude during voiced portions determined from natural speech.

junctions J_{output}:

$$h_\tau(n) = \frac{1}{|J_{output}|} \sum_{J \in J_{output}} p_J(n) \tag{7.7}$$

where J_{output} is a set of scattering junctions near the lips (see Figure 7.3). In order to generate an impulse response, a unit impulse is input at sample $n = 0$, split uniformly over the input junctions J_{input}:

$$p_J(0) = \begin{cases} 1 \,/\, |J_{input}|, & \text{if } J \in J_{input} \\ 0, & \text{otherwise} \end{cases} \tag{7.8}$$

In this framework, speaker characteristics can be controlled by using adaptation techniques, e.g., feeding speaker codes [209] to the DNN $\mathcal{H}$, as used in the standard DNN-based speech synthesis. In order to generate a smooth speech trajectory, the dynamic behaviour of the mesh is usually captured using the delta and delta-delta parameters (see Section 7.1) in the model. In the present study, these parameters are omitted for simplicity.

Once trained, the DNN can be used for TTS synthesis. In the synthesis stage, each speech frame $\mathbf{s}_\tau$ is produced by calculating the impulse response,

$\mathbf{h}_\tau$, of the DWM using the values of $\mathbf{Y}_\tau$ predicted by the DNN from the linguistic feature vector $\mathbf{l}_\tau$, and convolving the impulse response with a suitable excitation signal. The excitation signals used for this study are produced following the method in [10], where a pulse train is produced with frequency and amplitude calculated based on training data and using a separate model to the DNN. An example excitation signal for a complete utterance is illustrated in Figure 7.5 and shows how voiced and unvoiced sections of speech are implemented by the source model, with low-amplitude noise used during unvoiced sections rather than complete silence for improved naturalness. As in [200], noise can also be used in place of a pulse train to provide unvoiced excitation, such as whispering, to the vocal tract filter.

7.4 Pilot Study

7.4.1 Method

A pilot study was performed, using the DWM as described above without any additional constraints, save that all admittances generated should be positive. This choice was made in order to determine how the DNN-DWM system performed without guidance. Imposing the 2D raised-cosine mapping approach discussed in Section 5.2.1 was considered, but was believed to be too severe a constraint; for example, there is no capacity to model the nasal tract or other side branches using this method (nasal tracts can be modelled using the cosine-mapped 2D DWM, e.g. [148], but require an additional set of scattering junctions to be coupled to the mesh). Without a raised-cosine constraint, it should be possible for the model to approximate side branches, such as the nasal tract, by developing two channels of high admittance separated by a region of low admittance. However, by imposing very few constraints, non-physical aspects must also expected in the resulting admittance map—after all, without constraints the 2D DWM could be a model of a generic 2D membrane as much as a 2D representation of the vocal tract. Nevertheless, the results are expected to help illustrate the priorities

for future development of the DNN-DWM system.

The model was trained using the ATR B-set [210] of phonetically balanced Japanese speech from a male speaker. The database comprises 503 sentences, of which 450 were used for training the model, and the remainder used to test its performance. The speech signals were downsampled to 24 kHz and split into overlapping 25 ms frames with a 5 ms delay between frames. A 700-sample impulse response was calculated, corresponding to approximately 30 ms. At each time frame, τ, the 52-dimensional admittance vector $\mathbf{Y}_\tau$ was used in a DWM model to produce an impulse response, which was convolved with the excitation signal to produce a waveform that added to the continuous output vector using the overlap-add method.

The linguistic feature vector used as the network input had 411 dimensions, consisting of 408 linguistic features (including binary features such as phoneme identity and numerical features for contexts) and three duration features. The network had 3 hidden layers with 256 units per layer. The parameters of the network were randomly initialized, and were optimized using an Adam optimizer [211] with dropout [198] to maximise the log likelihood function. Sigmoid activation functions were used throughout the network.

7.4.2 Results

The 53 utterances remaining in the ATR database were used to test the DNN-DWM system. During informal testing, Japanese listeners found the sentences to be intelligible. The results that follow are based on the utterance 小さな鰻屋に熱気のようなものが漲る。, pronounced /ʧisanaunagijaninekkinojonamonogaminagiɯu/, which translates roughly as "the small eel restaurant is filled with something like a heated atmosphere". The associated audio files are provided in the accompanying material (see Appendix A).

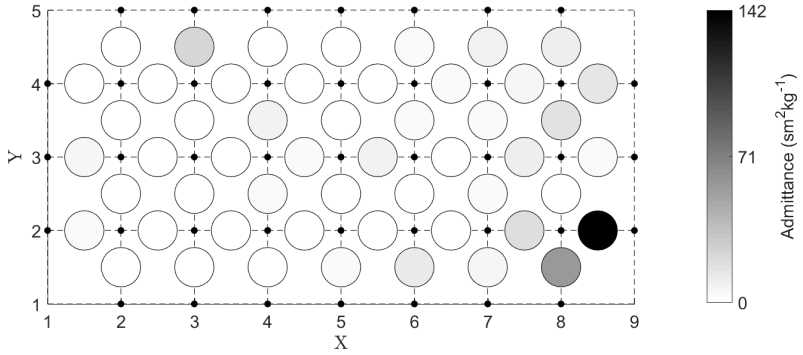
Vowels

Examples of admittance maps reconstructed from the DNN-generated vector $\mathbf{Y}_\tau$ for individual frames are provided for the vowels /a/, /i/ and /u/

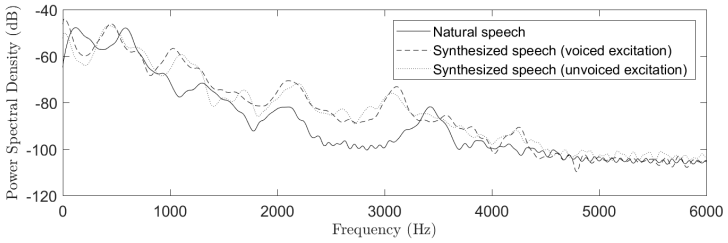
in Figure 7.6. Power spectral density (PSD) curves are also provided which compare the resynthesised speech frames (with both voiced and unvoiced excitation) to the original frame of natural speech. The PSD plots were calculated using a 4096-point FFT with a 300-sample window and 150-sample overlap, and smoothed with a 10-point moving-average filter using the MATLAB function `smooth` to remove the harmonic peaks that obscure the spectrum. Please note that throughout this chapter, a threshold has been applied to the admittance maps, so that the largest admittance value visible on all maps corresponds to 142 $\mathrm{s\,m^2\,kg^{-1}}$, which is the maximum admittance value present in several of the maps. Several other maps have a considerably higher maximum admittance, but only at one or two points in the map, and displaying these accurately unduly affects the resolution in the lower-admittance map regions. Furthermore, it is the *relative* admittance values that affect the simulation, so it is sufficient to note where an admittance is significantly larger than surrounding values rather than its exact value.

It is immediately apparent from Figure 7.6 that the admittance maps for different phonemes show several similarities, including a region of high admittance at $x = 8.5$, $y = 2$. As the model output is taken at $x = 8$, any high admittances beyond this may represent the expansion of the airway beyond the lips. Admittances generally appear to get larger towards the front of the mouth for all vowels—a pattern repeated for nasals and approximants as seen below—but show some phoneme-dependent variation in both dimensions. There also appears to be an area of high admittance centred on $x = 3$, $y = 4.5$ without an obvious physical interpretation; again this is present for approximants and nasal consonants, and additionally for fricative consonants, as will be seen below. It is suggested that these areas of high admittance may be local maxima obtained during the optimisation procedure, which might be addressed by introducing constraints into the model.

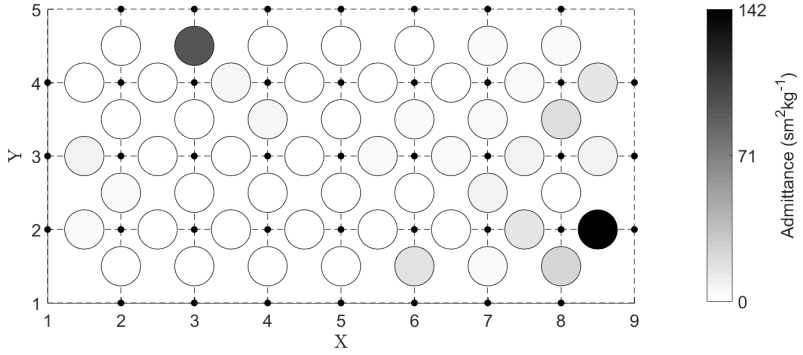
Comparing the DNN-generated admittance maps to the examples in Figure 7.4 that were calculated using the 2DD-DWM raised-cosine technique, it is clear that the DNN-DWM generated maps are less easily interpreted in terms of vocal tract behaviour. In Figure 7.6, regions of higher admittance



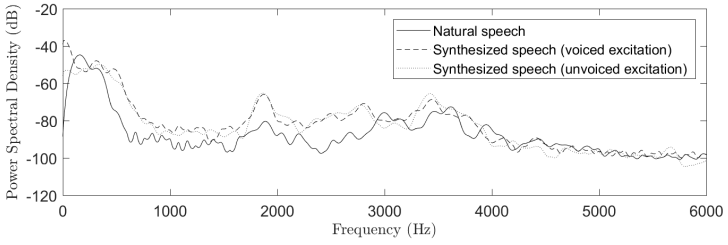
(a) /a/ admittance map



(b) /a/ PSD

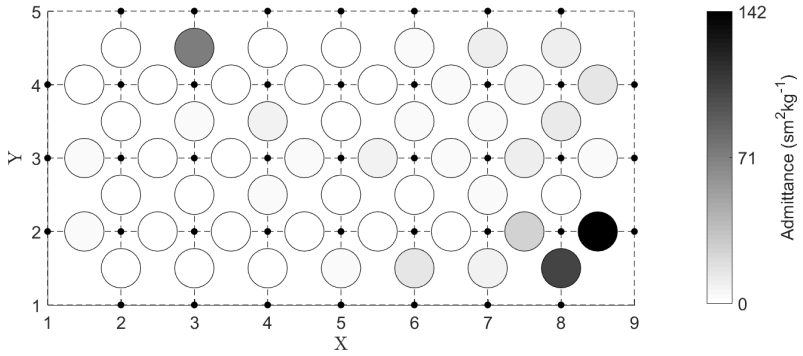


(c) /i/ admittance map

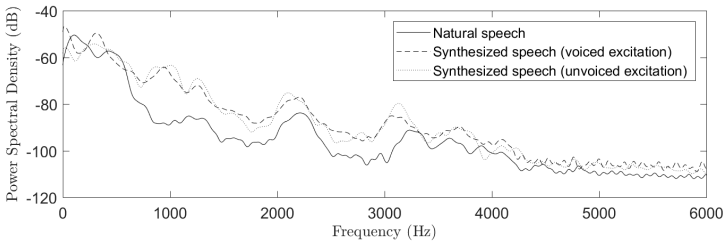


(d) /i/ PSD

Figure 7.6: Continued overleaf



(e) /u/ admittance map



(f) /u/ PSD

Figure 7.6: DNN-generated admittance maps and power spectral density (PSD) graphs for Japanese vowels /a/, /i/ and /u/. PSDs have been smoothed using a 10-point moving-average filter.

Phoneme	Formant error (Hz)				Formant error (%)				
	F1	F2	F3	F4	F1	F2	F3	F4	M.A.
/a/	-129	-264	-6	-317	-22.2	-20.4	-0.3	-9.2	13.0
/i/	-6	-12	-187	-170	-1.9	-0.6	-6.2	-4.7	3.4
/u/	-106	-229	-6	-170	-19.5	-19.5	-0.3	-5.3	12.5
							total M.A.		9.7

Table 7.1: Errors in formant frequencies for vowels synthesised using the DNN-DWM model, compared to those of natural speech. Formant frequencies are obtained from PSDs. M.A. is mean absolute error across all three formants.

are evident near the centreline and around the middle of the vocal tract ($3.5 < x < 7$) for the vowels /a/ and /u/. Regions of lower admittance are present behind these areas, suggesting more constriction at the back of the vocal tract and a relatively open front cavity, as expected for these vowels. The open regions towards the front of the vocal tract are not present in the admittance map for /i/, indicating a close-front vowel as expected, although the open back region, also characteristic of a /i/, does not appear to be present.

The formant errors for the three vowels, calculated based on the PSD functions, are presented in Table 7.1. It can be seen that, as in Chapter 5, the formant accuracy results appear to be phoneme-specific. In Tables 5.1 and 5.2, the 2DD-DWM approach obtained a mean absolute error of around 13% across 5 formants and 6 vowels; the DNN-DWM approach obtains a mean absolute error of 9.7% across 4 formants and 3 vowels. These results are not directly comparable, but do suggest that the formant accuracy in the mesh is surprisingly good given its comparatively poor spatial resolution (2.06 cm in the DNN-DWM model, compared to 1.24 mm in the 2DD-DWM model of Chapter 5). The only phoneme directly comparable to the 2DD-DWM results presented in Table 5.2 is /a/, and the formant errors are of similar magnitude in both cases. Mistuned higher formants may be attributed to the effects of dispersion error as discussed in Section 7.3.1. The PSD curves in Figure 7.6 indicate that the general trend of the simulated speech spectrum

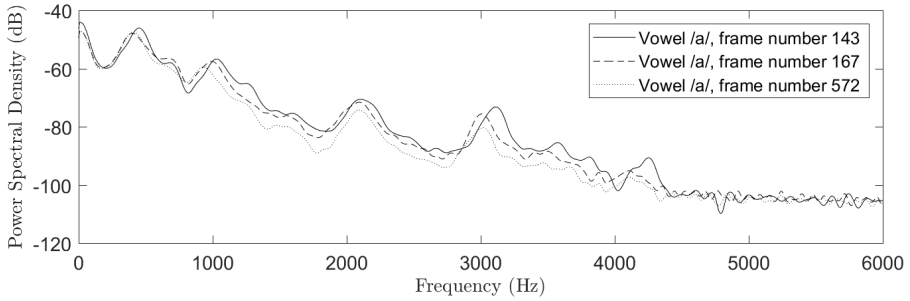


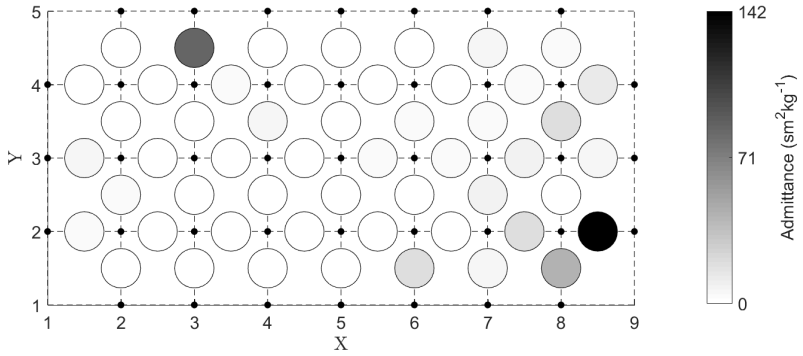
Figure 7.7: Power spectral density (PSD) graphs of voiced synthesised speech frames corresponding to the vowel /a/ at different times during the utterance, illustrating the consistency of the modelling. PSDs have been smoothed using a 10-point moving-average filter.

follows that of the recorded speech much more closely than the simulations presented in Figures 5.19–5.24. Further investigation is required to determine how this is achieved.

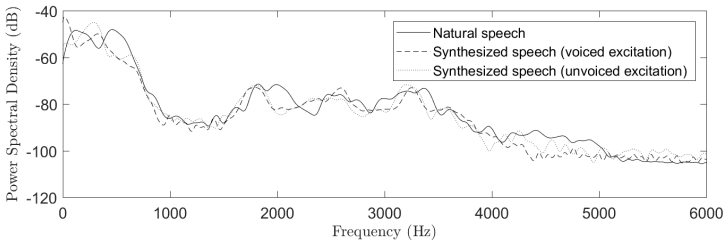
Since the input to the DNN-DWM system contains a large number of dimensions relating to context, it is of interest to determine how consistent the production of vowels is at different points during the utterance. Three frames that represent vowel /a/ have been selected: frames 143, 167 and 572 of the 859-frame synthesised signal, each corresponding to a different context for the vowel. The admittance map of vowel /a/ at frame 143 has already been provided in Figure 7.6(a), and the admittance maps in the other two contexts are very similar so have been omitted for brevity. The PSDs of the three synthesised vowels (using voiced excitation) can be seen in Figure 7.7, where PSDs have been calculated and smoothed as before. It is clear from this figure that formant reproduction is consistent across several vowel contexts. Future studies might consider the effect of different coarticulatory contexts on consonant modelling.

Approximants

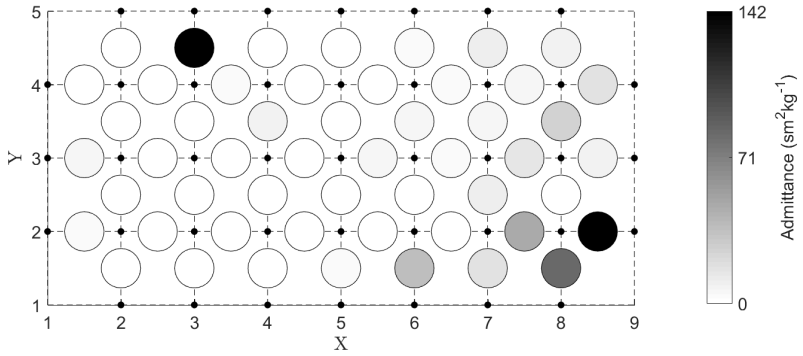
The utterance under study additionally contains two approximants, /j/ and /ɹ/, and the DNN-generated admittance maps and PSDs for these phonemes



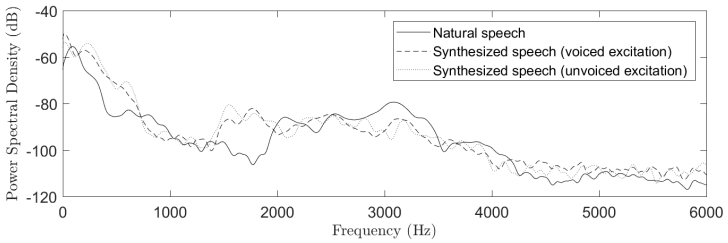
(a) /j/ admittance map



(b) /j/ PSD



(c) /ɹ/ admittance map



(d) /ɹ/ PSD

Figure 7.8: DNN-generated admittance maps and power spectral density (PSD) graphs for Japanese approximants /j/ and /ɹ/. PSDs have been smoothed using a 10-point moving-average filter.

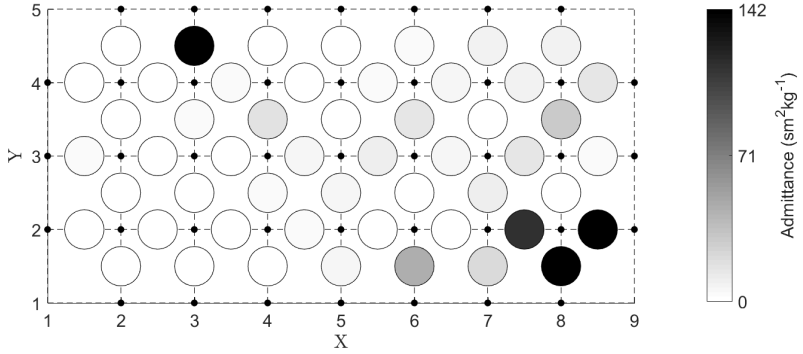
are presented in Figure 7.8. The erroneous local maxima described in the previous section are present in the admittance maps, along with the general trends of the vowel maps. As expected, the map for /j/ is similar to that of /i/, having a low admittance along most of the vocal tract length, with the closed front cavity well-represented, but the more open back cavity not apparent in the map. The map for /ɹ/ shows considerably higher admittance towards the front, especially for $y < 2.5$, although a physical interpretation of this difference is not obvious.

The PSDs of the approximants again illustrate that the model reproduces the losses in the vocal tract very well. The results for /j/ indicate that formant frequencies are well reproduced by the model. The PSD for /ɹ/ shows some significant differences in formant frequencies between the natural and synthesised speech frames. As there are fewer instances of the approximants in the test utterance it is difficult to make much of an assessment based on audition, but approximants in word-final syllables do appear to sound like their respective phonemes.

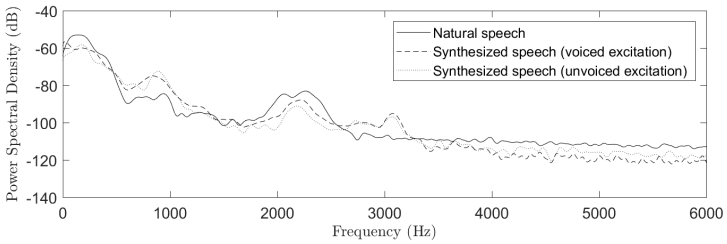
Nasals

Nasal consonants /m/ and /n/ are present in the utterance under study, and provide an opportunity to assess the performance of the DNN-DWM model for a system with a side branch. The DNN-generated admittance maps and PSDs are presented in Figure 7.9. The PSDs for the nasal consonants replicate the natural speech PSD particularly well, with the only major difference being that the synthesised versions of /n/ fail to reproduce the antiresonance at around 800 Hz present in the natural speech signal. This antiresonance is characteristic of an /n/ [12], so upon audition the /n/ phonemes do appear to sound more like /m/, but an acceptable nasal consonant is produced. The rest of the spectral content is well-reproduced in both cases.

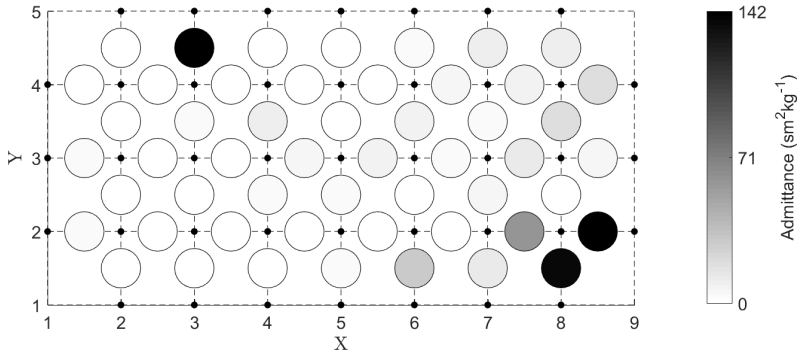
The DNN-generated admittance maps for the nasal consonants contain similar general features to the vowel and approximant admittance maps previously described. Both maps have higher admittances in the region $x \geq 7$ than the maps seen previously, and it is possible that this is an attempt by the



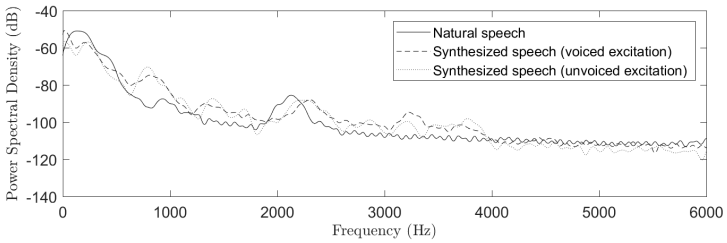
(a) /m/ admittance map



(b) /m/ PSD



(c) /n/ admittance map



(d) /n/ PSD

Figure 7.9: DNN-generated admittance maps and power spectral density (PSD) graphs for Japanese nasal consonants /m/ and /n/. PSDs have been smoothed using a 10-point moving-average filter.

system to reproduce the behaviour of a side branch by connecting the vocal tract model to the receiver position using more than one higher-admittance path.

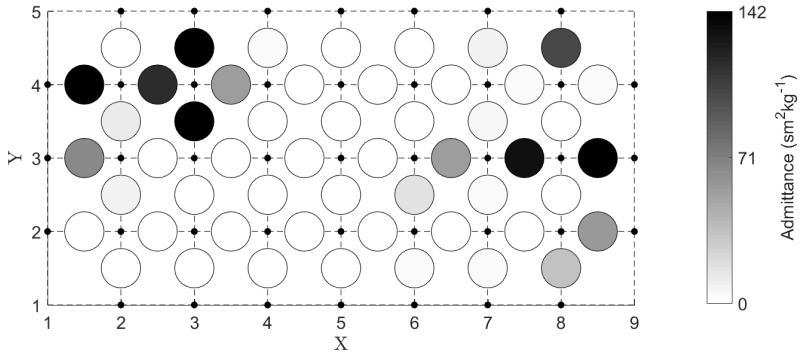
Fricatives

The final static phonemes present in the utterance under study are the fricatives /s/ and /ʃ/. In fact, /ʃ/ is not directly present, but the affricate /tʃ/ occurs at the start of the utterance, so later frames corresponding to this phoneme are assumed to represent /ʃ/. Figure 7.10 illustrates the DNN-generated admittance maps and PSDs for these phonemes. It is clear that several large errors exist in the PSD for /s/, particularly around 500 Hz and 4.2 kHz. Upon audition, the /s/ sounds like /f/ in the synthesised utterance, with both voiced and unvoiced excitation, indicating that although fricative noise is present, it is not correctly filtered by the model. The PSD of /ʃ/ is much closer to that of natural speech, and is readily identifiable within the synthetic speech.

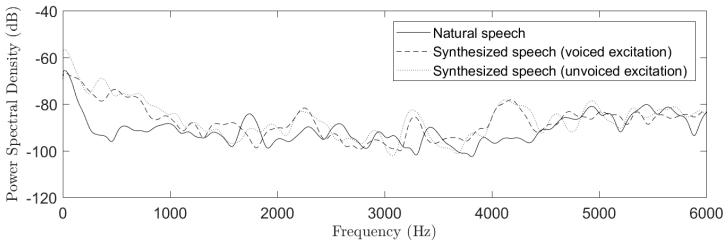
The admittance maps in Figure 7.10 are notable in that they do not feature such a large admittance maximum at $x = 8.5$, $y = 2$. This suggests that the unknown cause of this maximum is at least partially phoneme-dependent. Both fricatives also have areas of large admittance towards the back and front of the vocal tract, centred around $x = 3$, $y = 4$, with a region of lower admittance in between the front and back sections that may represent the constriction required in the production of a fricative.

Plosives

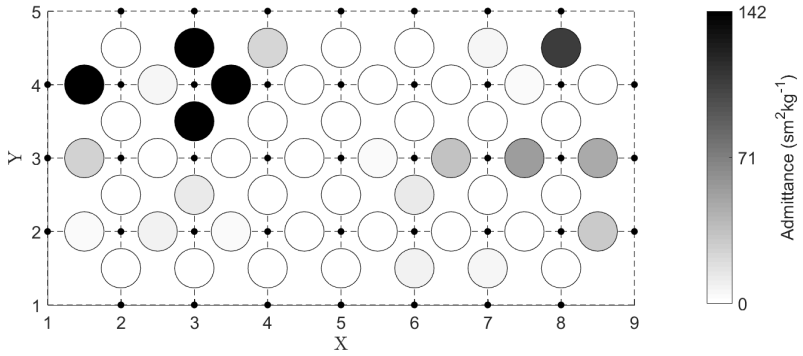
The utterance under study also contains two plosive consonants, /k/ and /g/. Due to the dynamic nature of these consonants, static admittance maps and PSD curves provide very little useful information. Instead, audition of the synthesised audio signal (see Appendix A) provides the most meaningful assessment of the modelling approach. In the author's opinion, the synthesised plosives all sound like /b/ when a voiced excitation is used, but the /k/ is



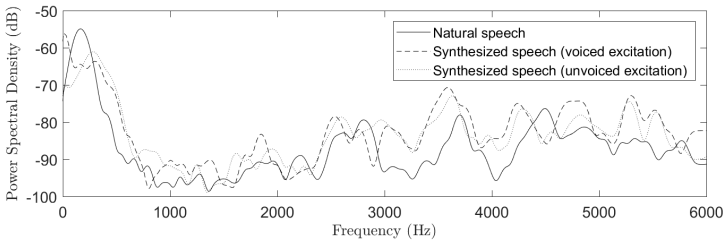
(a) /s/ admittance map



(b) /s/ PSD



(c) /ʃ/ admittance map



(d) /ʃ/ PSD

Figure 7.10: DNN-generated admittance maps and power spectral density (PSD) graphs for Japanese fricative consonants /s/ and /ʃ/. PSDs have been smoothed using a 10-point moving-average filter.

recognisable using an unvoiced excitation signal. This suggests that a mixed excitation signal is likely to be the most appropriate for such models in the future. The plosive nature of the phonemes are retained, even when the wrong plosive is identified, as the dynamic aspects such as the closure phase are well-reproduced. Additional constraints on the DWM parameters may improve the modelling of the vocal tract shape and therefore improve the formant transitions, leading to improved plosive identification in future versions of the model.

7.4.3 Summary

The results presented in this section illustrate that, while the DNN-DWM system produces reasonably accurate spectra for many phonemes, the admittance maps generated are difficult to interpret and are not always physically-related. This supports the findings of a previous study [206], which used genetic algorithms to estimate vocal tract shapes and found that physically realistic results were not always obtained.

The observations made above are largely generalizable to the respective phoneme categories across all test sentences: vowels are quite well estimated by the model, approximants and nasal consonants are less well estimated, and turbulent consonants are least well estimated. This may relate to the physical modelling paradigm used, as in general, physical models of vowel reproduction have been highly successful, but consonants have not yet been synthesized reliably. Further investigation into reliable consonant reproduction techniques is therefore expected to improve the model.

All of the $\mathbf{Y}$ vectors generated with the proposed technique display a much larger range of admittance values than the raised-cosine mapping technique illustrated in Figure 7.4, far exceeding the maximum characteristic acoustic admittance for a tube the size of a vocal tract. This occurs despite thresholding being applied to the maps as described previously; in fact several maps have a maximum impedance several orders of magnitude larger than that displayed. In addition, the DNN-generated admittance maps contain admit-

tance maxima that are present across multiple phoneme categories, which appear to be a consequence of the optimisation procedure getting stuck at local maxima. It is evident that, in order to obtain physically meaningful output, some constraints will be required upon the generated parameters, for example by including the positions and admittances of articulators like the teeth, hard palate and velum, although this is likely to require higher spatial resolution in the model. Nevertheless, the synthesised speech has a spectrum similar to natural speech and, according to informal listening tests, is sufficient for intelligibility. With further refinement it is anticipated that the quality of the synthesised speech will improve.

7.5 Discussion and Extensions

A number of extensions are obvious from the experiment described in the previous section. For example, the intelligibility of the model should be subject to perceptual tests in comparison with existing models in order to determine a baseline measure of performance. Furthermore, the range of admittances possible in the model should be constrained, perhaps to lie within the range given by the specific acoustic impedances of air and head tissue as defined in Section 5.4.2, 399 $\mathrm{Pa\,s\,m^{-3}}$ and 83666 $\mathrm{Pa\,s\,m^{-3}}$ respectively. A higher impedance may be necessary to model structures like bone and teeth, but will still result in a range of admittances with fewer orders of magnitude than that generated by the unconstrained DNN-DWM model. The reflection coefficients should be jointly estimated with the admittance values, as for a 2D model the reflection coefficient must vary along the vocal tract length in order to reproduce realistic 3D vocal tract behaviour [149]. A more appropriate initialisation technique, such as initialising the DNN weights with values that produce an admittance map representing a cylinder of average vocal tract radius, should be implemented. Finally, delta and delta-delta parameters should be incorporated to model dynamic behaviour as is commonly done in SPSS systems [59].

The 2D DWM implemented at present has no geometrical constraints except

its overall rectangular shape. A decision was made not to introduce further constraints for this proof-of-concept study, in order to determine how well the DNN would estimate the DWM parameters without them. While the results show some features that may be interpreted as similar to physical vocal tract behaviour, it is clear that some degree of constraint is needed in future versions of the model, perhaps based on MRI geometries. Without these, the DWM model is not necessarily a vocal tract model at all, but a 2D membrane that could learn to reproduce any sound given appropriate training data. As [196] puts it, "models with a large number of free parameters can describe an amazingly wide range of phenomena". The implication is that the proposed model could be improved by introducing some vocal-tract specific constraints, which should reduce the size of the parameter search during optimisation, making it less likely that the model will encounter local minima.

If the 2D DWM is retained as the modelling approach, several issues need to be addressed. First, the sampling frequency must be increased so that the valid frequency limit $f_s/4$ lies above the upper limit of human hearing. Practically, this may not be possible as it would introduce too many parameters for the DNN to train in a reasonable time, but increasing f_s so that the cut-off lies above 10 kHz would be sufficient to reproduce much of the high-frequency information that contributes to naturalness [79]. Furthermore, dispersion error in the mesh has a perceptual effect above around $0.15 \times f_s$, indicating that a higher sampling frequency may be needed.

One way to increase the sampling frequency without unduly affecting the DNN outputs would be to implement parametrisation of the admittance map so that the DNN does not need to estimate every individual admittance as an output parameter. Such parameters might be principal components of the admittance map, or some combination of spatial basis functions; alternatively the raised-cosine mapping of the 2DD-DWM method may be imposed so that only one admittance is required at distance x from the glottis in order to calculate the admittances across the whole mesh at x. The reduction in training time due to the comparatively small number of DNN outputs in these methods, however, is offset by the increased sampling frequency of the

mesh which requires more samples of impulse response to be generated, and a greater number of pressure updates within the mesh, for every iteration during the training phase.

One simple way of imposing the vocal tract shape onto the system would be to use a 1D rather than a 2D vocal tract model. This has the added benefit of requiring fewer parameters to describe the vocal tract, permitting greater resolution in the model without increasing the number of output parameters for the DNN to learn. This model might also include a nasal tract, coupled to the 1D vocal system at the appropriate location using a binary velum parameter (open/closed) or a continuous velar area. The 1D vocal tract model is computationally a lot simpler than even the 2D model, so the DNN training time is likely to improve. A first attempt at such a model is planned for the near future.

The discussion in this section illustrates that, even if 3D vocal tract models are parametrised to require a small number of control inputs, the long computation times required to obtain speech output are a significant hindrance to their implementation in a DNN model. Even a 2D DWM operating at 24 kHz (i.e. faster than real-time [147]) resulted in several weeks of DNN training time. Therefore, although the same techniques introduced in this chapter may theoretically be extended to 3D models with little alteration, this is unlikely to happen in the near future unless DNN training becomes significantly faster and the computational expense of calculating output from a 3D model is dramatically reduced. However, lower-dimensionality models may be leveraged instead, and with careful co-design of the vocal tract models and DNN systems, may begin to provide a viable alternative to existing TTS systems, as well as a means of estimating the vocal tract shape corresponding to input text, providing researchers with useful new information.

7.6 Conclusion

In this chapter, a speech synthesiser was produced that combines the 2D DWM technique with DNN-based SPSS TTS methods. Due to the inherent

complexity of the DWM approach and the long training times required for DNN systems, a highly simplified DWM vocal tract model with low spatial resolution and very few constraints was used. Despite this limitation, intelligible speech output was produced, and the formant error in vowel simulations was found to be comparable to that of the much higher resolution 2DD-DWM simulation in Chapter 5. These results suggest that similar models may provide a viable means of TTS synthesis in the future. Furthermore, with suitable constraints applied to the parameters generated by the DNN in order to guarantee vocal-tract-like behaviour, it may eventually be possible to obtain vocal tract shape information—if sufficient training data is available—for an individual without requiring time-consuming and expensive MRI scans.

The work presented in this chapter is a very early step in the process of designing a control system for complex 3D vocal tract models based on written text input. In its present state, the 3DD-DWM vocal tract model proposed in Chapter 5 has far too many control parameters, and too long a computation time, to be used in a DNN-based TTS system. However, as computational resources become ever more powerful, it is not unreasonable to believe that such systems are a possibility in the future. The next chapter discusses the conclusions that may be drawn from the entirety of this thesis, and presents a number of avenues for further research in order to achieve this goal.

Chapter 8

Conclusions and Further Work

> If a computer voice can successfully tell a joke, and do the timing and delivery [...] then that's the voice I want.
>
> Roger Ebert, 'Remaking my Voice' (TED Talk, 2011)

8.1 Thesis Summary

This thesis has presented a number of novel approaches for physical vocal tract modelling. In this section, a summary of the thesis and its contributions is provided on a chapter-by-chapter basis.

In Chapter 2, the acoustics of speech production was introduced. The human vocal mechanism is a complex system with many parts, and the effects of different aspects such as side branches and frequency-dependent losses were discussed here. The production of different phoneme types was also described. This thesis has focused upon the production of vowels, but developing speech synthesisers using physical models will also require the production of consonants to be addressed. This is discussed further below.

Chapter 3 introduced text-to-speech (TTS) synthesis, and the wide range of

speech synthesisers currently available. From the early formant synthesisers, through diphone synthesis, to state-of-the-art unit selection and statistical parametric techniques, synthetic speech technology has become ever more natural, but ever more removed from the physics of the vocal system. It is intuitively apparent that articulatory synthesis, which seeks to incorporate the properties of the real vocal system into a speech synthesiser, offers great potential and flexibility for natural synthetic speech output, produced using readily understandable control parameters. Articulatory synthesis has not yet become as widespread as other methods, as the computational power required to run the models—especially in real-time—has only recently started to become available. This thesis takes the view that computational power can be expected to keep increasing, making ever more accurate articulatory synthesis a real possibility for TTS systems in the near future.

Chapter 4 provided more detail on the physical modelling techniques available for application to the vocal system, and given the emphasis on naturalness in the current study, particularly focused upon the most accurate vocal tract models: those that incorporate the complete three-dimensional (3D) geometry of the vocal tract. A number of 3D vocal tract models—typically based on MRI scan data of the vocal tract—are available, but apart from the model proposed in this thesis, only a 3D FEM model developed during the EUNISON project [168] is capable of dynamic movement. This model requires very long computation times—over 1000 hours per second of output speech [11], depending on the implementation details—making it unsuitable for synthetic speech applications in the near future, despite being a useful tool for the study of vocal tract acoustics.

In Chapter 5, a model was introduced which aims to reduce the simulation times required by a 3D dynamic vocal tract model. Building upon the established 2D DWM [147] and 3D DWM [166] and FDTD [30, 164] models presented elsewhere, a novel heterogeneous 3D DWM vocal tract model was proposed which is capable of dynamic movement and requires significantly reduced computation—over 300× reduction in simulation time—compared to the 3D FEM model. This model also incorporated a novel method for

simulating the volume of air outside the mouth. The proposed model was compared to established 2D and 3D DWM vocal tract models and found to provide a reduction in overall formant error, although the first formant is consistently 30–40% too high in the proposed model, which appears to be due to the way the vocal tract walls are implemented. Additionally, a comparison with the dynamic 3D FEM technique was possible, as one diphthong synthesised using the method is available online [168]. Although it is difficult to draw conclusions on the basis of a single diphthong, results of the two methods appeared to be comparable in terms of formant reproduction, and both contained more energy in the frequency range above 5 kHz than recorded speech signals.

Objective results are useful in determining the similarity of synthesised speech to a recording, and such results may be related to naturalness, but do not provide a *general* measure of naturalness, nor indicate which features are perceptually relevant in such an assessment. Perceptual tests are therefore required, and these are presented in Chapter 6. Several pilot tests were performed in order to determine parameters of the model—how increased dimensionality affects the results, and the sampling frequency required—before a perceptual test was performed comparing the proposed model to the established 2DD-DWM vocal tract model and—for the one available diphthong—the 3DD-FEM model. The results suggest a clear improvement of the proposed model over the 2DD-DWM method in terms of naturalness, although the scores remain well below those assigned to natural speech. Listener comments indicate that one distinct cause of unnaturalness in the simulations is the poor modelling of losses above 5 kHz in the vocal system, resulting in too much high frequency energy in the synthesised vowels. Results also showed a strong dependence on the diphthong being modelled, suggesting that some diphthongs present more of a challenge for synthesis. Comparison with the FEM model resulted in comparable naturalness scores, although where frequencies above 5 kHz were filtered out, the FEM model was rated as more natural. This is probably due to the fact that the FEM modelling approach more accurately reproduces the vocal tract geometry [30], and once the confounding high frequencies were removed, listeners were able to perceive this

difference. High frequency losses are clearly a priority for all 3D vocal tract modelling techniques, whether FDTD/DWM-based or FEM-based.

The ability of a modelling technique to produce output speech sounds from input MRI data is an important test, but with MRI data being costly, difficult to obtain and not available for all subjects, a better control method is required if synthetic speech is to be reliably produced. Chapter 7 presented the results of an initial enquiry into the use of a statistical parametric TTS approach to control a DWM vocal tract model. This novel method leverages existing developments in SPSS methods, such as text pre-processing and excitation generation, which have been shown to produce more natural-sounding synthetic speech. A DNN was used to learn the admittances of a 2D DWM model—more complex models are not currently viable using this system, but may be in the future—for each context-specific phoneme in the training dataset. Tests on previously unseen data indicate that the model generalises well enough to produce intelligible speech. The admittances generated by the model, however, do not show much physical relevance, suggesting that greater constraints are required on the model output. Plans are in place to develop a much more highly-constrained vocal tract model as the output to the DNN system, with the aim of not only producing a TTS synthesiser but also of estimating vocal tract shapes directly from input text, removing the requirement for MRI scanning.

8.2 Novel Contributions

This research, as described above, has resulted in the following contributions to the field:

Dynamic 3D DWM Vocal Tract Model The 3DD-DWM model presented in Chapter 5 is a novel approach to vocal tract modelling based on MRI vocal tract geometries.

Radiation Volume Modelling The radiation volume model detailed in Section 5.5 is novel in its use of locally-reacting walls at the bound-

ary and its application in a DWM-based model.

Model Validation The objective comparisons in Chapter 5 illustrate that the proposed model reproduces formants two to five more accurately than dynamic 2D and static 3D DWM vocal tract models. Furthermore, the perceptual tests indicate that the proposed model is considered to produce diphthongs that are significantly more natural than the dynamic 2D DWM vocal tract model.

Model Comparison The proposed model is compared to a state-of-the-art dynamic 3D FEM vocal tract model [11] and found to be both objectively and subjectively comparable.

Combined DNN-DWM Model The DNN-DWM model proposed in Chapter 7 is, to the author's knowledge, the first attempt to control an articulatory synthesiser using statistical parametric speech synthesis methods, and produces intelligible speech.

These contributions enable the hypothesis to be revisited in light of the results obtained.

8.3 Hypothesis

The hypothesis informing this thesis is as follows:

A detailed, three-dimensional dynamic digital waveguide mesh model of the vocal tract can produce more natural synthetic diphthongs than the two-dimensional dynamic digital waveguide mesh model.

The perceptual test results in Chapter 6 clearly support this hypothesis, as the proposed 3DD-DWM model output was rated as significantly more natural than the output of the 2DD-DWM model, for every diphthong studied, with a large effect size. However, the perceptual test results also indicated that the diphthongs synthesised using the 3DD-DWM were still significantly less natural than recorded speech, indicating that further model development is required. Listener comments suggest that the additional high-frequency

energy in the spectrum of the synthesised samples contribute considerably to perceived unnaturalness, so addressing this must be a priority for future work.

The results of Chapter 7 also suggest that the DWM method may be extended beyond diphthongs, and incorporated into a complete TTS system capable of whole utterances, with a few adjustments to the simulation method. Although this combined DNN-DWM technique will not be available for use with 3D vocal tract models for some time due to the large computational costs involved, it does illustrate the application of physical vocal tract models to practical speech synthesis problems rather than the study of the acoustic properties of isolated vowels.

The results of this thesis in their current form mostly indicate potential rather than concrete gains. While the hypothesis is supported in that the 3DD-DWM produces more natural *diphthongs* than the 2D method, the real priority for future development is synthetic *speech*. The 3DD-DWM system has the potential to generate more natural-sounding synthetic speech than 2D methods, and comparable results to other, more computationally expensive 3D methods, but only once all phonemes, including consonants, can be modelled. Similarly, the combined DNN-DWM system has the potential to become a natural-sounding TTS system in the future, with strengths particular to physical modelling such as the ability to implement speaker characteristics through parameters like vocal tract size, once appropriate constraints are developed. The end goal for this research, envisaged for the far future, is a 3D vocal tract model—controlled using SPSS or other, as yet undeveloped, techniques—based on an 'average' 3D vocal tract model which is personalised to match a desired speaker's vocal tract by means of a short adaptation period, perhaps informed by characteristics such as age, size and gender. Such a system has applications that extend far beyond TTS, into medical, forensic, and creative fields. In order to achieve these goals, a great deal of work will be required and some directions that this might take are outlined in the next section.

8.4 Future Work

Frequency-Dependent Losses

Objective results in Chapter 5 suggested that synthesised vowels contained more energy in the higher frequency range (above 5 kHz) than was present in recorded speech. However, it was the listener comments from the perceptual tests in Chapter 6 that illustrated how much this additional energy affects the naturalness of the synthesised signals. High frequency energy is lost in the vocal tract for many reasons, and [41, p. 2-13] provides a comparison of the different loss mechanisms and their relative magnitudes, indicating that the effect of viscosity, head conduction and radiation losses all increase with frequency. Since a radiation volume and realistic facial geometry is incorporated into the proposed model, but viscous and heat conduction losses are not, the latter are considered a priority in the development of more accurate frequency-dependent losses in the model. This should reduce the high frequency error in the synthesised vowels and contribute to increased naturalness.

One method to determine the contribution of all loss factors would be to calculate the vocal tract transfer function (VTTF) of the real human vocal tract for comparison with modelled VTTFs. The human VTTF can be estimated using broadband noise excitation and an impedance-matching procedure following the method in [212]. By comparison with simulated VTTFs, it may be possible to design a filter that accurately reproduces vocal tract losses. Comparing the results across different phonemes and subjects, using both objective and perceptual measures, would indicate how well such a filter generalises.

In the long term, frequency-dependent losses should be incorporated by accurately modelling the tissue properties, including complex frequency-dependent admittance values, within the different tissues surrounding the vocal tract.

Turbulence

In its present form, the model is only able to reproduce vowels or vowel-like sounds such as approximants. In order to model fricative and plosive consonants, a model of automatic turbulence generation is required in the model. As noted in Section 2.6.1, turbulence occurs when the Reynolds number exceeds a certain critical value. The Reynolds number is calculated using flow variables, which may be calculated throughout the mesh (e.g. [30]) or—more efficiently where flow values are only required at certain points throughout the mesh—calculated from pressure values at any given point [213].

A complete model of turbulence is an unrealistic expectation for physical models in the near future, as the precise behaviour of turbulent flow is still not well understood. However, turbulence can be approximated by a noise source inserted slightly downstream of the constriction in the vocal tract [82], which is triggered automatically when the critical Reynolds number is exceeded and has an amplitude proportional to the Reynolds number. Automatic generation of turbulence using this approach should also lead to appropriate noise generation during plosives. Automatically triggering a noise source in this way is more complex in 3D than the 1D models used in [82], since the cross-sectional area of the tract is not governed by a single parameter, but with careful construction such a model should be possible without significant alterations to the proposed method.

Nasal Tract

In order to produce the full range of phonemes, it is also necessary to incorporate a nasal tract into the model. At present, MRI scans do not provide sufficient resolution to capture all the detail of the nasal tract and particularly the paranasal sinuses which are separated only by thin walls [26]. Computed tomography (CT) images offer better resolution, but utilise x-rays which present a radiation risk and are therefore not suitable for use on human subjects except where there is also a medical need. Therefore, modelling must

rely either on low-resolution data or existing nasal tract models such as [26] that are not subject-specific.

One advantage when it comes to modelling the nasal tract is that the geometry is largely fixed: with the exception of the velum, which opens and closes to adjust coupling with the oral tract, the nasal tract does not vary during speech as the rest of the vocal tract does, so a single scan may be sufficient to produce a nasal tract model. Tests are planned for the future to determine how much detail is required in the nasal tract model to achieve more natural-sounding synthetic speech.

Vocal Tract Data Collection

As noted above, MRI data does not provide sufficient resolution to capture the details of the nasal tract. There are also a number of other issues with MRI data collection. As described in Section 5.1.3, the unnatural position and loud noise in the scanner can lead to hyperarticulation and other effects that may result in non-representative vocal tract shapes. Furthermore, MRI scans are expensive, time-consuming, and not suitable for subjects with tattoos, metal implants or claustrophobia. Alternatives, such as CT scans mentioned above, also have their own disadvantages. In the long term, it would be desirable to avoid medical imaging altogether, and determine vocal tract shape based on acoustic information alone. This problem is known as acoustic-articulatory inversion and has been widely studied (e.g. [214]), so far without sufficient progress to determine personalised vocal tract characteristics. In future, it may be that an 'average' 3D vocal tract model can be developed, and aspects of this model optimised rather than attempting to infer the complete vocal tract shape from speech. This will remove the need for medical imaging and make personalised vocal tract models more accessible.

In the short term, MRI technology must be leveraged to obtain as much information about the vocal tract as possible. One area where this is particularly important is the identification of the teeth, which appear no different from air in MRI scan data. Recently, a new set of MRI scans was taken

by the author for a future project, and before any speech articulations were captured, the subjects were asked to push their tongue against and between their teeth so that the detailed 3D shape could be identified by contrast with the tongue tissue. It is intended that a similar technique to [29] will be used to superimpose these teeth shapes onto the MRI scans and therefore improve the accuracy of the simulations.

Modelling Paradigm

As discussed previously, the DWM approach was chosen for the proposed model as it offers good stability characteristics, but at the expense of increased run-time and memory requirements [201] compared to FDTD implementations. One area for future development will therefore be the production of an FDTD version of the model, which will require careful design to avoid instability but will further reduce the computational expense of the model, which is already superior to FEM-based approaches. Future work is also planned in collaboration with the developers of the 3D dynamic FEM model [11] to undertake a more rigorous comparison of the two methods incorporating objective and perceptual tests across a range of different sounds.

Control Model

Plans for the extension of the SPSS-based control model proposed in Chapter 7 have been elaborated upon in Section 7.5, but alternative control methods also present a number of future possibilities. Methods controlled using articulatory movements, such as the silent speech system proposed in [32], or even the neural commands intended to produce such movements, such as those used in brain-computer interfaces [117], offer a natural and intuitive way to control a vocal tract model, for example using muscle activations as recently demonstrated in [173].

The proposed 3D model currently features hundreds of thousands of control parameters, making it unsuitable for use with any of the control methods described above. Therefore, a priority for future development will be the

parametrisation of the admittance map into articulatory or gestural parameters with the intention of eventually incorporating the 3D model into a complete speech synthesis system.

8.5 Closing Remarks

There remains much to be done in the development of a natural-sounding synthetic voice. This thesis has illustrated the potential of three-dimensional vocal tract models for the generation of natural-sounding synthetic vowels, and the possibility of combining such a system with the highly successful approaches of statistical parametric speech synthesis. However, until a model of the vocal tract is developed that can produce every sound that the real human vocal tract is capable of; until a control method is developed that is suitable not only for text-to-speech applications but also for other input mechanisms required by those who cannot input text; and until all of these things can be implemented in real time, the challenges of voice synthesis have not yet been met. This research is likely to take many years, but will provide significant rewards and may one day allow researchers to reunite a patient with their *own* voice.

Appendix A

Index of Accompanying Multimedia Files

The accompanying CD consists of the following files:

- **audio** *Folder containing all audio files*
  - dnn-dwm *Audio files of recorded and synthesised speech corresponding to DNN-DWM study (Section 7.4)*
  - **final test** *Audio samples corresponding to final perceptual test (Section 6.4)*
  - **pilot1** *Audio samples corresponding to first pilot test (Section 6.2)*
  - **pilot2** *Audio samples corresponding to second pilot test (Section 6.3)*
  - **VTTFs** *Vocal tract transfer functions associated with Chapter 5*
    - **mesh extent** *VTTFs comparing mesh extent (Section 5.5.1)*
    - **monophthongs** *VTTFs for assessment of monophthongs (Section 5.6.2)*
    - **occluded branches** *VTTFs of vocal tract with occluded side branches (Section 5.6.2)*
    - **source position** *VTTFs comparing source position (Section 5.5.2)*

    - gaussian input.wav *Input signal for VTTF calculations (7.4)*

- **data** *Folder containing all data, including MATLAB scripts and figures*
  - **objective** *MATLAB scripts for objective comparisons in Chapter 5*
  - **subjective** *MATLAB scripts for subjective comparisons in Chapter 6*
  - **synthesis** *MATLAB scripts for synthesis of samples, including 3D vocal tract data in file 'final meshes for simulation.mat'*
- thesis.pdf *Electronic copy of this thesis*

Appendix B

Listener Comments from Perceptual Tests

This section lists the listener comments for perceptual tests in Chapter 6. Comments are presented verbatim, including any grammatical and spelling mistakes. Where possible, the sample to which the participant is referring has been indicated in square brackets, but due to the randomisation techniques used it was not always possible to determine which was the relevant sample.

B.1 First Pilot

This test compared 1D, 2D and simplified 3D DWM simulations and recordings of diphthongs using the MUSHRA methodology. An audio reference was not provided, but participants were given an example word to indicate what the vowel was supposed to sound like. Participants were permitted to comment on each individual diphthong comparison.

Comments during example question

Clip 3 *[3DS-DWM]* sounds like it's ringing a lot

Clip 3 *[3DS-DWM]* high frequency artefact

1 *[1D-KL]* has ahigher freq increasing pitch element as dipthong descends / 2 *[2DD-DWM]* high frequency hiss / 3 *[3DS-DWM]* has ahigher freq increasing pitch element as dipthong descends / 4 *[recording]* very human like

My feeling about 3 *[3DS-DWM]* sounding more like bay might be down to regional difference - in my accent the 'a' part is slightly longer than in some places. Not sure though, they're pretty similar.

clip 3 *[3DS-DWM]* sounded a bit like it had been put through a flanger so not natural but was identifiable as 'ay'. THe larger amount of high frequency almost breathiness adds to it sounding more like ay than the first one

Diphthong /eɪ/

1 *[3DS-DWM]* has definite buzz

Diphthong /aɪ/

Slight glitch at the end of clip 4 *[1D-KL]* particularly off-putting (although it is still the least natural)

4 *[1D-KL]* most human although there is a discernible note that decreased in pitch

higher frequency on clip 2 *[3DS-DWM]* made it feel more realistic to me

Diphthong /ɔɪ/

Clip 1 *[3DS-DWM]* sounds breathier than clip 3 *[2DD-DWM]* which makes it sound more human

2 *[1D-KL]* is a Rastafarian and has a high q formant that decreases in pitch to produce a discernible note

in synthesized samples, F2 seems not to reach all the way up to the /i/ position, letting the glide end seemingly at an 'uh' sound.

Diphthong /ɪə/

Aliasing-style effect is very obvious on clip 4 *[1D-KL]*

clip 4 *[1D-KL]* sounds like ingressive breath

4 *[1D-KL]* would be best but has as 'woop'

The human has a different accent to me, so I could mark them as not saying it right but I won't as I don't think that's what it's getting at. I'll listen for the same pronunciation in the robots.

There is a tone sweep in clip 4 *[1D-KL]*. But I don't know whether I should take that into account

I think the 3rd *[3DS-DWM]* sounded more artificial and therefore less like the eer sound than the second *[2DD-DWM]*. The 4th *[1D-KL]* sounded like it had been put through a flanger hence low realism and lower similarity rating

Diphthong /eə/

Clips 1-3 *[1D-KL, 3DS-DWM, 2DD-DWM]* sound like they have aliased components? When the tone drops in the diphthong, I can hear a noise component moving in the other direction.

most convincing vowels so far

1, 3 and 2 *[1D-KL], [2DD-DWM], [3DS-DWM]* (least to most) have high frequency hiss that increases in pitch

Strange phaser/breathiness to clip 1 *[1D-KL]* makes it seem less human.

I think there's something about the tone that makes the middle one sound a bit more human. Doesn't seem to be how close it is to the proper sound that does it, or at least not for me.

clip 1 *[1D-KL]* and 3 *[2DD-DWM]* sound 'metallic' - perhaps an abundance of high frequency information in the source sound?

Diphthong /ʊə/

Sounds like synth'd vowels are too neutral (eur..) than rounded? (oor)

Synthesised examples sound more like 'eurgh'

all except 4 *[recording]* sound like eurgh

The robot ones just sound disappointed, not like they're saying 'boar'

to me the first three *[2DD-DWM], [1D-KL], [3DS-DWM]* sound more like errr than orrr

apart from recorded sample, vowels seem closer to 'uhr' than 'oar'.

Diphthong /əʊ/

High frequency components again. Synth'd vowels sound more like "err"

slightly confused as I don't think I know the word beau, but asumming from listening that it is the same phoneme as in 'low'

only recorded sample seems to end in /u/.

Diphthong /aʊ/

Strange high frequency components

Clip 2, 3 and 4 *[2DD-DWM], [1D-KL], [3DS-DWM]* sound like a different accent. Maybe Scottish or Irish?

Sounded more like 'Oh', or something.

B.2 Second Pilot

This test compared 3DD-DWM simulations with different sampling frequencies (384 kHz and 960 kHz) to recordings, for both monophthongs and diphthongs, using a paired comparison methodology. Participants were given the opportunity to comment at the end of the test after all comparisons had been made.

Test-final comments

Sometime difficult to tell the difference between some of the synthesised speech examples.

The similar ones were very similar indeed, and seemed to differ in the centre frequency of the high frequency (4 kHz-ish?) windy/squeeky resonance. Thanks - interesting!

Very difficult for the 'very different' sounding samples.

lots obvious; a few no perceptible diff between the 2 tests; only two or three synthetic and different and more tricky to make choice.

Very interesting. Sometimes only way to tell difference is very subtle pitch shifts, and I chose the one closest to the reference. Also there were some with some bad high frequency artefacts and I chose the ones that sounded smoother relative to the reference voice.

about 70% were really obvious!

Some of the answers, although very different to the example, were seemingly identical to each other and were harder to discern differences in.

I couldn't hear the difference between some of them. In a couple that were very similar the one I chose changed depending on what I listened to (e.g. the frequency of the noise or the quality of the vowel)

Some of the options were very similar (I couldn't tell the difference), I but very different from the example sound.

B.3 Final Test

The final test compared diphthongs synthesised using the 2DD-DWM and 3DD-DWM to recordings, without a reference recording for comparison, using the MUSHRA methodology. Additionally, comparisons were made between 2DD-DWM, 3DD-DWM and 3DD-FEM simulations for the diphthong /ɑi/ with four different cutoff frequencies; these simulations were shorter in duration than those compared to a recording. Participants were permitted

to comment on each individual diphthong comparison and to provide general comments at the end of the test.

Comments during example question

They were very fast so not a great deal of information could be gathered. All three *[2DD-DWM, 3DD-DWM, 3DD-FEM]* did have a mechanical quality to them though which gave them away slightly.

Diphthong /eɪ/

I can't put my finger on what, but the third example had something about it that made me doubt that it was human. Neither of the other two examples have the twang at the end of the diphthong to make it sound like a human, but the first one was very close.

Diphthong /aɪ/

2nd sample sounds very close to human voice

After hearing the human sound first, the following two sounds definitely sounded more robotic than before.

Diphthong /ɔɪ/

The first example was closer in pitch, but the movement of the diphthong wasn't quite there. The third example had the diphthong maybe starting on a slightly incorrect vowel sound, which made it sound synthesised.

Diphthong /ɪə/

The word ear naturally sounds higher in pitch even if it isn't when compared to other words and so the second one sounds more real than the first.

Diphthong /eə/

Phasey sounding resonance on the lowest ranked sample *[2DD-DWM]*.

The second one sounded better synthesised, but the third example had the depth that the second one was missing.

Diphthong /ʊə/

Last two have a sharpness to them that makes it unnatural.

The first one sounded like it had quite a bit of white noise in it which detracted from the otherwise well constructed diphthong

Diphthong /əʊ/

Until hearing the second example, I initially thought that the first example was human. After hearing the second example, I thought it sounded synthesised.

Diphthong /aʊ/

The second example was very well synthesised, it just lacked a little depth in the sound.

Diphthong /ɑi/, 16 kHz cut-off

After hearing all the other examples, it has made me much more aware of a computer sound and these are able to be perceived as human sounds, but it is obvious that they're not.

Diphthong /ɑi/, 16 kHz cut-off

[no comments given]

Diphthong /ɑi/, 8 kHz cut-off

Same as the previous three examples. They can be perceived as sounds, but it is obvious that it isn't a human making them.

Diphthong /ɑi/, 5 kHz cut-off

since this is the first set of 3, not sure where to position them absolutely on the scale cos nothing else to compare to (yet)

These are starting to sound more human than the past few examples due to losing the mechanical feeling behind the sound.

Test-final comments

Some of them I could clearly pick out a human reference, some of them much harder to distinguish. Big revealing issues for me were the sound of resonances, almost like going overboard on a finite Q-factor boost at certain frequencies.

Lack of high frequencies in some examples contributes to them sounding less natural. Also some that felt less natural would have comb filter-y / aliasing effects udible.

The most natural ones could well have been completely real, however the reason I gave no sound 100% natural is that maybe it could have been compression technique / low bit rate of recording.

Only timbrel differences effected the most natural sounding ones

Where there was a human reference I tended to compare the naturalness of the others with this. When there wasn't, and the target dipthong was less obvious, I tended to compare each one with itself, i.e. using a more 'absolute' scale of naturalness.

I found this really interesting and hopefully it will help me to make my synthesis better in my project because it seems like the natural sound can be achieved.

I tended to base them on each other in terms of levels of naturalness. So I would listen o them all and increase/decrease the scale based on the most natural one. This was not completely intentional but I noticed I was doing it. It was hard to define the naturalness of them, unless there were obvious human recordings playing.

some of them sounded like they might be natural if they were a muffly recording, but not like natural speech you'd hear in person

With the very short samples it is harder to judge

Appendix C

List of Acronyms

AAC	augmentative and alternative communication
ANN	artificial neural network
ASR	automatic speech recognition
BCI	brain-computer interface
BEM	boundary element method
CFL	Courant-Friedrichs-Lewy
CSA	cross-sectional area
CT	computed tomography
DNN	deep neural network
DWM	digital waveguide mesh
EMA	electromagnetic articulography
FDTD	finite-difference time-domain
FEM	finite element method
GPU	graphics processing unit
HCI	human-computer interaction
HMM	hidden Markov model
IPA	International Phonetic Association
KL	Kelly-Lochbaum
LF	Liljencrants-Fant
LPC	linear predictive coding
LRW	locally reacting wall

LTI	linear time invariant
MFCC	mel-frequency cepstral coefficient
MOS	mean opinion score
MRI	magnetic resonance imaging
MUSHRA	multiple stimulus with hidden reference and anchors
PSD	power spectral density
SPSS	statistical parametric speech synthesis
TLM	transmission line matrix
TTS	text-to-speech
VOCA	voice output communication aid
VTTF	vocal tract transfer function

Appendix D

List of Symbols

A_k	cross-sectional area of tube section k
A_x	cross-sectional area of a duct at position x
c_{medium}	speed of sound in the propagation medium
D	number of dimensions in system
Δx	size of discrete step in x dimension
Δy	size of discrete step in y dimension
Δz	size of discrete step in z dimension
Δt	size of discrete step in continuous time, t
f	frequency variable in continuous frequency
f_s	sampling frequency
G	normalised admittance parameter
G_k	normalised admittance parameter between junction k and $k+1$
$\mathbf{h}_\tau$	impulse response of 2D DWM at frame τ
$H(\omega)$	vocal tract transfer function
$\mathcal{H}(x)$	nonlinear function represented by a DNN
J	scattering junction
J_{input}	set of input junctions in a DWM model
J_{nei}	scattering junction neighbouring junction J
J_{output}	set of output junctions in a DWM model
k	discrete index in first spatial dimension (x)
K	mesh size (number of scattering junctions) in x dimension

l	discrete index in second spatial dimension (y)
L	mesh size (number of scattering junctions) in y dimension
$\mathbf{l}_\tau$	linguistic feature vector at frame τ
λ	Courant number
m	discrete index in third spatial dimension (z)
M	mesh size (number of scattering junctions) in z dimension
μ	viscosity coefficient
n	time index in discrete time
N	number of waveguides connected at each DWM scattering junction
p	local acoustic pressure
p_k^+	right-going travelling wave component in tube section k in 1D DWM model
p_k^-	left-going travelling wave component in tube section k in 1D DWM model
$p^+_{J,J_{nei}}(n)$	pressure incident at junction J from neighbouring junction J_{nei} at time step n in multi-dimensional DWM model
$p^-_{J,J_{nei}}(n)$	pressure output by junction J towards neighbouring junction J_{nei} at time step n in multi-dimensional DWM model
$p_B(n)$	acoustic pressure at DWM boundary junction B at time step n
$p_J(n)$	acoustic pressure at DWM scattering junction J at time step n
p_m^n	instantaneous acoustic pressure at time step $t = n\Delta t$ and spatial location $x = k\Delta x$ for a system with one spatial dimension (FDTD notation)
$p^n_{k,l,m}$	instantaneous acoustic pressure at time step $t = n\Delta t$ and spatial location $x = k\Delta x$, $y = l\Delta y$, $z = m\Delta z$ for a system with three spatial dimensions (FDTD notation)
R	reflection coefficient
Re	Reynolds number
$r_{glottis}$	reflection coefficient at glottis end of 2D DWM domain
r_{lips}	reflection coefficient at lip end of 2D DWM domain
r_{sides}	reflection coefficient at sides of 2D DWM domain
r_x	radius of a duct at position x

ρ_{medium}	density of the propagation medium
$\mathbf{s}_\tau$	windowed speech frame at frame τ
$\mathbf{\Sigma}_{Y_\tau}$	covariance matrix based on $\mathbf{h}_\tau$
t	time variable in continuous time
τ	frame number for DNN-DWM simulations
U	volume velocity
v	particle velocity
x	location in first spatial dimension
y	location in second spatial dimension
Y	acoustic admittance
$\mathbf{Y}$	matrix of admittance values for a multi-dimensional DWM model
Y_i	admittance of waveguide connecting scattering junctions J and J_{nei}
Y_{medium}	specific acoustic admittance of propagation medium
$\mathbf{Y}_\tau$	DNN-generated admittance map for frame τ
z	location in third spatial dimension
Z	acoustic impedance
Z_{medium}	specific acoustic impedance of propagation medium
Z_x	characteristic acoustic impedance of a duct at position x

References

[1] W. T. Fitch, "The evolution of speech: a comparative review," *Trends Cogn. Sci.*, vol. 4, no. 7, pp. 258–267, Jul. 2000.

[2] N. T. Uomini and G. F. Meyer, "Shared brain lateralization patterns in language and Acheulean stone tool production: a functional transcranial doppler ultrasound study," *PLoS One*, vol. 8, no. 8, Aug. 2013, e72693.

[3] O. Sacks, *Seeing Voices: A Journey Into the World of the Deaf*, University of California Press, Berkeley and Los Angeles, CA, 1989.

[4] R. Pieraccini, *The Voice in the Machine: Building Computers That Understand Speech*, MIT Press, Cambridge, MA, 2012.

[5] B. Bryson, *Mother Tongue: The Story of the English Language*, Penguin Books, London, UK, 1990.

[6] J. L. Flanagan, "Voices of men and machines," *J. Acoust. Soc. Am.*, vol. 51, no. 5, pp. 1375–1387, May. 1972.

[7] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.*, vol. 82, no. 3, pp. 737–793, Sep. 1987.

[8] S. J. Winters and D. B. Pisoni, "Perception and comprehension of speech synthesis," in *Encylopedia of Language and Linguistics, 2nd Edition*, K. Brown, Ed., pp. 31–49. Elsevier Science, Ltd., Oxford, UK, 2006.

[9] L. Blomert and H. Mitterer, "The fragile nature of the speech-perception deficit in dyslexia: natural vs. synthetic speech," *Brain Lang.*, vol. 89, no. 1, pp. 21–26, Apr. 2004.

[10] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[11] M. Arnela, O. Guasch, S. Dabbaghchian, and O. Engwall, "Finite element generation of vowel sounds using dynamic complex three-dimensional vocal tracts," in *Proc. 23rd Int. Congr. Sound Vib.*, Athens, Greece, July 2016.

[12] K. N. Stevens, *Acoustic Phonetics*, The MIT Press, Cambridge, MA, 1998.

[13] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics, 4th Ed.*, John Wiley & Sons, Inc., New York, NY, 2000.

[14] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics, 4th Ed.*, John Wiley & Sons, Inc., New York, NY, 2000.

[15] S. Bilbao, *Numerical Sound Synthesis: Finite Difference Schemes and Simulation in Musical Acoustics*, John Wiley and Sons, Ltd., West Sussex, UK, 2009.

[16] J. O. Smith, *Physical Audio Signal Processing for Virtual Musical Instruments and Digital Audio Effects*, W3K Publishing, `http://books.w3k.org`, 2012.

[17] M. Kleiner, *Electroacoustics*, Taylor & Francis Group, LLC, Boca Raton, FL, 2013.

[18] J. L. Flanagan, *Speech Analysis Synthesis and Perception, 2nd Ed.*, Springer-Verlag, Berlin, Germany, 1972.

[19] K. Johnson, *Acoustic and Auditory Phonetics, 3rd Ed.*, John Wiley and Sons, Ltd., West Sussex, UK, 2012.

[20] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *J. Acoust. Soc. Am.*, vol. 100, no. 1, pp. 537–554, July 1996.

[21] M. Arnela, O. Guasch, and F. Alías, "Effects of head geometry simplifications on acoustic radiation of vowel sounds based on time-domain finite-element simulations," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 2946–2954, Oct. 2013.

[22] P. Mermelstein, "Articulatory model for the study of speech production," *J. Acoust. Soc. Am.*, vol. 53, no. 4, pp. 1070–1082, 1973.

[23] G. Fant, *Acoustic Theory of Speech Production*, Mouton & Co. N.V., The Hague, The Netherlands, 1960.

[24] P. Rubin, T. Baer, and P. Mermelstein, "An articulatory synthesizer for perceptual research," *J. Acoust. Soc. Am.*, vol. 70, no. 2, pp. 321–328, Aug. 1981.

[25] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," in *Proc. 4th Int. Congr. Acoust.*, Copenhagen, Denmark, Aug. 1962, IAC, pp. 1–4.

[26] T. Kitamura, H. Takemoto, H. Makinae, T. Yamaguchi, and K. Maki, "Acoustic analysis of detailed three-dimensional shape of the human nasal cavity and paranasal sinuses," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 3472–3476.

[27] K. M. Dawson, M. K. Tiede, and D. H. Whalen, "Methods for quantifying tongue shape and complexity using ultrasound tmaging," *Clin. Linguist. Phon.*, vol. 30, no. 3, pp. 328–344, 2016.

[28] D. W. McRobbie, E. A. Moore, M. J. Graves, and M. R. Prince, *MRI from Picture to Proton*, Cambridge University Press, Cambridge, UK, 2003.

[29] H. Takemoto, T. Kitamura, H. Nishimoto, and K. Honda, "A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions," *Acoust. Sci. Tech.*, vol. 25, no. 6, pp. 468–474, 2004.

[30] H. Takemoto, P. Mokhtari, and T. Kitamura, "Acoustic analysis of the vocal tract during vowel production by finite-different time-domain method," *J. Acoust. Soc. Am*, vol. 128, no. 6, pp. 3724–3738, Dec. 2010.

[31] W. J. Hardcastle and F. Gibbon, "Electropalatography and its clincal applications," in *Instrumental Clinical Phonetics*, M. J. Ball and C. Code, Eds., pp. 149–193. Whurr Publishers Ltd., London, UK, 1997.

[32] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "A silent speech system based on permanent magnet articulography and direct synthesis," *Comput. Speech Lang.*, vol. 39, no. C, pp. 67–87, Sept. 2016.

[33] A. Rugchatjaroen and D. M. Howard, "Flexibility of cosine impedance function in 2-D digital waveguide mesh for plosive synthesis," in *Proc. ChinaSIP*, Sapporo, Japan, July 2014.

[34] P. W. Schonle, K. Grabe, P. Wenig, J. Hohne, J. Schrader, and B. Conrad, "Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, vol. 31, no. 1, pp. 26–35, May 1987.

[35] S. D'Amario and H. Daffern, "Using electrolayngography and electroglottography to assess the singing voice: a systematic review," *Psychomusicology: Music, Mind and Brain*, June 2017.

[36] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Quarterly Progress and Status Report, Dept. Speech Music Hearing, KTH (STL-QPSR)*, vol. 26, no. 4, pp. 1–13, 1985.

[37] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.*, vol. 49, no. 2, pp. 583–590, Feb. 1971.

[38] Birkholz. P., D. Jackèl, and B. J. Kröger, "Simulation of losses due to turbulence in the time-varying vocal system," *IEEE Trans. Audio Speech Language Process.*, vol. 15, no. 4, pp. 1218–1226, May. 2007.

[39] J. Clark and C. Yallop, *An Introduction to Phonetics and Phonology*, Basil Blackwell Ltd., Oxford, UK, 1990.

[40] T. Kitamura, K. Honda, and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoust. Sci. Tech.*, vol. 26, no. 1, pp. 16–26, 2005.

[41] J. Liljencrants, *Speech Synthesis with a Reflection-Type Line Analog*, Ph.D. thesis, KTH, Stockholm, Sweden, 1985.

[42] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.*, vol. 24, no. 2, pp. 175–184, Mar. 1952.

[43] International Phonetic Association, "Full IPA chart," `www.internationalphoneticassociation.org/content/full-ipa-chart`, 2015, [Accessed: Jul. 31, 2017].

[44] D. Jones, *An outline of English phonetics, 9th Ed.*, W. Heffer and Sons, Cambridge, UK, 1960.

[45] O. Guasch, M. Arnela, R. Codina, and H. Espinoza, "A stabilized finite element method for the mixed wave equation in an ALE framework with application to diphthong production," *Acta Acust. united Ac.*, vol. 102, no. 1, pp. 94–106, Jan. 2016.

[46] B. H. Story and K. Bunton, "An acoustically-driven vocal tract model for stop consonant production," *Speech Communication*, vol. 87, pp. 1–17, Mar. 2017.

[47] B. V. Tucker and M. Ernestus, "Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon," *Mental Lexicon*, vol. 11, no. 3, pp. 375–400, 2016.

[48] W. J. Hardcastle and N. Hewlett, *Coarticulation: Theory, Data and Techniques*, Cambridge University Press, Cambridge, UK, 1999.

[49] Alberta Phonetics Laboratory, "Reduction examples," `aphl.artsrn.ualberta.ca/?page_id=323`, [Accessed: May 29, 2017].

[50] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[51] N. Campbell, "Evaluation of speech synthesis," in *Evaluation of Text and Speech Systems*, L. Dybkjær, H. Hemsen, and W. Minker, Eds., pp. 29–64. Springer, Dordrecht, The Netherlands, 2007.

[52] A. W. Black and K. A. Lenzo, "Building synthetic voices," `http://festvox.org/bsv/`, 2014, [Accessed: Jun. 17, 2016].

[53] S. King, "A beginners' guide to statistical parametric speech synthesis [online]," `http://www.cstr.ed.ac.uk/downloads/publications/2010/king_hmm_tutorial.pdf`, 2010, [Accessed: Jun. 30, 2017].

[54] A. Parlikar and A. W. Black, "Data-driven phrasing for speech synthesis in low-resource languages," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 4013–4016.

[55] CSTR University of Edinburgh, "The Festival speech synthesis system," `www.cstr.ed.ac.uk/projects/festival/`, 2014, [Accessed: Aug. 25, 2014].

[56] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, Springer-Verlag, Berlin, Germany, 1976.

[57] J. Beskow, "Formant synthesis demo," `www.speech.kth.se/wavesurfer/formant`, 2001, [Accessed: Jun. 30, 2017].

[58] M. Schroder, "Emotional speech synthesis: A review," in *Proc. 7th European Conference on Speech Communication and Technology (EUROSPEECH)*, Aalborg, Denmark, Sept. 2001, pp. 561–564.

[59] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[60] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Atlanta, Georgia, May 1996, pp. 373–376.

[61] Y. O. Kong L. Latacz and W. Verhelst, "Unit selection synthesis using long non-uniform units and phonemic identity matching," in *Proc. ISCA Workshop Speech Synth.*, Bonn, Germany, Aug. 2007, pp. 270–275.

[62] A. W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Proc. EUROSPEECH*, Rhodes, Greece, 1997.

[63] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, Vancouver, Canada, May 2013, pp. 7962–7966.

[64] D. Jurafsky and J. H. Martin, *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall PTR, Upper Saddle River, NJ, 2009.

[65] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[66] T. Yoshimura, K. Tokuda, T. Masuko, T. kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH*, Budapest, Hungary, Sept. 1999, pp. 2347–2350.

[67] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction," *Acoust. Sci. Tech.*, vol. 33, no. 1, pp. 1–5, Jan. 2012.

[68] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, Y. Guan, R. Hu, K. Oura, Y-J. Wu, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis—analysis and application of TTS systems built on various ASR corpora," *IEEE Trans. Audio Speech Language Process.*, vol. 18, no. 5, pp. 984–1004, July 2010.

[69] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Communication*, vol. 52, no. 5, pp. 394–404, May 2010.

[70] K. Oura, J. Yamagishi, M. Wester, S. King, and K. Tokuda, "Analysis of unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using KLD-based transform mapping," *Speech Communication*, vol. 54, no. 6, pp. 703–714, July 2012.

[71] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, and V. Strom, "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis," *Speech Communication*, vol. 52, no. 2, pp. 164–179, Feb. 2010.

[72] K. Oura, A. Mase, T. Yamada, S. Muto, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Recent development of the HMM-based singing voice synthesis system—Sinsy," in *Proc. 7th ISCA Speech Synthesis Workshop*, Kyoto, Japan, Sept. 2010, pp. 211–216.

[73] J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "HMM-based speech synthesiser using the LF-model of the glottal source," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Prague, Czech Republic, May 2011, pp. 4704–4707.

[74] T. Toda and S. Young, "Trajectory training considering global variance for HMM-based speech synthesis," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 4025–4028.

[75] O. Watts, G. E. Henter, T. Merritt, Z. Wu, and S. King, "From HMMs to DNNs: where do the improvements come from?," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Shanghai, China, Mar. 2016, pp. 5505–5509.

[76] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Trajectory training considering global variance for speech synthesis based on neural networks," in *Proc. Int. Conf. Acoust. Speech and Signal Process.*, Shanghai, China, Mar. 2016, pp. 5600–5604.

[77] K. Tokuda and H. Zen, "Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis," in *Proc. Int. Conf. Acoustics Speech and Signal Process.*, Brisbane, Australia, Apr. 2015, pp. 4215–4219.

[78] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vanyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: a generative model for raw audio," *arXiv:1609.03499v2*, Sept. 2016.

[79] B. B. Monson, E. J. Hunter, A. J. Lotto, and B. H. Story, "The perceptual significance of high-frequency energy in the human voice," *Front. Psychol.*, vol. 5, no. 587, pp. 1–11, June 2014.

[80] C. H. Shadle and R. I. Damper, "Prospects for articulatory synthesis: a position paper," in *Proc. 4th ISCS Workshop Speech Synthesis*, Blair Atholl, Scotland, 2001, pp. 121–126.

[81] B. J. Kroger and P. Birkholz, "Articulatory synthesis of speech and singing: state of the art and suggestions for future research," in *Multimodal Signals*, A. Esposito et al., Ed., pp. 306–319. Springer-Verlag, Berlin, Germany, 2009.

[82] J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," *Bell System Tech. J.*, vol. 54, no. 3, pp. 485–506, Mar. 1975.

[83] C. H. Coker, "A model of articulatory dynamics and control," *Proc. IEEE*, vol. 64, no. 4, pp. 452–460, Apr. 1976.

[84] B. H. Story, "A parametric model of the vocal tract area function for vowel and consonant simulation," *J. Acoust. Soc. Am.*, vol. 117, no. 5, pp. 3231–3254, May 2005.

[85] P. Birkholz, B. J. Kroger, and C. Neuschaefer-Rube, "Articulatory synthesis and perception of plosive-vowel syllables with virtual consonant targets," in *Proc. INTERSPEECH*, Makuhari, Japan, Sept. 2010, pp. 1017–1020.

[86] B. J. Kroger P. Birkholz and C. Neuschaefer-Rube, "Model-based reproduction of articulatory trajectories for consonant-vowel sequences," *IEEE Trans. Audio Speech Language Process.*, vol. 19, no. 5, pp. 1422–1433, July 2011.

[87] P. Birkholz, L. Martin, Y. Xu, S. Scherbaum, and C. Neuschaefer-Rube, "Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis," *Comput. Speech Lang.*, vol. 41, pp. 116–127, Jan. 2017.

[88] T. Toda, A. W. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis," in *Proc. 5th ISCA Speech Synth. Workshop*, Pittsburgh, PA, June 2004, pp. 31–36.

[89] Z.-H. Ling, K. Richmond, and J. Yamagishi, "Articulatory control of HMM-based parametric speech sysnthesis using feature-space-switched multiple regression," *IEEE Trans. Audio Speech Language Process.*, vol. 21, no. 1, pp. 205–217, Jan. 2013.

[90] S. Aryal and R. Gutierrez-Osuna, "Data-driven articulatory synthesis with deep neural networks," *Comput. Speech Lang.*, vol. 36, pp. 260–273, Mar. 2016.

[91] A. J. Gully, T. Yoshimura, D. T. Murphy, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Articulatory text-to-speech synthesis using the digital waveguide mesh driven by a deep neural network," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 234–238.

[92] Y. Xu, A. Lee, W.-L. Wu, X. Liu, and P. Birkholz, "Human vocal attractiveness as signaled by body size projection," *PLoS One*, vol. 8, no. 4, Apr. 2013, e62397.

[93] J. D. Leongomeza, V. R. Mileva, A. C. Littleb, and S. C. Roberts, "Perceived differences in social status between speaker and listener affect the speaker's vocal characteristics," *PLoS One*, vol. 12, no. 6, June 2017, e0179407.

[94] S. Creer, P. Enderby, S. Judge, and A. John, "Prevalence of people who could benefit from augmentative and alternative communication (AAC) in the UK: determining the need," *Int. J. Lang. Commun. Disord.*, vol. 51, no. 6, pp. 639–653, Nov. 2016.

[95] S. E. Stern, "Computer-synthesized speech and perceptions of the social influence of disabled users," *J. Lang. Soc. Psychol.*, vol. 27, no. 3, pp. 254–265, Sep. 2008.

[96] H. T. Bunnell and R. Patel, "VocaliD: Personal voices for augmented communicators," *J. Acoust. Soc. Am.*, vol. 135, no. 4, pp. 2390, Apr. 2014.

[97] Anne Rowling Regenerative Neurology Clinic, "The Speak:Unique voicebank research project," `http://annerowlingclinic.com/Speak_Unique_research.html`, 2017, [Accessed: Jun. 30, 2017].

[98] I. Frodsham, "Yorkshireman battling motor neurone disease will keep his accent even when he can no

longer speak as experts design him a robotic voice," `http://www.dailymail.co.uk/news/article-4192846/Yorkshireman-battling-MND-keeps-accent-t-speak.htm`, 2017, [Accessed: Jun. 30, 2017].

[99] H. Cryer and S. Home, "User attitudes towards synthetic speech for talking books," *RNIB Centre for Accessible Information*, May 2009.

[100] Y-P. P. Chen, C. Johnson, P. Lalbakhsh, T. Caelli, G. Deng, D. Tay, S. Erickson, P. Broadbridge, A. El Refaie, W. Doube, and M. E. Morris, "Systematic review of virtual speech therapists for speech disorders," *Comput. Speech Lang.*, vol. 37, pp. 98–128, May 2016.

[101] Y. Ohkawa, M. Suzuki, H. Ogasawara, A. Ito, and S. Makino, "A speaker adaptation method for non-native speech using learners' native utterances for computer-assisted language learning systems," *Speech Communication*, vol. 51, no. 10, pp. 875–882, Oct. 2009.

[102] T. Arai, "Vocal-tract models and their applications in education for intuitive understanding of speech production," *Acoust. Sci. Tech.*, vol. 37, no. 4, pp. 148–156, 2016.

[103] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," *Comput. Speech Lang.*, vol. 19, no. 1, pp. 55–83, Jan. 2005.

[104] S. Lattner, B. Maess, Y. Wang, M. Schauer, K. Alter, and A. D. Friederici, "Dissociation of human and computer voices in the brain: evidence for a gestalt-like perception," *Human Brain Mapping*, vol. 20, no. 1, pp. 13–21, Sept. 2003.

[105] C. Mayo, R. A. J. Clark, and S. King, "Listeners' weighting of acoustic cues to synthetic speech naturalness: a multidimensional scaling analysis," *Speech Communication*, vol. 53, no. 3, pp. 311–326, Mar. 2011.

[106] C. R. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, "Quality prediction of synthesized speech based on perceptual quality dimensions," *Speech Communication*, vol. 66, pp. 17–35, Feb. 2015.

[107] D.-Y. Huang, "Prediction of perceived sound quality of synthetic speech," in *Proc. APSIPA Annu. Summit and Conf.*, Xi'an, China, 2011.

[108] ITU-T, "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," P.862, 2001.

[109] H. C. Nusbaum, A. L. Francis, and A. S. Henly, "Measuring the naturalness of synthetic speech," *Int. J. Speech Technology*, vol. 2, no. 1, pp. 7–19, May 1995.

[110] S. Hawkins, S. Heid, J. House, and M. Huckvale, "Assessment of naturalness in the protosynth speech synthesis project," in *IEEE Colloq. Speech Synthesis*, London, UK, May 2000.

[111] D. B. Pisoni, "Perception of synthetic speech," in *Progress in Speech Synthesis*, J. P. H. van Santen, R. Sproat, J. Olive, and J. Hirschberg, Eds., pp. 541–560. Springer-Verlag, New York, NY, 1997.

[112] C. R. Paris, M. H. Thomas, R. D. Gilson, and J. P. Kincaid, "Linguistic cues and memory for synthetic and natural speech," *Human Factors*, vol. 42, no. 3, pp. 421–431, Fall 2000.

[113] P. M. Evitts and J. Searl, "Reaction times of normal listeners to laryngeal, alaryngeal and synthetic speech," *J. Speech Language and Hearing Research*, vol. 49, no. 6, pp. 1380–1390, Dec. 2006.

[114] X. Wang, Z.-H. Ling, and L.-R. Dai, "Concept-to-speech generation with knowledge sharing for acoustic modeling and utterance filtering," *Comput. Speech Lang.*, vol. 38, pp. 46–67, July 2016.

[115] P. A. Taylor, "Concept-to-speech synthesis by phonological structure matching," *Phil. Trans. R. Soc. Lond. A*, vol. 358, no. 1769, pp. 1403–1417, Apr. 2000.

[116] J. S. Brumberg and F. H. Guenther, "Development of speech prostheses: current status and recent advances," *Expert Rev. Med. Devices*, vol. 7, no. 5, pp. 667–679, Sept. 2010.

[117] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, "Real-time control of an articulatory-based speech synthesiser for brain-computer interfaces," *PLoS Computational Biology*, vol. 12, no. 11, Nov. 2016, e1005119.

[118] M. S. Hawley, S. P. Cunningham, P. D. Green, P. Enderby, R. Palmer, S. Sehgal, and P. O'Neill, "A voice-input voice-output communication aid for people with severe speech impairment," *IEEE Trans. Neural Syst. and Rehab. Eng.*, vol. 21, no. 1, pp. 23–31, Jan. 2013.

[119] T. Arai, "Mechanical vocal-tract models for speech dynamics," in *Proc. INTERSPEECH*, Makuhari, Japan, Sep. 2010, pp. 1025–1028.

[120] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Acoust. Speech and Signal Process.*, vol. 35, no. 7, pp. 955–967, July 1987.

[121] K. Yee, "Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media," *IEEE Trans. Antennas and Propagation*, vol. 14, no. 3, pp. 302–307, May 1966.

[122] K. R. Kelly, R. W. Ward, S. Treitel, and R. M. Alford, "Synthetic seismograms: a finite-difference approach," *Geophysics*, vol. 41, no. 1, pp. 2–27, 1976.

[123] S. Bilbao, *Numerical Sound Synthesis: Finite Difference Schemes and Simulation in Musical Acoustics*, John Wiley and Sons, Ltd., West Sussex, UK, 2009.

[124] A. Southern, T. Lokki, and L. Savioja, "The perceptual effects of dispersion error on room acoustic model auralization," in *Proc. Forum Acusticum*, Aalborg, Denmark, 2011, pp. 1553–1558.

[125] J. O. Smith, "Physical modeling using digital waveguides," *Comput. Music J.*, vol. 16, no. 4, pp. 74–91, 1992.

[126] D. Murphy, A. Kelloniemi, J. Mullen, and S. Shelley, "Acoustic modeling using the digital waveguide mesh," *IEEE Signal Process. Mag.*, vol. 24, no. 2, pp. 55–66, Mar. 2007.

[127] P. Chobeau, *Modeling of sound propagation in forests using the transmission line matrix method*, Ph.D. thesis, Université du Maine, Le Mans, France, 2014.

[128] Y. Lam, "The boundary element method," *Lecture Notes*, 2013.

[129] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLoS One*, vol. 8, no. 4, Apr. 2013, e60603.

[130] H. K. Dunn, "The calculation of vowel resonances, and an electrical vocal tract," *J. Acoust. Soc. Am*, vol. 22, no. 6, pp. 740–753, Nov. 1950.

[131] K. N. Stevens, S. Kasowski, and C. G. M. Fant, "An electrical analog of the vocal tract," *J. Acoust. Soc. Am*, vol. 25, no. 4, pp. 734–742, July 1953.

[132] M. H. L. Hecker, "Studies of nasal consonants with an articulatory speech synthesizer," *J. Acoust. Soc. Am*, vol. 34, no. 2, pp. 179–187, Feb. 1962.

[133] J. Flanagan and L. Landgraf, "Self-oscillating source for vocal-tract synthesizers," *IEEE Trans. Audio and Electroacoust.*, vol. 16, no. 1, pp. 57–64, Mar. 1968.

[134] C. H. Coker, "A model of articulatory dynamics and control," *Proc. IEEE*, vol. 64, no. 4, pp. 452–460, Apr. 1976.

[135] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Communication*, vol. 1, no. 3–4, pp. 199–229, dec 1982.

[136] P. Birkholz and D. Jackèl, "Influence of temporal discretization schemes on formant frequencies and bandwidths in time-domain simulations of the vocal tract system," in *Proc. INTERSPEECH*, Jeju Island, Korea, 2004, pp. 1125–1128.

[137] P. Birkholz, D. Jackèl, and B. J. Kroger, "Simulation of losses due to turbulence in the time-varying vocal system," *IEEE Trans. Audio Speech and Language Process.*, vol. 15, no. 4, pp. 1218–1226, May 2007.

[138] P. Birkholz, "VocalTractLab," `http://www.vocaltractlab.de/`, 2016, [Accessed: Mar. 26, 2016].

[139] P. Cook, *Identification of Control Parameters in an Articulatory Vocal Tract Model, with Applications to the Synthesis of Singing*, Ph.D. thesis, Stanford University, Stanford, CA, 1991.

[140] V. Valimaki and M. Karjalainen, "Improving the Kelly-Lochbaum vocal tract model using conical tube sections and fractional delay filtering techniques," in *Proc. Int. Conf. Spoken Lang. Process.*, Yokohama, Japan, Sep. 1994, pp. 615–618.

[141] B. H. Story, *Physiologically-Based Speech Simulation using an Enhanced Wave-Reflection Model of the Vocal Tract*, Ph.D. thesis, University of Iowa, Iowa City, IA, 1995.

[142] H. Matsuzaki, K. Motoki, and N. Miki, "Computation of the acoustic characteristics of simplified vocal-tract models by 3-D finite element method," in *Proc. Int. Symp. Commun. and Inf. Tech.*, Sapporo, Japan, Oct. 2004, pp. 894–899.

[143] M. Arnela, R. Blandin, S. Dabbaghchian, O. Guasch, F. Alías, X. Pelorson, A. Van Hirtum, and O. Engwall, "Influence of lips on the production of vowels based on finite element simulations and experiments," *J. Acoust. Soc. Am.*, vol. 139, no. 5, pp. 2852–2859, May 2016.

[144] M. Arnela, S. Dabbaghchian, R. Blandin, O. Guasch, O. Engwall, A. Van Hirtum, and X. Pelorson, "Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds," *J. Acoust. Soc. Am.*, vol. 140, no. 3, pp. 1707–1718, Sep. 2016.

[145] T. J. Thomas, "A finite element model of fluid flow in the vocal tract," *Comput. Speech Lang.*, vol. 1, no. 2, pp. 131–151, Dec. 1986.

[146] J. Mullen, D. M. Howard, and D. T. Murphy, "Waveguide physical modelling of vocal tract acoustics: flexible formant bandwidth control from increased model dimensionality," *IEEE Trans. Audio Speech Language Process.*, vol. 14, no. 3, pp. 964–971, May 2006.

[147] J. Mullen, D. M. Howard, and D. T. Murphy, "Real-time dynamic articulations in the 2D waveguide mesh vocal tract model," *IEEE Trans. Audio Speech Language Process.*, vol. 15, no. 2, pp. 577–585, Feb. 2007.

[148] A. Rugchatjaroen and D. M. Howard, "An evaluation of rectilinear digital waveguide mesh in modelling branch tube for English nasal synthesis," *Appl. Acoust.*, vol. 122, pp. 1–7, Dec. 2017.

[149] M. Arnela and O. Guasch, "Two-dimensional vocal tracts with three-dimensional behavior in the numerical generation of vowels," *J. Acoust. Soc. Am.*, vol. 135, no. 1, pp. 369–379, Jan. 2014.

[150] J. Wei, W. Guan, D. Q. Hou, D. Pan, W. Lu, and J. Dang, "A new model for acoustic wave propagation and scattering in the vocal tract," in *Proc. INTERSPEECH*, San Francisco, CA, Sep. 2016.

[151] Y. Kagawa, R. Shimoyama, T. Yamabuchi, T. Murai, and K. Takarada, "Boundary element models of the vocal tract and radiation field and their response characteristics," *J. Sound. Vib.*, vol. 157, no. 3, pp. 385–403, Sept. 1992.

[152] Y. Kagawa, Y. Ohtani, and R. Shimoyama, "Vocal tract shape identification from formant frequency spectra—a simulation using three-

dimensional boundary element models," *J. Sound. Vib.*, vol. 203, no. 4, pp. 581–596, June 1997.

[153] C. Lu, T. Nakai, and H. Suzuki, "Finite element simulation of sound transmission in vocal tract," *J. Acoust. Soc. Jpn. (E)*, vol. 14, no. 2, pp. 63–72, 1993.

[154] H. Matsuzaki, T. Hirohku, , N. Miki, and N. Nagai, "Analysis of acoustic characteristics in the vocal tract with inhomogeneous wall impedance using a three-dimensional FEM model," *Electron. and Commun. in Japan*, vol. 77, no. 10, pp. 27–36, 1994.

[155] S. El-Masri, X. Pelorson, P. Saguet, and P. Badin, "Vocal tract acoustics using the transmission line matrix (TLM) method," in *Proc. Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, Oct. 1996, pp. 953–956.

[156] S. El-Masri, X. Pelorson, P. Saguet, and P. Badin, "Development of the transmission line matrix method in acoustics applications to higher modes in the vocal tract and other complex ducts," *Int. J. Numer. Model.*, vol. 11, no. 3, pp. 133–151, May 1998.

[157] X. Pelorson, P. Badin, P. Saguet, and S. El-Masri, "Numerical simulation of the vocal tract acoustics using the TLM method," in *Proc. Forum Acusticum*, Seville, Spain, Sep. 2002.

[158] H. Matsuzaki, , N. Miki, and Y. Ogawa, "FEM analysis of sound wave propagation in the vocal tract with 3-D radiational model," *J. Acoust. Soc. Jpn. (E)*, vol. 17, no. 3, pp. 163–166, 1996.

[159] H. Matsuzaki, , N. Miki, and Y. Ogawa, "3D finite element analysis of Japanese vowels in elliptic sound tube model," *Electron. and Commun. in Japan*, vol. 83, no. 4, pp. 43–51, 2000.

[160] A. Hannukainen, T. Lukkari, J. Malinen, and P. Palo, "Vowel formants from the wave equation," *J. Acoust. Soc. Am.*, vol. 122, no. 1, pp. EL1–EL7, July 2007.

[161] P. Švancara, J. Horáček, and L. Pešek, "Numerical modelling of effect of tonsillectomy on production of Czech vowels," *Acta Acust. United Acust.*, vol. 92, no. 5, pp. 681–688, Sep. 2006.

[162] T. Vampola, J. Horáček, and J. G. Švec, "FE modeling of human vocal tract acoustics. part I: production of Czech vowels," *Acta Acust. United Acust.*, vol. 94, no. 3, pp. 433–447, May 2008.

[163] Y. Wang, H. Wang, J. Wei, and J. Dang, "Acoustic analysis of the vocal tract from a 3D physiological articulatory model by finite-difference time-domain method," in *Proc. Int. Conf. Automat. Control and Artificial Intell.*, Xiamen, China, Mar. 2012, pp. 329–333.

[164] Y. Wang, H. Wang, J. Wei, and J. Dang, "Mandarin vowel synthesis based on 2D and 3D vocal tract model by finite-difference time-domain method," in *Proc. APSIPA Annu. Summit and Conf.*, Hollywood, CA, Dec. 2012, pp. 1–4.

[165] M. Speed, D. T. Murphy, and D. M. Howard, "Three-dimensional digital waveguide mesh simulation of cylindrical vocal tract analogs," *IEEE Trans. Audio Speech Language Process.*, vol. 21, no. 2, pp. 449–454, Feb. 2013.

[166] M. Speed, D. Murphy, and D. Howard, "Modeling the vocal tract transfer function using a 3D digital waveguide mesh," *IEEE Trans. Audio Speech Language Process.*, vol. 22, no. 2, pp. 453–464, Feb. 2014.

[167] M. D. A. Speed, *Voice Synthesis using the Three-Dimensional Digital Waveguide Mesh*, Ph.D. thesis, University of York, York, UK, 2012.

[168] EUNISON, "Eunison - extensive unified-domain simulation of the human voice," `http://fp7eunison.com/`, 2016, [Accessed: Jul. 31, 2017].

[169] O. Guasch, M. Arnela, A. Pont, J. Baiges, and R. Codina, "Finite elements in vocal tract acoustics: generation of vowels, diphthongs

and sibilants," in *Proc. Acoustics 2015 Hunter Valley*, Hunter Valley, Australia, Nov. 2015.

[170] A. Katsamanis and P. Maragos, "Fricative synthesis investigations using the transmission line matrix method," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 3741, May 2008.

[171] R. Wilhelms-Tricarico, "Physiological modeling of speech production: methods for modeling soft-tissue articulators," *J. Acoust. Soc. Am.*, vol. 97, no. 5, pp. 3085–3098, May 1995.

[172] J. Dang and K. Honda, "Construction and control of a physiological articulatory model," *J. Acoust. Soc. Am.*, vol. 115, no. 2, pp. 853–870, Feb. 2004.

[173] S. Dabbaghchian, M. Arnela, O. Engwall, O. Guasch, I. Stavness, and P. Badin, "Using a biomechanical model and articulatory data for the numerical production of vowels," in *Proc. INTERSPEECH*, San Francisco, CA, Sep. 2016, pp. 3569–3573.

[174] O. Engwall, "Are static MRI measurements representative of dynamic speech? Results from a comparative study using MRI, EPG and EMA," in *Proc. INTERSPEECH*, Beijing, China, Oct. 2000, pp. 17–20.

[175] 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, Nov. 2008.

[176] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. 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, July 2006.

[177] G. R. A. S. Sound & Vibration A/S, "Head and torso simulators," `http://www.gras.dk/products/head-torso-simulators-kemar.html`, 2016, [Accessed: Sep. 7, 2017].

[178] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," `http://www.praat.org/`, 2017, [Accessed: Jul. 31, 2017].

[179] D. Aalto, A. Huhtala, A. Kivela, J. Malinen, P. Palo, J. Saunavaara, and M. Vainio, "How far are vowel formants from computed vocal tract resonances?," *arXiv:1208.5962v2 [math.DS]*, Oct. 2012.

[180] M. Fleischer, S. Pinkert, W. Mattheus, A. Mainka, and D. Murbe, "Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall," *Biomech. Model Mechanobiol.*, vol. 14, no. 4, pp. 719–733, Aug. 2015.

[181] I. R. Titze, "Parameterization of the glottal area, glottal flow, and vocal fold contact area," *J. Acoust. Soc. Am.*, vol. 75, no. 2, pp. 570–580, Feb. 1984.

[182] S. Siegel and N. J. Castellan, *Nonparametric Statistics for the Behavioural Sciences, 2nd Ed.*, McGraw-Hill, New York, NY, 1988.

[183] A. Vargha and H. D. Delaney, "A critique and improvement of the CL common language effect size statistics of McGraw and Wong," *J. Educ. Behav. Stat.*, vol. 25, no. 2, pp. 101–132, Jun. 2000.

[184] ITU-T, "A method for subjective performance assessment of the quality of speech voice output devices," P.85, 1994.

[185] J. Kreiman, B. R. Gerratt, and M. Ito, "When and why listeners disagree in voice quality assessment tasks," *J. Acoust. Soc. Am.*, vol. 122, no. 4, pp. 2354–2364, Oct. 2007.

[186] ITU-R, "Method for the subjective assessment of intermediate quality level of audio systems," BS.1534-3, 2015.

[187] G. E. Henter, T. Merritt, M. Shannon, C. Mayo, and S. King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *Proc. INTERSPEECH*, Singapore, Sept. 2014, pp. 1504–1508.

[188] J. P. H. van Santen, "Perceptual experiments for diagnostic testing of text-to-speech systems," *Comput. Speech Lang.*, vol. 7, no. 1, pp. 49–100, Jan. 1993.

[189] M. Schoeffler, F.-R. Stöter, H. Bayerlein, B. Edler, and J. Herre, "An experiment about estimating the number of instruments in polyphonic music: a comparison between internet and laboratory results," in *Proc. Int. Soc. Music Inform. Retrieval Conf.*, Curitiba, Brazil, Nov. 2013, pp. 389–394.

[190] U.-D. Reips, "Standards for internet-based experimenting," *Experimental Psychology*, vol. 49, no. 4, pp. 243–256, Feb. 2002.

[191] M. K. Wolters, C. B. Isaac, and S. Renals, "Evaluating speech synthesis intelligibility using Amazon Mechanical Turk," in *Proc. 7th ISCA Speech Synthesis Workshop*, Kyoto, Japan, Sept. 2010, pp. 136–141.

[192] "Qualtrics," `http://www.qualtrics.com`, 2017, [Accessed: Sep. 10, 2017].

[193] L. E. Harris and K. R. Holland, "Using statistics to analyse listening test data: some sources and advice for non-statisticians," in *Proc. 25th IoA Conf. Reproduced Sound*, Brighton, UK, Nov. 2009, pp. 294–309.

[194] E. C. Schwab, H. C. Nusbaum, and D. B. Pisoni, "Some effects of training on the perception of synthetic speech," *Human Factors*, vol. 27, no. 4, pp. 395–408, Aug. 1985.

[195] C. Donalek, "Supervised and unsupervised learning [slides]," `http://www.astro.caltech.edu/~george/aybi199/Donalek_classif1.pdf`, 2011, [Accessed: Sep. 10, 2017].

[196] M. A. Nielsen, *Neural Networks and Deep Learning*, Determination Press, 2015, `http://neuralnetworksanddeeplearning.com/index.html`.

[197] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (DNN) for parametric TTS synthesis," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Florence, Italy, May 2014.

[198] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Machine Learning Res.*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.

[199] A. van den Oord, S. Dieleman, and H. Zen, "WaveNet: a generative model for raw audio," `https://deepmind.com/blog/wavenet-generative-model-raw-audio/`, 2016, [Accessed: Sep. 10, 2017].

[200] K. Tokuda and H. Zen, "Directly modeling voiced and unvoiced components in speech waveforms by neural networks," in *Proc. Int. Conf. Acoustics Speech and Signal Process.*, Shanghai, China, Mar. 2016, pp. 5640–5644.

[201] M. Karjalainen and C. Erkut, "Digital waveguides versus finite difference structures: equivalence and mixed modeling," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 7, pp. 978–989, June 2004.

[202] J. O. Smith, "On the equivalence of the digital waveguide and finite difference time domain schemes," *arXiv:physics/0407032v4*, July 2004.

[203] D. Murphy, A. Kelloniemi, J. Mullen, and S. Shelley, "Acoustic modeling using the digital waveguide mesh," *IEEE Signal Process. Mag.*, vol. 24, no. 2, pp. 55–66, Mar. 2007.

[204] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. Int. Conf. Acoustics Speech and Signal Process.*, Salt Lake City, UT, May 2001, pp. 805–808.

[205] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by

style interpolation and morphing," *Proc. IEICE Trans. Information and Systems*, vol. E88–D, no. 11, pp. 2484–2491, 2005.

[206] C. Cooper, D. Murphy, D. Howard, and A. Tyrrell, "Singing synthesis with an evolved physical model," *IEEE Trans. Audio Speech Language Process.*, vol. 14, no. 4, pp. 1454–1461, July 2006.

[207] S. A. Van Duyne and J. O. Smith, "Physical modeling with the 2-D digital waveguide mesh," in *Proc. Int. Comput. Music Conf*, Tokyo, Japan, Sep. 1993.

[208] K. Dzhaparidze, *Parameter estimation and hypothesis testing in spectral analysis of stationary time series*, Springer-Verlag, New York, NY, 1986.

[209] N. Hojo, Y. Ijima, and H. Mizuno, "An investigation of DNN-based speech synthesis using speaker codes," in *Proc. INTERSPEECH*, San Francisco, CA, Sep. 2016, pp. 2278–2282.

[210] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, Aug. 1990.

[211] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[212] J. Epps, J. R. Smith, and J. Wolfe, "A novel instrument to measure acoustic resonances of the vocal tract during phonation," *Meas. Sci. Technol.*, vol. 8, no. 10, pp. 1112–1121, Oct. 1997.

[213] J. Botts and L. Savioja, "Integrating finite difference schemes for scalar and vector wave equations," in *Proc. Int. Conf. Acoustics Speech and Signal Process.*, Vancouver, Canada, May 2013, pp. 171–175.

[214] K. Richmond, "Acoustic-articulatory inversion," `http://www.cstr.ed.ac.uk/research/projects/inversion/`, 2009, [Accessed: Sep. 25, 2017].