Articulatory-Based English Consonant Synthesis in 2-D Digital Waveguide Mesh

Anocha Rugchatjaroen

PhD

Department of Electronics
University of York

June, 2014

Abstract

In articulatory speech synthesis, the 3-D shape of a vocal tract for a particular speech sound has typically been established, for example, by magnetic resonance imaging (MRI), and this is used to model the acoustic output from the tract using numerical methods that operate in either one, two or three dimensions. The dimensionality strongly affects the overall computation complexity, which has a direct bearing on the quality of the synthesized speech output.

The digital waveguide mesh (DWM) is a numerical method commonly used in room acoustic modelling. A smaller space such as a vocal tract, which is about 5 cm wide and 16.5-18 cm long in adults, can also be modelled using DWM in one, two and three dimensions. The latter requires a very dense mesh requiring massive computational resources; these requirements are lessened by using a lower dimensionality (two rather than three) and/or a less dense mesh. The computational cost of 2-D digital waveguide modelling makes it a practical technique for real-time synthesis in an average PC at full (20 kHz) audio bandwidth. This research makes use of a 2-D mesh with the advantage of the availability and flexibility of existing boundary modelling and the raised-cosine impedance control to study the possibilities of using it for English consonant synthesis.

The research was organized under the phonetic ‘manner’ classification of English consonants as: semi-vowel, nasal, fricative, plosive and affricate. Their production has been studied in terms of acoustic pressure wave propagation. Meshing topology was fixed to being a 4-port scattering 2-D rectilinear waveguide mesh for ease of understanding and mapping to the tract shape.

As the characteristic of consonant production requires vocal tract articulation variations that are quite unlike vowels, this research adopts the articulatory trajectories using electromagnetic (mid-sagittal) articulograph (EMA) data from *mngu0* to guide the change of cross-sectional vocal tract area. Generally, articulatory trajectories have been used to improve the accuracy of speech recognition and synthesis in recent decades. This research adopts the trajectories to control coarticulation in consonant synthesis to demonstrate that a 2-D digital waveguide mesh (DWM) is able to simulate the formant transition accurately. The formant transitions in the results are close acoustically to natural speech and are based on controlling articulation for four places of articulation. Positions of lip, tongue tip, tongue body and tongue dorsum are inversely mapped to their corresponding cross-sectional areas. Linear interpolation between them enabled all tract movements to be modelled. The results show that tract movements are best modelled as non-linear coarticulation.

Contents

List of Figures

List of Tables

To my grandmother
Lamoon Rugchatjaroen

Acknowledgements

I would like to thank the Office of the Civil Service Commission (OCSC) for funding my PhD course at the University of York, and my supervisor Prof. David M. Howard for his kind support and guidance throughout my studentship. I would also like to thank my colleague in York, Matthew Speed, who provided very useful tools, MRI data and great research advice. I would also like to thank the EMA research team at the Centre for Speech Technology Research in the University of Edinburgh for publishing mngu0 EMA corpus.

Thanks to my great friends Aglaia Foteinou, Sorrel Harriet, Christin Kirchhuebel and Andrew Chadwick for their full support, kind suggestions and advice in both personal and work issues. They have made my life in York warm and colourful.

I also feel lucky to have my Thai best friends in the UK, Permtip Buaphet and Areeya Limmanonukul for their great friendship, support and advice, and also my Thai friends, Rungnapa Tagun and Paramita Punwong for nice times dining together. I would also like to thank my English-Thai consultant Pojanut Suthipinittharm for all of her quick responses and kind help.

I owe it to Imke van Heerden, Carrie Ching-Yi Liang, Takako Nonaka and Frank Soboczenski, my great house-mates, for great friendship and nice international dining.

I truly appreciate the inspirations and supports from my family: my grandmother, Lamoon, my parents, Preecha and Ing-orn, my brother, Krish Rugchatjaroen, and my sister-in-law, Metinee Sangkrajang.

Last but not least, I would like to thank my best friend Nitass Sutaveephamochanon and his family, who always provide me with warm suggestions, great vacation trips, Thai dried food and financial support.

Declaration

I hereby declare that this thesis is entirely my own work and all contributions from outside sources, through direct contact or publications, have been explicitly stated and referenced. I also declare that some parts of this research have been presented previously, at conferences. These publications are listed in Chapter 1.

Chapter 1
Introduction

Speech production has been studied for hundreds of years. A variety of approaches have been used, from the physical, biological and mechanical to the electronic, in the attempt to understand and to mimic human speech production. Focusing on what occurs in the vocal tract, one approach to understanding what happens during speech production involves the study of acoustic phonetics which is an area of analysing and describing the acoustic characteristics of the speech signal in relation to speech production. In this work, the acoustic output is calculated by simulating pressure propagation in the vocal tract using the two-dimensional digital waveguide mesh (2-D DWM).

Motivation

In early attempts to study speech production, mechanical blowing instruments were adjusted to mimic the sound of a human vowel [1]. Today, directly mimicking human speech production is not such a common approach due to its complexity, and the general market requirement for text-to-speech synthesis. Computer and computational performances are now of sufficiently high quality to enable speech synthesis by various other approaches, including concatenative speech synthesis, HMM (hidden Markov model-based speech synthesis), and deep neural network-based speech synthesis [2]. However, the study of speech production physically is still essential in biological or medical research to evaluate the output from a compromised vocal tract [3]. Therefore, sound simulation is an approach to be considered. It can be done by applying numerical methods to simulate the physics of the output acoustic signal. Finite-difference time-domain (FDTD), finite volume and other numerical methods can work successfully for the simulation [4], but they require very high computational resources to produce natural-sounding speech. The idea is to explore whether there is any possibility of synthesizing sound by considering the underlying physics with an acceptable level of accuracy but with sufficient speed to work in real-time, and this makes the 2-D DWM a good choice with which to experiment. Mullen's research looked at this between 2004 and 2006 in relation to evaluating the 2-D DWM for real-time vowel synthesis [5]. He also conducted an experiment on smoothing the signal propagation when articulating by weighting the impedance by cosine function. Here, this approach is extended to include the remaining consonant phones of English and to establish whether this numerical method is practical enough to be used in articulatory speech synthesis.

Hypothesis

The hypothesis of this research is that articulatory-based English consonant synthesis can be achieved using a 2-D digital waveguide mesh (DWM). A 2-D DWM will be used in the simulation of wave propagation in a virtual oral tract coupled as appropriate to a virtual nasal tract. More functions such as dynamic tube shaping, additional noise sourcing in the middle of the tract and duration controlling will also be implemented. The proposed system should be able to synthesize vowels, nasals, fricatives, plosives, affricates and semi-vowels. To achieve this, a list of research subtopics has been drawn up as follows:

- Simulating the tracts using a more precise area function for more dense scattering junctions from existing MRI data (to be more precise in wave propagation simulation).

- Inserting white noise injection (to be used for fricatives and plosives).

- Adopting articulation trajectories for synthesizing consonant-vowel (CV) sounds.

After all of these milestones and the characteristics of English consonants have been studied, the proposed system will be evaluated by comparing the results' characteristics with their theoretical acoustic characteristics. A perception test will be performed on a set of synthesized CV sounds in order to explore a percentage of perceptibility on formant transition.

Proposed Approach

This research proposes a study of using 2-D DWM in synthesizing English consonants. Firstly, the nasal model is attached to the existing oral tract to enable simulation of all consonants involved in speech production. Secondly, an external noise source is implemented behind the appropriate place of constriction and, finally, the articulation is developed by adopting experimental data for articulatory trajectories into the area function. In addition, articulatory details relating to specific configurations for each manner of articulation of English consonants are included.

The setting of the proposed system works separately for each manner of articulation to analyse the resulting acoustic output characteristics. Therefore, the set of target characteristics for each manner and place of articulation of English consonants is studied. The centre frequencies of bursts, range of frication frequencies, formant transitions, and anti-formant frequencies will be considered as appropriate to synthesize the representative acoustic characteristics of English consonants.

Scope

This proposed work studies the 2-D DWM acoustic pressure propagation for English phonemes, but does not cover tones, intonation, stress or duration. In addition, the vowel synthesis part will use Jack Mullen’s synthesizer [5].

The proposed system will be tested in terms of its resonation. The set of appropriate articulatory characteristics this research will analyse includes the centre frequency of bursts, formant transition, anti-formant frequencies and range of frication frequency response which are the results from tube resonances. It should be noted that some of the acoustic characteristics which are not from resonation, such as the voice onset time, will not be included in this research.

The proposed system will consolidate the previous successful system with noise source in the vocal tract and add a nasal tract system, together with appropriate timing controls, into a laboratory prototype of an articulatory English speech synthesizer. The target performance of the proposed system is compared to some well-known synthesizers after Childers [6].

Table 1.1. Target performance of the proposed articulatory speech synthesis instrument compared with some successful articulatory speech synthesizers (adapted from Childers [7]).

	Flanagan	**Maeda**	**Childers**	**Proposed Synthesizer**
Model of excitation	Self-oscillating two-mass	A slit	LF	LF
Jitter and Shimmer models included	No	No	Yes	No
Noise source at the glottis	No	No	Yes	Yes
Noise source in the vocal tract	Every section	No	Centre of, or downstream or upstream from, or distributed along the constriction	Downstream from the constriction
Excitation in the vocal tract	No	No	Yes	Yes

Overall, this research reports on experiments to characterize the acoustic outputs of English consonant synthesis using the 2-D DWM technique. Previous research by Jack Mullen [5] provides the starting point, and this is modified for each type of English phone. The results from this study show the system’s accuracy, in both spectral and perceptual terms, in synthesizing speech sounds and demonstrates its capacity to implement a synthesizer for English phones.

Contribution through published papers

A. Rugchatjaroen and D. M. Howard, "A STUDY ON DYNAMIC VOCAL TRACT SHAPING FOR DIPHTHONG SIMULATION USING A 2D DIGITAL WAVEGUIDE MESH", Proc. of the 15th Int. Conference on Digital Audio Effects (DAFx-12), York, UK, September 17-21, 2012.

A. Rugchatjaroen and D. M. Howard, "THE ACOUSTICS OF CONSTRICTION IN A VOCAL TRACT MODEL USING 2D DIGITAL WAVEGUIDE MODELLING", 10th International Seminar on Speech Production, Cologne, Germany, 2014.

A. Rugchatjaroen and D. M. Howard, "FLEXIBILITY OF COSINE IMPEDANCE FUNCTION IN 2-D DIGITAL WAVEGUIDE MESH FOR PLOSIVE SYNTHESIS", in 2014 2nd IEEE China Summit & International Conference on Signal and Information Processing, Xi' an, China, 2014.

Chapter 2
Speech Synthesis

In general human speech communication, the information to be communicated is sent from the sender's brain via signalling articulators through the nervous system to the muscles. The speech is produced and transferred through the air around the speaker, then the listener's ears perceive the air vibrations and signal the nervous system the receiver's brain before being translated into information. This process is called the speech chain [7].

The speech chain encapsulates the processes from the speaker's intention all the way through to the listener's understanding. There are ten stages in the chain: intention, meaning, utterance, articulatory plan, articulation, sound, auditory response, word sequence, meaning and understanding [8]. Speech synthesis intervenes in the process of transformation from utterance to articulatory plan to articulation and to sound.

In the mid-19th century, Alexander Ellis started to use the Roman alphabet to describe phonetic symbols in his pronunciation transcription work for any language [9]. Paul Passy published the first IPA alphabet in 1888 [10] and then in 1946 the first speech visualization sound spectrograph was invented at Bell Laboratories [8, 11]. The first British English synthesized speech by rule was generated in 1964 [12]. In 1970, Gunnar Fant published his ground-breaking work on the physics of speech sound production: *Acoustic Theory of Speech Production* [13]. Subsequently electronic tools for articulatory studies were invented: electropalatography (EPG) for measuring the degree of tongue-palate contact in 1972 [14, 15]; magnetic resonance imaging (MRI) for imaging the articulators in 1974 [16];

laryngography for measuring vocal fold contact in 1977 [17]; electromagnetic articulography (EMA) for recording mid-sagittal position of articulators in 1992 [18]; and dynamic/real-time MRI in 2010 [19].

Before exploring the details of speech synthesis, a general overview of the physics of sound is provided in the next section on sound waveforms.

2.1 Sound Waveforms

Sound is an audible waveform that is created by one or more vibrating objects that initiate a sequence of pressure changes to particles in the surrounding medium (air, solid, liquid). If the frequency of vibration is in the audible range (normally 20 Hz to 20 kHz but changing with age and noise abuse [20]), then the disturbance is heard as sound. The movement of particles during sound transmission is in the same direction as the propagation of the sound itself, and it is therefore called a longitudinal wave.

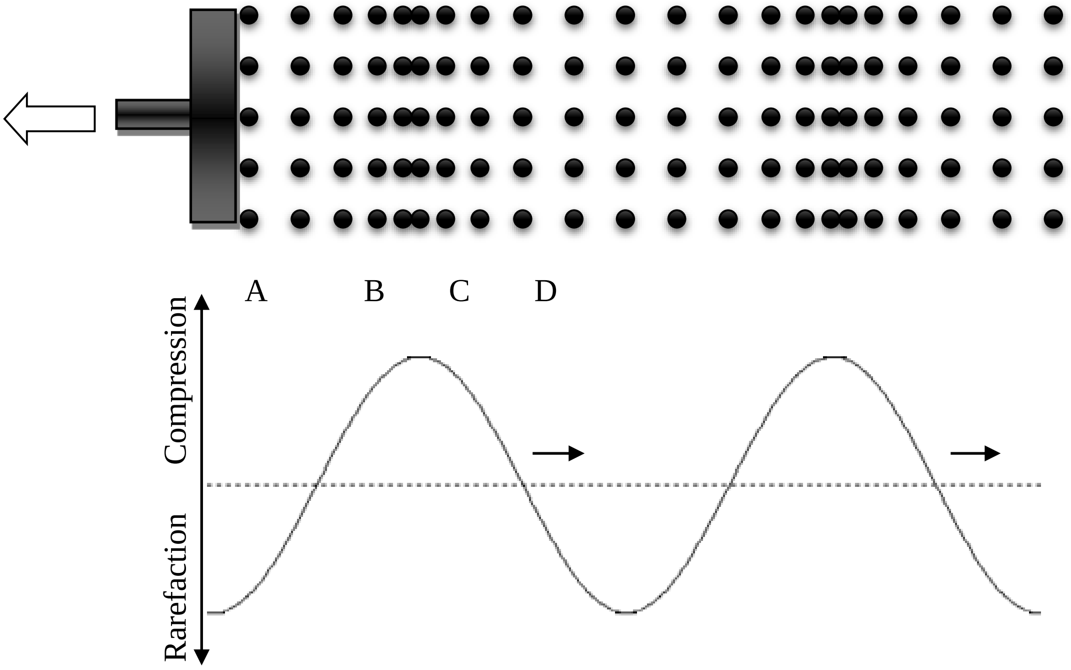
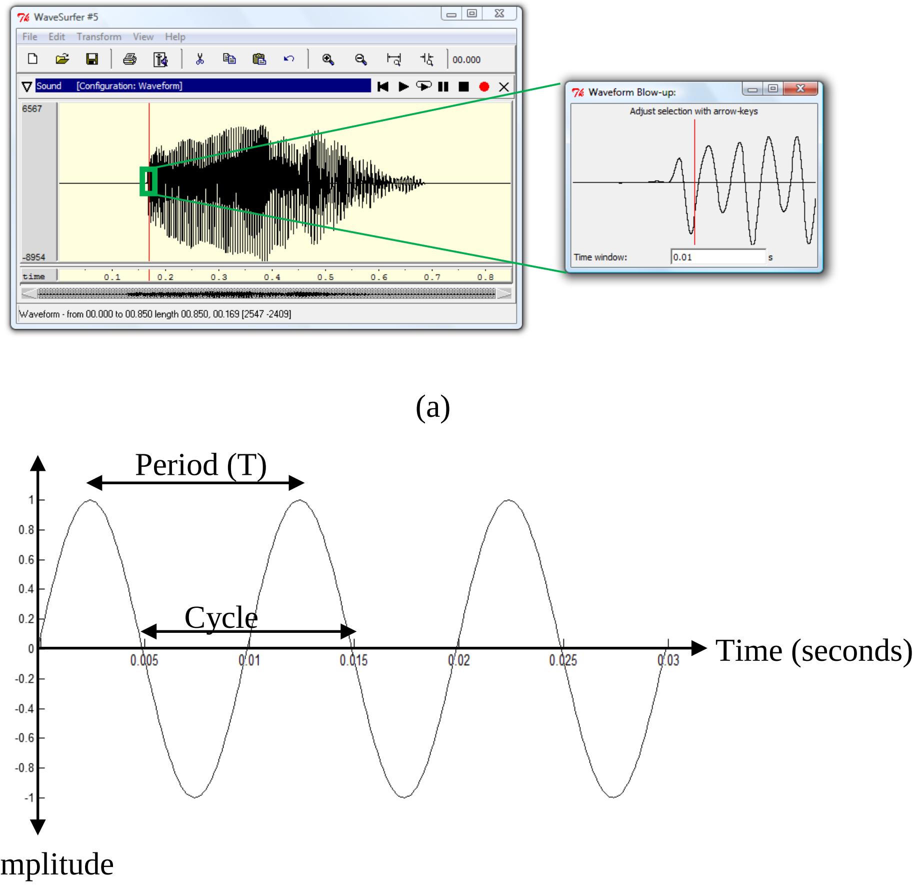


Figure 2.1. A snapshot of the movement of molecules simulation during sound transmission.

The pressure moves longitudinally with the velocity. If we set up a recording system and record a sequence of sound pressures along with their time stamp, we can plot the fluctuation of the pressure against time in the x axis which is an ordinary representation of sound (waveform view).



Waveform - from 00.000 to 00.850 length 00.850, 00.169 [2547 -2409]

Time window: 0.01 s

(a)

(b)

Figure 2.2. (a) Speech waveform of phone /u/ on Wavesurfer – an Open Source tool for sound visualization and manipulation by KTH [21] and (b) An example of a 100 Hz waveform.

In the waveform view, it clearly shows important wave characteristics such as the period (T), cycle, etc. In a simple periodic wave which has only uniform pressure movement, the period of the wave is the distance between a repeating feature on the waveform such as positive peak, negative peak or zero crossing; a cycle is from the time between the repeating features. The fundamental frequency is the number of cycles per second. For a complex periodic wave, the waveform contains more than one frequency component. Joseph Fourier demonstrated mathematically that any wave shape can be synthesized as a combination of sine waves with appropriate amplitudes, frequencies and phases which are known as Fourier components [22].

If we let the simulation in Figure 2.1 continue, and a set of repeating force is given, the pressure propagation can be plotted in the time domain. A time gap between each push is called a period (T) and the distance between the repeating pulses is called the wavelength (λ). For example, if the object on the left-hand side of the figure vibrates and pushes the particles every 5 milliseconds (ms), the period of the sound of the simulation will be 5 ms. The wavelength varies according to the velocity of sound and equals to

$$\lambda = c\ T \qquad 2.1$$

where *c* is the velocity of sound. The velocity of sound depends on the medium itself and the temperature. For example, the velocity of sound in air at 20 degrees centigrade at sea level is 344 m/s. Looking back at the previous simulation, if we assume that it occurs in the air at sea level, the wavelength would be equal to 344 x 0.005 = 1.72 m. Moreover, a number of periods of sound in a second equals its frequency (*f*) which is calculated by

$$f = 1\ /\ T \qquad 2.2$$

Therefore the velocity of sound could be calculated by

$$c = f\ \lambda \qquad 2.3$$

and the frequency can also be calculated by

$$f = c\ T = c\ /\ \lambda \qquad 2.4$$

Further details of the velocity of sound in other mediums can be found in Howard and Angus, 2009 [23].

The velocity also has an important relationship with pressure (*p*) – a 90^{o} phase difference relationship. If we look at the particles in Figure 2.1 more closely, when they start moving from the left-hand side to the right by being pushed by an object on the left, the velocity rises because the particles are moving apart but the pressure becomes lowest (as the particles at A in Figure 2.1). After the particles have been moving for a while the velocity reaches a maximum and the pressure becomes positive (at B in Figure 2.1). Increased density then makes the particles move more slowly which, in turn, causes a decrease in the velocity. When the pressure reaches the maximum the velocity reaches zero (at C in Figure 2.1) the particles begin moving back, which makes the velocity negative and the pressure becomes lower and lower because the particles are moving apart again (at D in Figure 1). This situation makes the pressure have a 90^0 phase lag compared with the velocity (Howard and Murphy, 2008 [20], Johnson, 2003, [23]).

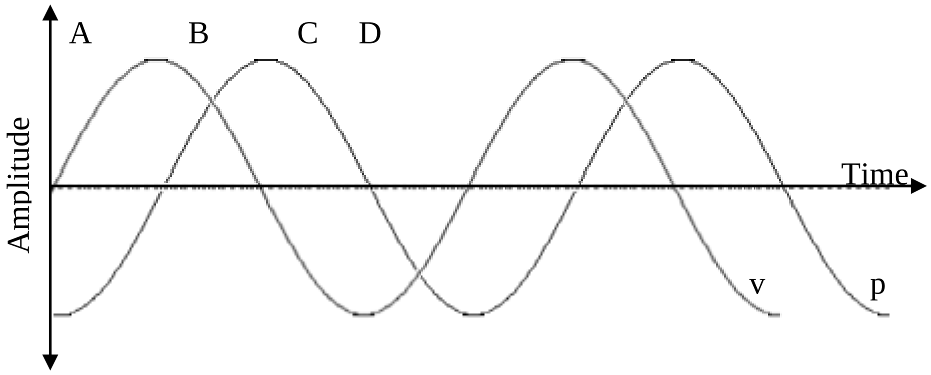


Figure 2.3. Plotted relationships between pressure and velocity (90^0 phase difference).

Considering the relationship between pressure and velocity in terms of the density and springiness of the propagating medium, a medium molecule moves faster in low density and weak spring for a given pressure amplitude. The relationship is

$$\frac{Pressure\ amplitude}{Velocity\ amplitude} = Z_{acoustic} \qquad 2.5$$

where $Z_{acoustic}$ is known as the acoustic impedance. Considering the propagation through a medium, the velocity depends on springiness and density of the medium.

$$v = \sqrt{\frac{E}{\rho}} \qquad 2.6$$

where E is Young's modulus value measuring the force needed to compress a medium (Nm^{-2}) and ρ is the density of the medium. In air, Young's modulus and the density has to be considered as a gas. The velocity of sound needs to be considered in the adiabatic gas law equation which considers pressure in a volume of the gas as a constant (γ) and Young's modulus for air is given by

$$E_{gas} = \gamma P \qquad 2.7$$

where P is the pressure of the gas (N m^{-2}) and γ is 1.4 for air. For density of a gas, it is given by

$$\rho_{gas} = \frac{m}{V} = \frac{PM}{RT} \qquad 2.8$$

where m is the mass of the gas (kg), M is the molecular mass of the gas (kg mole^{-1}), R is the gas constant (8.31 J K^{-1} mole^{-1}) and T is the absolute temperature (K).

Then the velocity of sound in the gas can be considered after temperature T and constant γ, R and M .

$$v_{gas} = \sqrt{\frac{E_{gas}}{\rho_{gas}}} = \sqrt{\frac{\gamma RT}{M}} \qquad 2.9$$

Assuming γ can be ignored, equation 2.5 can be rewritten by dividing equations 2.7 by 2.6 as:

$$Z_{acoustic} = \sqrt{\rho^2 \left(\frac{E}{\rho}\right)} = \rho c \ \text{ kg m}^{-2}\text{ s}^{-1} \qquad 2.10$$

Then the impedance can be considered in a specific context such as a tube as

$$Z_{acoustic\ tube} = \frac{\rho c}{A_{tube}} \ \text{ kg m}^{-4}\text{ s}^{-1} \qquad 2.11$$

where $A_{tube}$is the tube area. This equation will be considered later in Chapter 4 for acoustic simulation in a tube.

2.2 Harmonics and Resonance

In complex periodic waves, components are based on a fundamental frequency (f0) such that each has a frequency that is an integer multiple of f0 and its own amplitude and phase. Figure 2.4 shows an example of the harmonics of a 100 Hz f0 wave. All of them are 0.03 seconds long. The top wave shows the first harmonic at 100 Hz. The following two show the second and third harmonics at 200 and 300 Hz respectively, with lower amplitude. The harmonic number is called after the integer that is used to multiply with f0. The last waveform results from mixing the top three sine waves. Their amplitudes are 0.5, 0.43 and 0.4. The relation between frequency and amplitude can also be plotted and is known as a power spectrum.

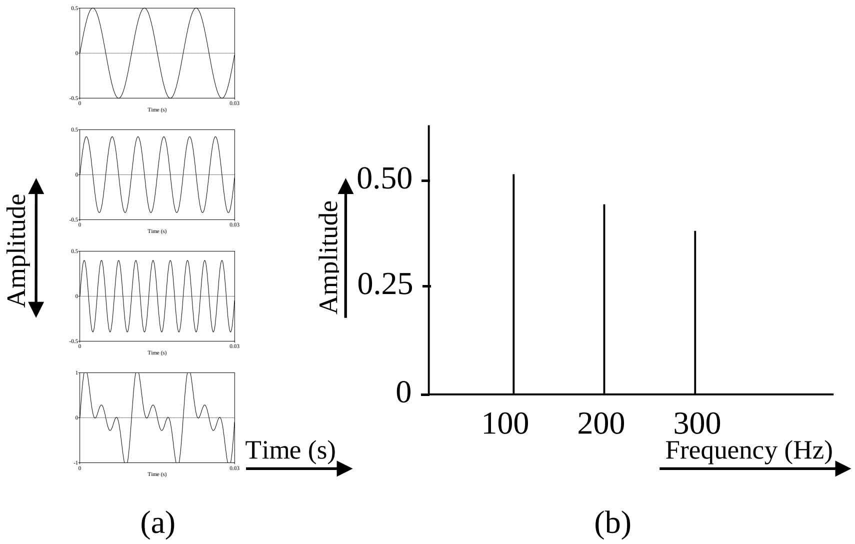


Figure 2.4. (a) An example of the 1st, 2nd and 3rd harmonics of 100 Hz sine wave and their sum. (b) A spectrum of waveforms in (a).

The power spectrum plot (spectrogram) shows the amplitude of frequencies analysed from the waveform, usually using fast Fourier transform (FFT) algorithm or a linear prediction coefficient (LPC). The plot depicts the resonances found in analysing sound. Generally, in a speech sound, the complication of source and resonation causes some peaks and/or dips in its resonances. The peak resonance/frequency is known as a formant. The formant frequencies are named according to the order in which they appear from low to high frequency: the peak at the lowest frequency is considered as the natural frequency of that sound and called F0, then the next peak (in higher frequency) is called F1 and then F2 and so on. These formants are one of the most important characteristics of speech sound and allow us to distinguish different vowels.

The resonation occurs when a travelling wave hits a boundary and then reflects back to create one or more complete loops. The standing wave or resonance then requires a suitable wavelength to travel a suitable distance between two boundaries. For example, the first or the lowest resonance of a wave travelling between two hard boundaries is when a wave with wavelength λ travels distance L where L = λ/2. The wavelength of the lowest is then equal to 2L which has c/2L of frequency (from equation 1.4) where c is the velocity of sound. The other wavelengths which correspond to the same proportion of L would also be resonated, such as λ, 3λ/2, 2λ and vice versa. In the same way, a travelling wave between bound-unbound boundaries will have the resonance at wavelength λ/4, 3λ/4, 5λ/4 and so on, and therefore the resonance frequency could be at (2n + 1)/4L where n is 0, 1, 2, 3,.., ∞ [23].

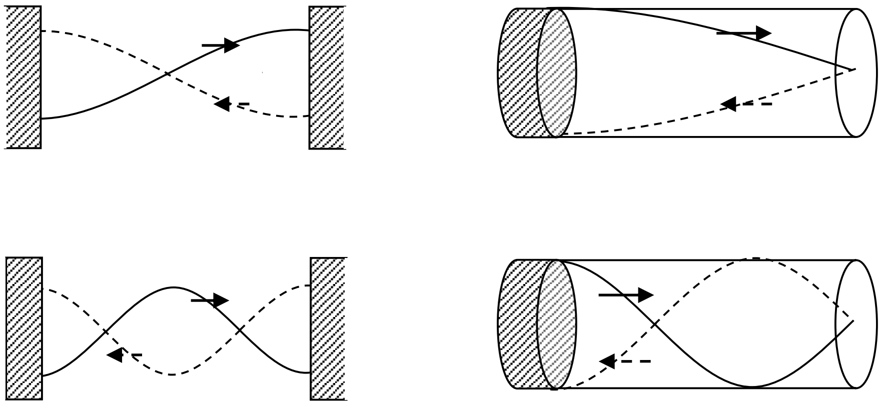


Figure 2.5. On the left are the first two resonances between two hard boundaries. On the right are the first two resonances in a one-sided open tube.

In the time domain, the waveform can be written as a function of amplitude $x(t)$ at time *t*. For a representation of sine wave, the function can be written as

$$x(t) = \sin(t) = \cos\left(t - \frac{\pi}{2}\right) = \cos(t + \emptyset) \qquad 2.12$$

when $\emptyset$ is $[-\frac{\pi}{2}, \frac{3\pi}{2}]$.

From the sine wave in Figure 2.2, a single frequency periodic wave equation can be written as

$$x(t) = x(t+T) = x(t+2T) = x(t+3T) = \cdots \qquad 2.13$$

The repetition can be considered in terms of frequency F instead of period T with the angular frequency ω which equals to

$$\omega = 2\pi F = \frac{2\pi}{T} \qquad 2.14$$

It can then be written as

$$x(t) = A\cos(\omega t + \emptyset) \qquad 2.15$$

In general, for a periodic wave which contains more than one frequency, the lowest frequency is called the fundamental frequency F_0 and the period of one cycle is called the fundamental period $T_0 = 1/F_0$. Then the angular frequency ω_0 of F_0 equals to $1/(2\pi T_0)$ and the harmonic frequency is in a series of F_0, $2F_0$, $3F_0$, …

Equation 2.15 can be expanded to a combination form of multiplication of fundamental frequency for a general periodic wave as

$$x(t) = a_0\cos(0\times\omega_0 t + \emptyset_0) + a_1\cos(1\times\omega_0 t + \emptyset_1) + a_2\cos(2\times\omega_0 t + \emptyset_2) + \cdots \quad 2.16$$

Considering the first term as a constant because $\cos(\emptyset_0)$ is a constant, then equation 2.16 can be written as

$$x(t) = A_0 + \sum_{k=1}^{\infty} a_k\cos(k\omega_0 t + \emptyset_k) \qquad 2.17$$

Equation 2.17 is a specific form of the Fourier series which is often used to analyse a periodic wave/signal. Note that for non-periodic signals, the Fourier transform is used instead [2].

2.3 Acoustic Representation in the Human Voice

The human voice, whether in speech, singing or exclamation, works as a carrier in delivering a message from a speaker to a listener. Frequencies, formants, duration, tone or even a short pause all carry some meaning in human perception. To adjust these attributes the human vocal apparatus, including all organs from the diaphragm up to the mouth and nostrils, is involved in modifying the power source, sound source and sound modifiers. The power source in speech is the human breath. Muscles around the ribs work together with the diaphragm, sucking air into the lungs and slightly pushing air back out through the glottis and vocal tract. Any constriction in the tract establishes the sound source which can be classified as voiced, voiceless or mixed, depending on the manner of the constriction. Figure 2.6 from Howard, 2008 [23] shows ideal sound sources (voiced and voiceless) with their spectra.

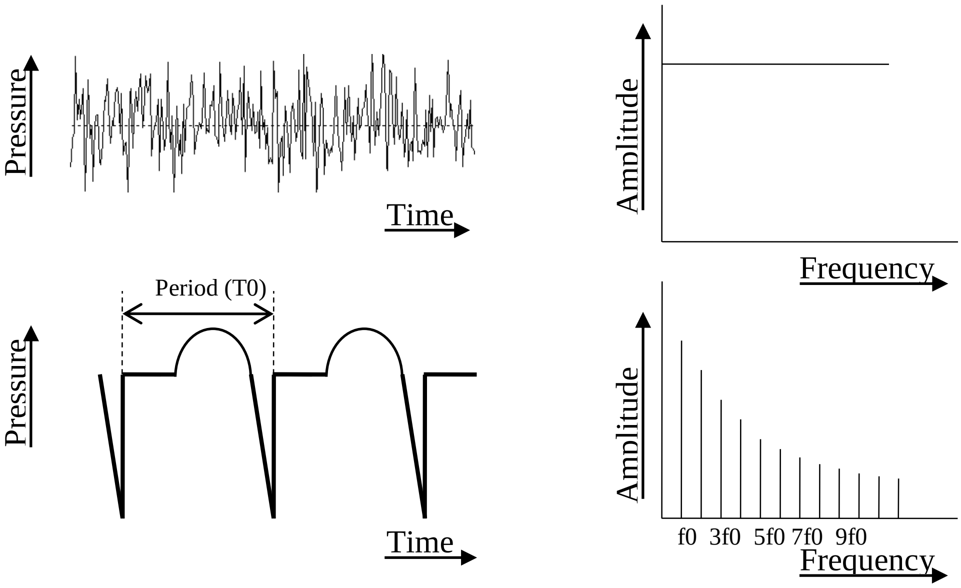


Figure 2.6. Idealized voiceless sound source (upper) and voiced sound source (lower) and their spectra (right) from [20].

These two types of sound source are classified by a repetition pattern of the waveform which is caused by cyclic vibrations of the vocal folds. The upper non-repetitive wave is an example of an idealized voiceless sound source with equal amplitude across the frequency range, while the lower example shows the repetition of the voiced sound source. A period (T_0) in the voiced sound is the time taken for a pattern of pressure fluctuation or for a cycle of vocal folds vibration.

The sound modifier in a human vocal system is the vocal tract. The shape or volume of each part in the tract resonates at different frequencies, and therefore the frequencies of the output voice have different amplitudes. There are peaks in the spectrum which are called formants, and the relationship between the formant frequencies and the tract shape is explained by perturbation theory (see more detail in Chapter 4 section 4.2).

2.4 Introduction to Speech Synthesis

Human speech communication is the transfer of information from one person to others via speech. The transfer process from a speaker's brain to the arrival of the message in the listener's brain is known as the speech chain [7]. Each part of the chain can be implemented by various simulation methods or devices but, in terms of speech synthesis, it takes place only in a part of the speaker's role as a generator of speech.

There have been several attempts to synthesize speech in the past. The first mechanical speaking machine was recorded by Charles Darwin in 1806, based on his grandfather's experiment in 1771 mentioned in [8]; there is not much detail of its engineering. The second that was fully recorded was Kratzenstein's vowel resonators [24]. Christian Gottlieb Kratzenstein, professor of physiology at the Imperial Academy of St Petersburg successfully resonated vowel-like sounds at a constant pitch when his equipment was activated by a reed in 1779. Twelve years later, the first recorded success in synthesizing connected speech was achieved by von Kempelen in Vienna [25].

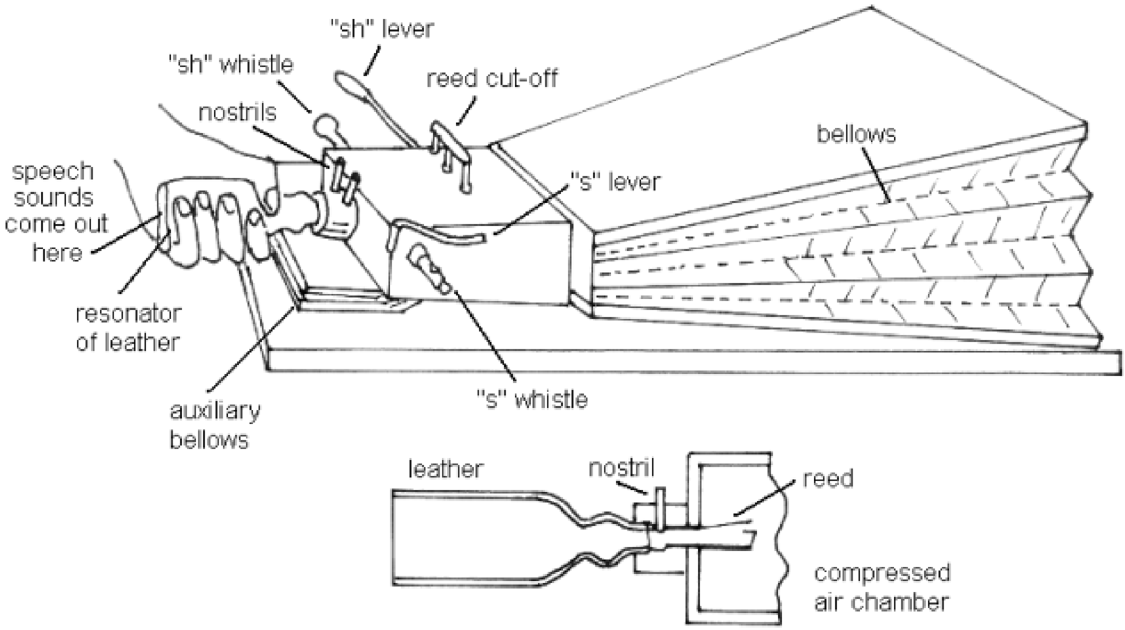


Figure 2.7. Structure of Wheatstone's reconstruction of von Kempelen's speaking machine from Flanagan, 1972.

In 1845 Joseph Faber introduced a device which was suitable for singing synthesis [26]. It was a model of the tongue and a pharyngeal cavity whose shape could be controlled via a key board. In the next century, R. R. Riesz's talking mechanism successfully produced the word 'cigarette' in 1937. His device was shaped like the human vocal tract with rubber and metal with ten control keys to support two hands. It is a mechanical articulatory speech synthesizer that can produce fairly good speech with a well-trained operator. A few years later, the first commercial electronic device, the VODER, was developed and introduced by Homer Dudley (1939) [27] which produced human-like speech and created considerable based on articulatory synthesis interest in the artificial speech production research area. In 1989, the talking machine by Martin Riches was introduced. It contains 32 pipes with air valves, wind chests, magazine bellows, blower and a computer which operated as a user interface to control valve movements for each wind chest (a picture of the machine can be found in [8]). Figure 2.8 shows some milestones of speech synthesis from its beginning until now after Sami Lemmetty [28] and some more recent milestones from [2, 6].

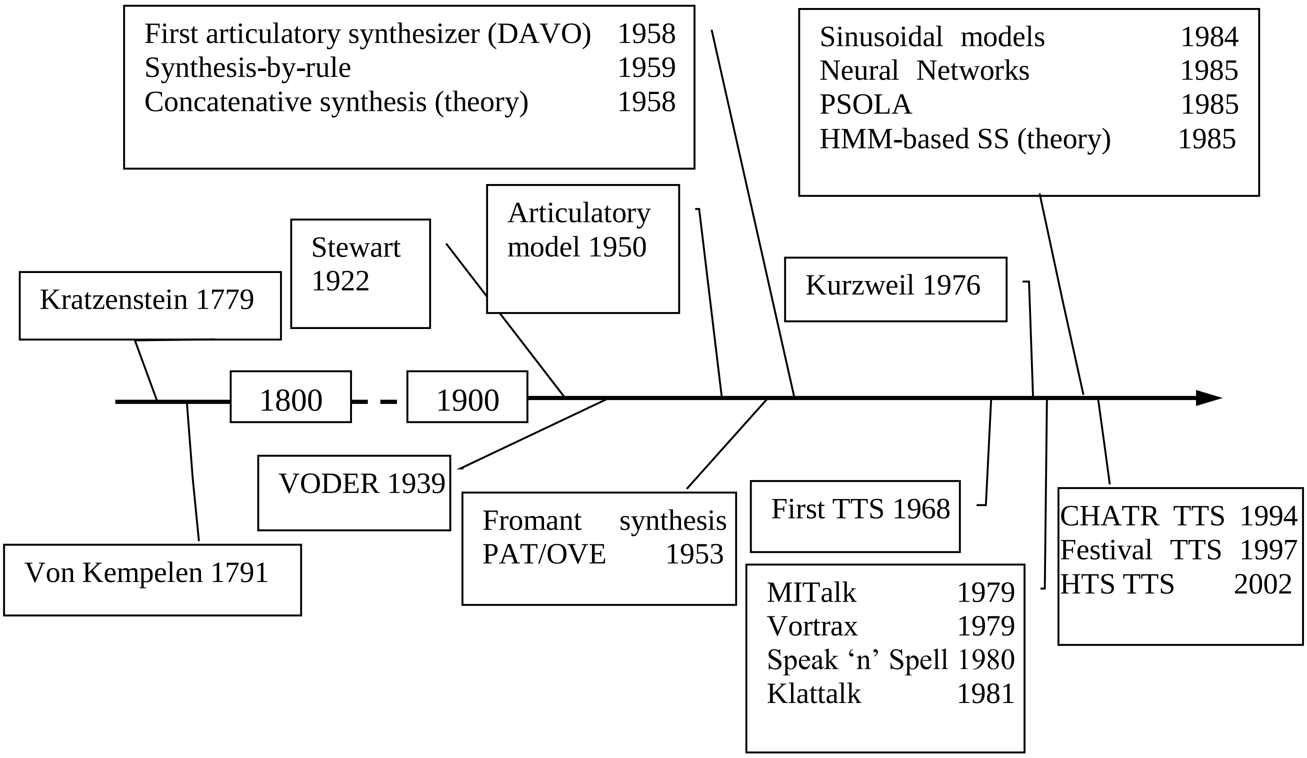


Figure 2.8. Some milestones in speech synthesis adapted from [28].

All of the speech synthesis methods are usually classified into three groups:

- Articulatory synthesis, which attempts to model the human speech production system directly.
- Formant synthesis, which attempts to model pole and zero frequencies of the speech signal by source-filter modelling.
- Concatenative synthesis, which attempts to concatenate different lengths of prerecorded samples from natural speech.

The concatenative methods including HMM-based [29] are the most commonly used in today's text-to-speech synthesis system. The approaches that are used in today's market speech synthesizers are HMM-based + STRAIGHT, hybrid HMM-based + unit selection, parallel HMM, and deep neural network-based (more details can be found in [30, 31]). However, the articulatory method still has the potential for higher quality implementations, especially better co-articulation, in the future.

2.5 Introduction to Articulatory-based Speech Synthesis

Articulatory speech synthesis is used to transform the articulation to sound in the speech chain. It involves understanding the articulator movements from the sound generators to speech signals at the lips and nostrils. The articulators of the human speech apparatus can be sketched in a mechanical view as in Figure 2.9 after Childers [6]. A balloon on the left-hand side represents the lungs which push pressure P_s to the mechanical vocal fold model where the pressure flow vibrates. U_G represents the velocity of the propagated pressure from the glottis. It passes through the oral tract and/or the nasal tract and is then released at the nostrils and lips.

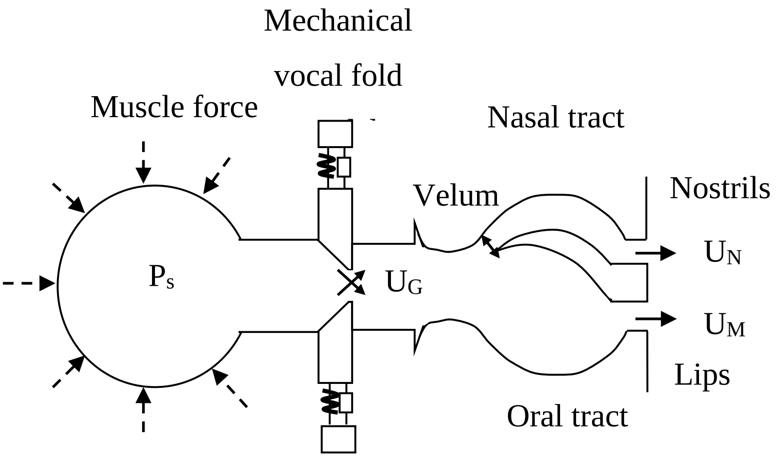


Figure 2.9. A mechanical model after [6].

In the upper vocal tract the articulation perturbs the source from the glottis. To simulate the articulation, Childers described the modelling as two separated parts (the articulatory model and the acoustic model), as shown in Figure 2.10 [6]. For the articulatory model, the vocal tract is viewed as a structure of small ducts with corresponding cross-sectional areas that are used as parameters to represent the vocal tract characteristics in the acoustic modelling. Each cross-sectional area is basically used as an electrical transmission line by simulating the shape of the vocal tract with sometimes the area function changing by time for the articulation. Steps for co-articulation are specified by changes of the articulators from frame to frame over the synthesizing process. All of the data for mid-sagittal distance modelling are observed by photography, x-ray technology or magnetic resonance imaging (MRI) [6, 32].

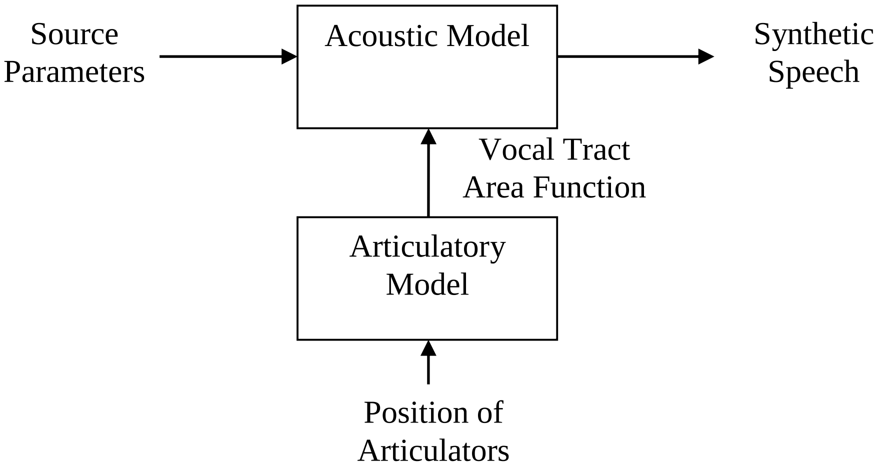


Figure 2.10. A model of articulatory speech synthesis after [6].

In practice the first articulatory synthesizer, named DAVO (Dynamic Analogue of the VOcal tract), was introduced in 1958 by George Rosen from the Massachusetts Institute of Technology. DAVO uses the transmission line for acoustic modelling [33]. Figure 2.11 shows a block diagram of its control system in (a) and a photograph of DAVO the synthesizer in (b). In figure 2.11b, the principal unit arrays are allocated in racks. From left to right they are: rack 1 – the function generators; rack 2 – the timer with time-selection matrix; rack 3 – the buzz and noise generators and part of the configuration matrix; rack 4 – the configuration matrix; rack 5 – the transmission line; and rack 6 – power supplies for the transmission line [33].

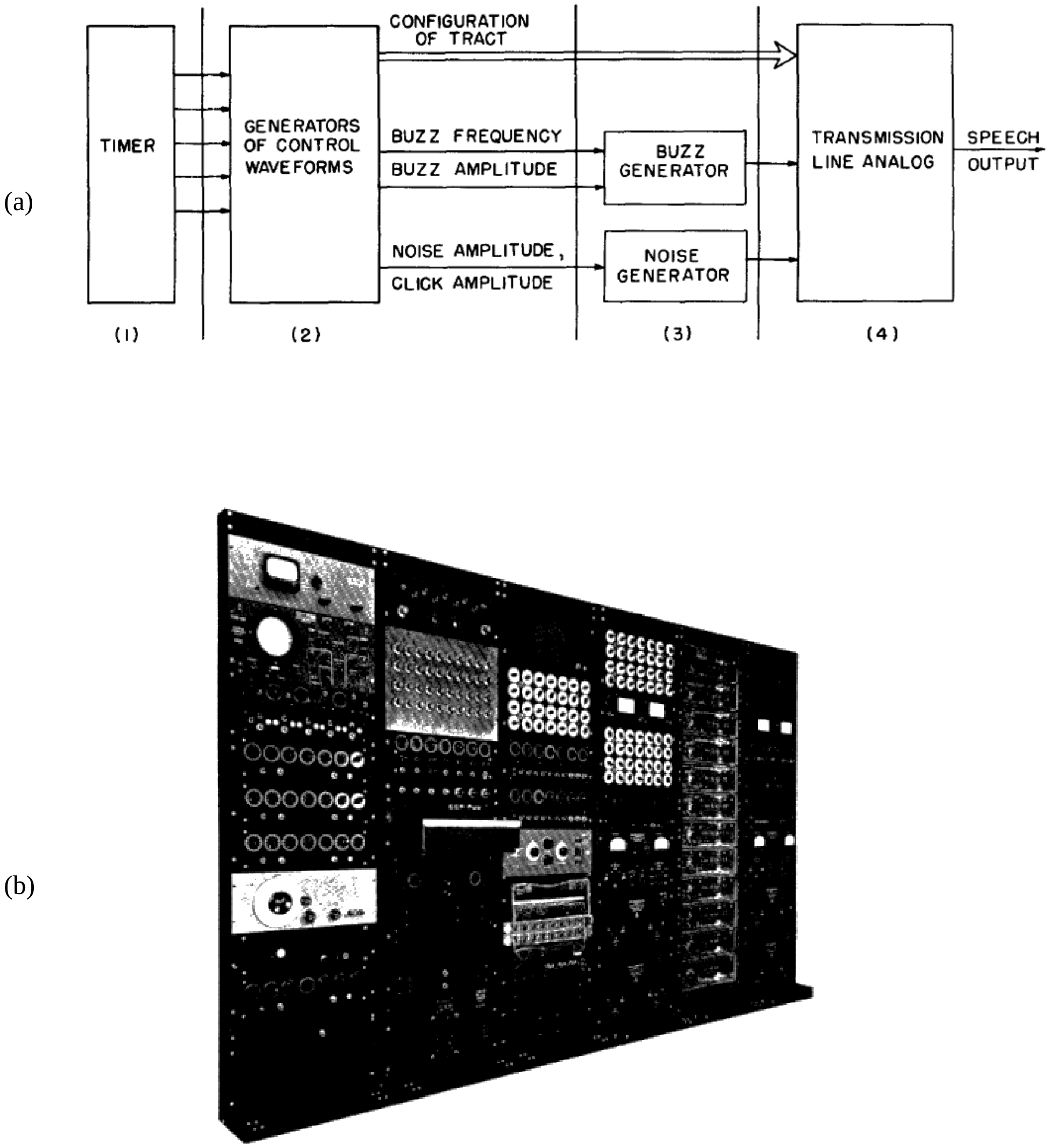


Figure 2.11. (a) Block diagram of the control system and (b) Image of Dynamic Analogue of the VOcal tract (DAVO) - the first articulatory synthesizer from George Rosen [33]

2.6 Tube models and Wave Scattering

From the mechanical to the electronic and/or computerised versions, the human vocal tract has been modelled in terms of a resonator. The tube model represents the tract using cross-sectional areas as parameters. From the parameters, various simulations of the sound propagation have been observed to help the understanding of acoustic behaviour in speech – resonance, formants and some aspects of the relation are between articulatory configurations and their acoustic consequences.

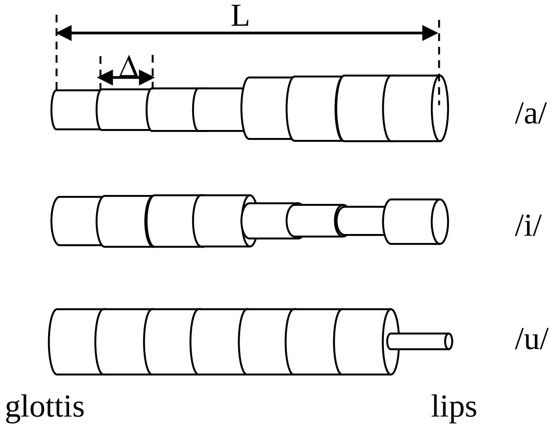


Figure 2.12. Approximated tube models of vocal tract shapes for vowels /a/, /i/ and /u/.

The Kelly-Lochbaum model [34] is one of the very first digital speech synthesis models that uses the idea of a transmission delay line to simulate acoustic wave propagation in the vocal tract by matching the resonance function at the spatial coordinate x into the delay function. The coordinate x in his model can be viewed as an index of each tube in the series of N concatenated tubes. The series of tubes that is used in visualization of the vocal tract modelling is shown in Figure 2.12. In the figure, N is 8 and each of them is an equal length of Δ; hence, for the total length L, the production of ΔN has to be L. In a situation when N is big and Δ is small, the simulation becomes more accurate but at a high computational cost.

In each spatial tube, the simulation concerns velocity and pressure of the travelling wave in terms of u(x,t) and p(x,t), respectively. At the position x at time t, two acoustic states are considered as a summation of their left-travelling and right travelling components with pressures usually denoted as p^r and p^l and velocities as u^r and u^l. Figure 2.13 shows the solution for the travelling wave from [35]. The travelling wave components are assumed to be implemented in the discrete-time T which can represent the state of wave components in every Δ or, in other words, it is used to represent travelling pressure for each of the durations Δ/*c* in the figure where *c* is the speed of sound. Then the pressure at either end is the summation of the leftward- and rightward-travelling components and travels through the pair of digital delay lines as in Figure 2.13(b).

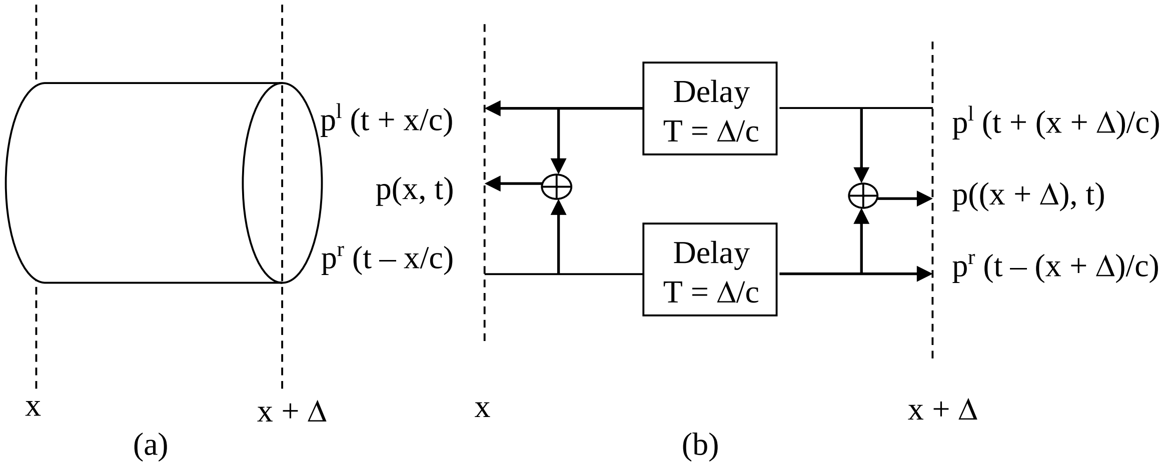


Figure 2.13. (a) An acoustic tube and (b) A representation of the travelling wave solution after Bilbao [35].

Figure 2.13 shows the simulation in a single uniform acoustic tube but the vocal tract model considers the simulation in a series of concatenated tubes with each of them representing a different size of cross-sectional area; hence the acoustic behaviour at the junction has to be discussed. Figure 2.14 (after [35]) clearly depicts a situation at a junction (a) where the areas of tube are different. The subscription *i* is used to index tubes in the series, while A is the area and R is the reflection parameter.

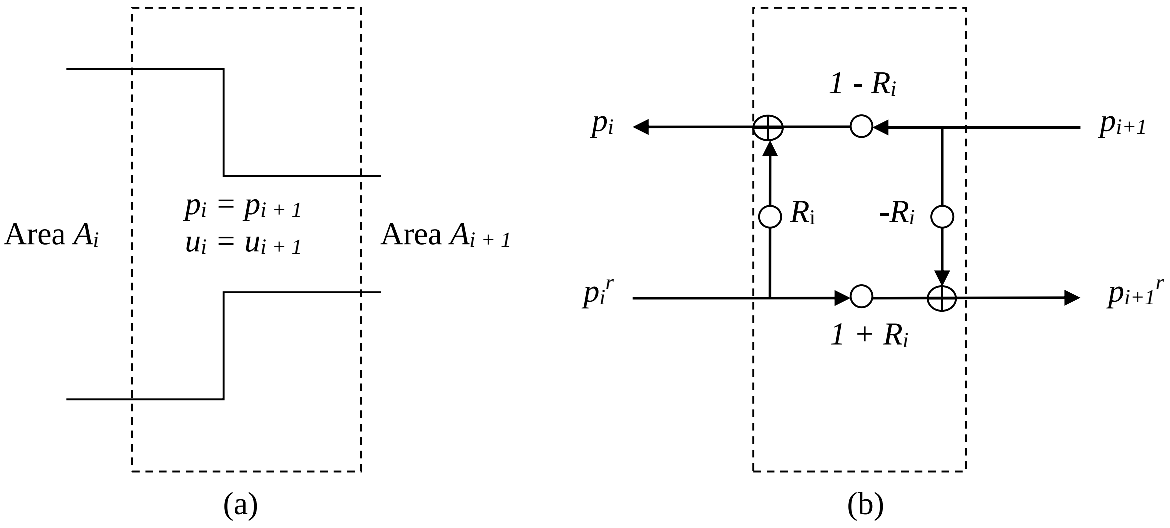


Figure 2.14. (a) The junction between the i^{th} and $(i+1)^{th}$ acoustic tubes in the Kelly-Lochbaum vocal tract model and (b) The resulting scattering junction for pressure waves after Bilbao [35].

Based on the assumption of summation between left- and right-travelling components, the general solution for components at position i can be written as

$$p_i = p_i^l + p_i^r \qquad u_i = Y_i(p_i^l - p_i^r) \qquad 2.18$$

$$p_{i+1} = p_{i+1}^l + p_{i+1}^r \qquad u_{i+1} = Y_{i+1}(p_{i+1}^l - p_{i+1}^r) \qquad 2.19$$

where Y_i is the admittance of the i^{th} tube which is defined by

$$Y_i = A_i/\rho c. \qquad 2.20$$

The equation for Kelly-Lochbaum's model in Figure 2.14 can also be written as the scattering components at a junction.

$$p_i^l = R_i p_i^r + (1-R_i)p_{i+1}^l \qquad p_{i+1}^r = (1+R_i)p_i^r - R_i p_{i+1}^l \qquad 2.21$$

where R_i is

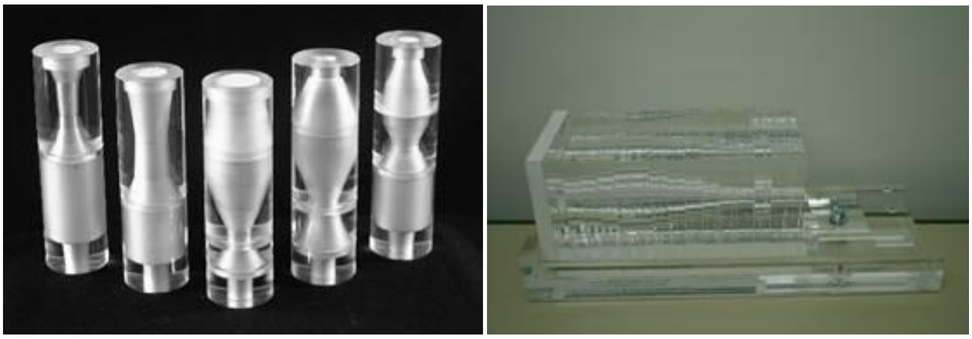
$$R_i = \frac{Y_i - Y_{i+1}}{Y_i + Y_{i+1}} \qquad 2.22$$

The reflection parameter is a level of admittance between two tubes, which means that the left- and right-travelling components are part-reflected and part-transmitted through each junction according to the reflection parameter R. Note that R is [0, 1] which means no reflection when 0 and fully transmitted when 1 [35]. Here, from the above equations, the simulation can then be exited at one end (*i*=0) by periodic and/or non-periodic signal and then pass the signal through the model and radiate the output signal at the other end.

2.7 Vocal tract modelling with some examples

Vocal tract modelling has been discussed in two ways: the parametric and the articulatory models. The direct parametric model bypasses the articulatory model in Figure 2.10 and uses the vocal tract area function (as in Kelly-Lochbaum) in vocal tract simulation and then calculates the corresponding acoustic characteristics. This is very different from the articulatory models which are based on the concept that all parameters physically correlate with human articulatory structures and replicate observed articulatory movement [36].

The parametric vocal tract model is based on the observed vocal tract area. The most straightforward simulation of the tract is to model the tract physically (using resin or acrylic). An example of this straight simulation is Arai’s work. He proposed two types of modelling – the cylinder-type and plate-type models, which are 50 mm diameter sculpted acrylic cylinders as in Figure 2.15 (a) and 10 mm radius curve in a step-wise fashion as in Figure 2.15 (b), respectively [37].



(a) (b)

Figure 2.15. Arai vocal tract models from [37].

The step-wise modelling in Figure 2.15 (b) uses sets of concatenated 10 mm radius acrylic plate to model the vocal tract. In this step-wise fashion, each acrylic plate is 75 mm x

75 mm x 10 mm with a hole in the centre. The model is set by placing appropriate plates side-by-side. The holes in the plates then form a tube.

Another example of the articulatory model is the articulatory synthesizer included within Praat [38]. The system is based on Functional Phonology [39], which makes use of the actions of relevant muscles whose function is to vary the shape of the vocal tract. Together with timing control, the tensions of these muscles become input variables to the system that control sequentially the movement of the articulators.

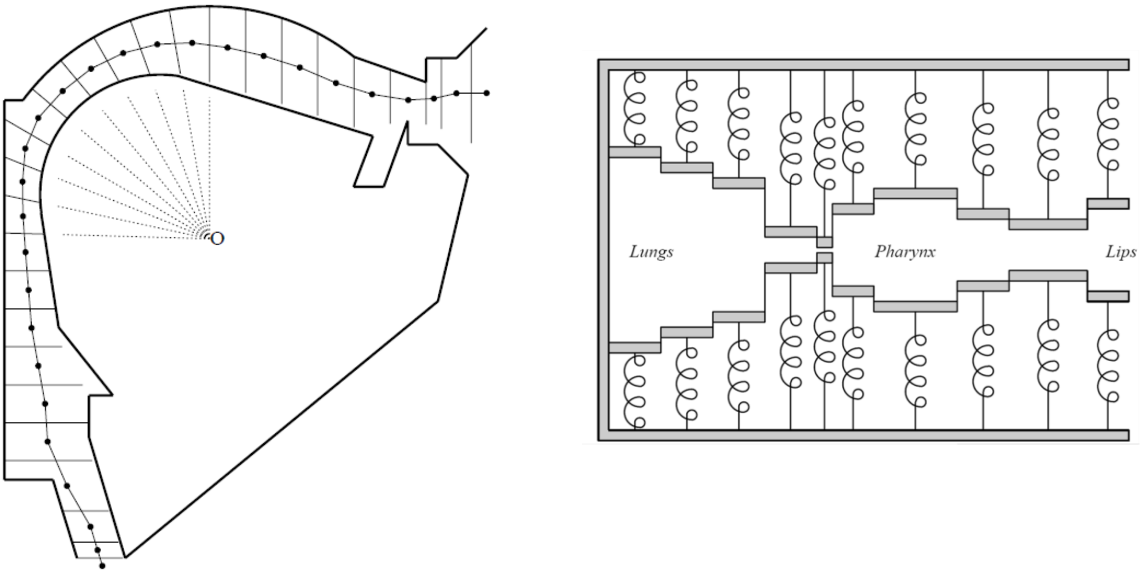
The algorithm in Praat articulatory speech synthesizer was introduced in 1993 by Paul Boersma in Berlin 1993 [40].

In the Praat model, there are 27 concatenated tubes with flexible walls, time-varying lengths of tract regions but fixed length in the sub glottal region. The glottis consists of two tubes. The mesh points are then at the central cross sections of the tubes in Figure 2.16.

Boersma simulates the physics of air particles through the momentum density *p* (kg/m2s), the mass line density *e* (kg/m), mass flow *J* (kg/s), the continuous pressure *Q* (N/m2) and the resistance of Hagen-Poiseuille *R* in the lungs, bronchi, trachea, glottis and vocal tract. The springiness at the walls are considered in terms of the mass of wall *m*, the mean excess pressure in tube *p*, the spring force *F* and the tissue stiffness. All equations are shown in the paper [38] with updating steps through the simulation and more details are given in Chapter 3 Section 3.12 of [39]. With the inclusion of all of these physical properties, his model can generate a glottal source and also noise turbulence anywhere in the vocal tract. Table 2.1 shows a system comparison between Boersma’s Praat articulatory synthesis and Mullen’s 2-D DWM articulatory synthesis.

Table 2.1. Comparison between using 2D-DWM and that in Praat.

	Boersma’s Praat	**Mullen’s 2-D DWM**
Modelling	Model the momentum density, the mass line density, the mass flow and the continuous pressure	Model the pressure
Sound source generator	Simulate glottal source and noise turbulent generator	Use external sound source; currently the LF model
Meshing	Concatenated tubes - based on Mermelstein [40]	Rectilinear (four ports connected at a junction)



(a) (b)

Figure 2.16. (a) Twenty seven concatenated tubes for the pharyngeal and oral cavities and (b) the simplified mid-sagittal view of our model of the speech apparatus (not drawn to scale) of Praat from [39].

Praat mathematically uses finite difference to implement the time varying of the aerodynamics and myoelastics of the concatenated tubes ([39] page 91). The mass *e*, the momentum *p*, the mass flow *J* and the continuous pressure *Q* are derived step by step in Chapter 3 Section 3.12 of [39] and give the final output in Pa (N/m^2) at 40 centimetres from the head as

$$sound(t) = \frac{4\pi}{0.4}\left(\sum\nolimits_{M=nose,lip}\frac{J_{M+}^{n} - J_{M+}^{n-1}}{\Delta t} + \sum\nolimits_{m=every\ tube} 1000\ \rho_0 \Delta x_m^n\ \Delta z_m^n \Delta y_m^{n-\frac{1}{2}}\right) \qquad 2.25$$

where Boersma claims the novelties of his Praat synthesizer to be:

- The entire speech apparatus is modeled in the same way.
- Tube lengths can vary as functions of time, so that we can faithfully model speech sounds that crucially depend on longitudinal movements.
- The meshing algorithm is resistant to the wildest movements.

In 2009, Professor Atsuo Takanishi of Waseda University presented his WT-7RII Waseda Talker No. 7 Refined II. It is a robot talker that has a human-like speech production mechanism from lung to lips including nasal cavity. Professor Takanishi claims that it can produce sounds with similar acoustic characteristics to its adult male model [41].

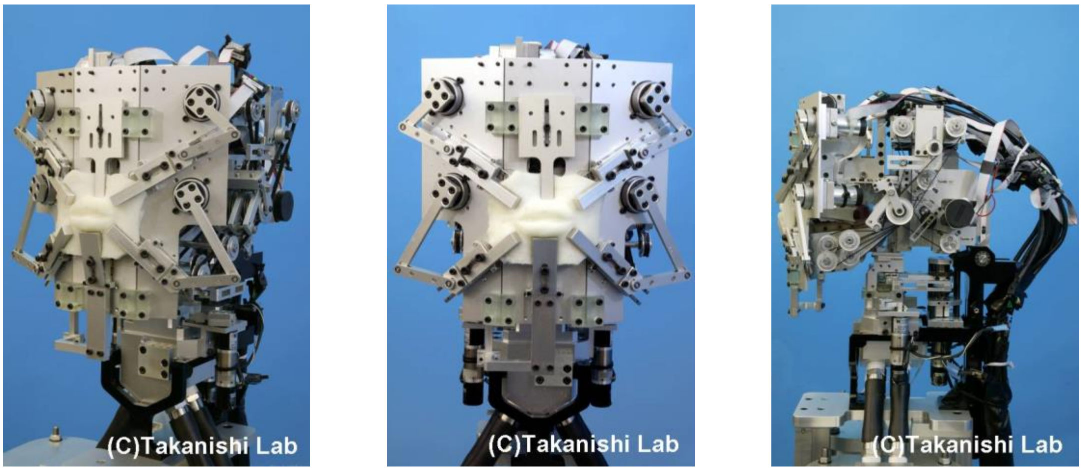
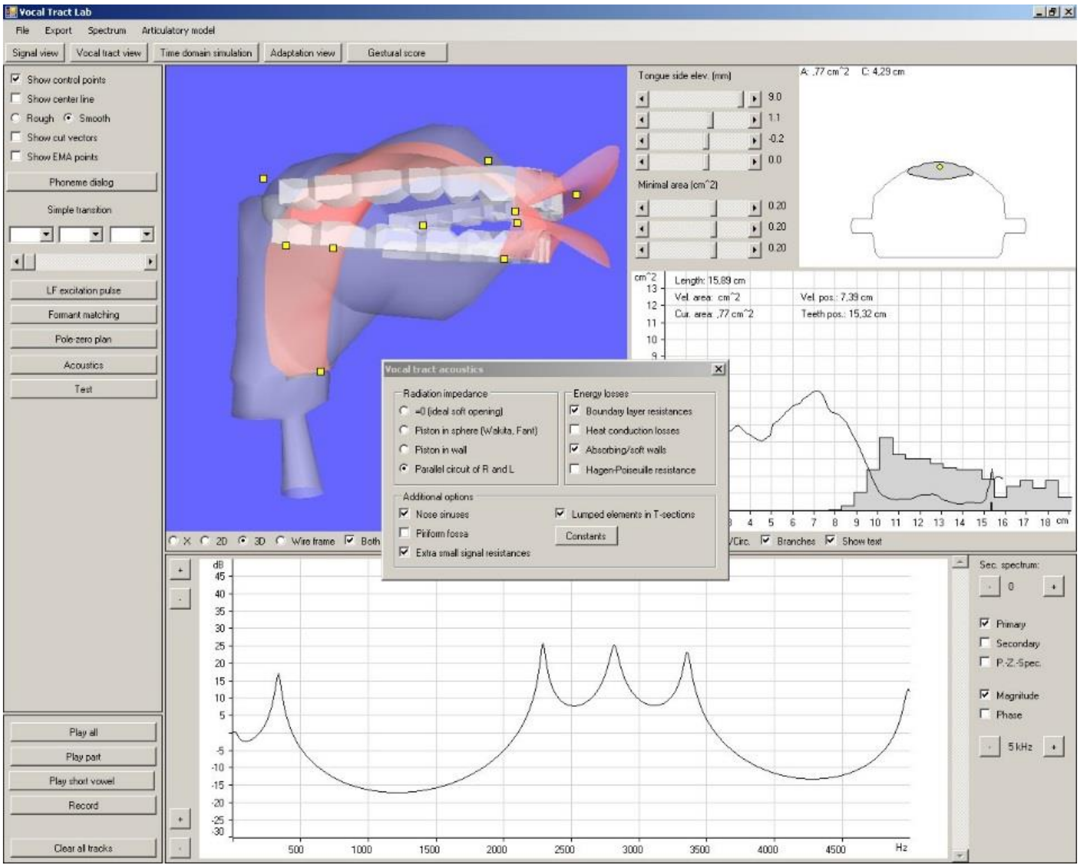


Figure 2.17. Waseda Talker from [41]

The Takanishi lab research team developed the anthropomorphic talking robot Waseda Talker with the aim of addressing the difficulty of simulating articulation through a combination of speech science and humanoid robot technology. They modelled the vocal cords and vocal tract in three-dimensional models and claimed that they could produce more human-like speech production at that time (2009). Its mechanical form contains all speech production models from lung through to the lips and nasal cavity [41]. The size is the same as an adult male model (no details of the model were provided). They adopted EMA data to control the articulation when producing continuous speech. However, they mentioned the limitation of motors in the robot in that they cannot work fast enough to capture all human speech production. A video recording of its movement and synthesized sound can be found in [41].

In September 2013, Peter Birkholz published his updated version of VocalTractLab v 2.1, a vocal tract simulator, on www.vocaltractlab.de. His software tool is currently available free of charge. It can demonstrate the vocal tract mechanism in speech production in a three-dimensional model. The model simulates the surfaces of articulators and vocal tract walls together with their interaction with both volume flow and pressure distribution. The software is very flexible and lets users control parameters themselves, such as vocal tract control points with panels showing graphs of the volume velocity transfer function, vocal tract input impedance and spectrum of the glottal source. The software can also show the animation of the acoustic simulation (similarly to Praat but with more nasal details). The timing scale parameter control in the software makes it capable of synthesizing an utterance (multiple phones). There are more features allowing the user to control the synthesizer which can be found in the software manual [42].


☑ Show control points
□ Show center line
○ Rough ● Smooth
□ Show cut vectors
□ Show EMA points

Tongue side elev. (mm): 9.0, 1.1, -0.2, 0.0
Minimal area (cm^2): 0.20, 0.20, 0.20

A: ,77 cm^2 C: 4,29 cm

Length: 15,89 cm
Vel. area: cm^2
Cur. area: ,77 cm^2
Vel. pos.: 7,39 cm
Teeth pos.: 15,32 cm

○ X ○ 2D ● 3D ○ Wire frame ☑ Both

/Circ. ☑ Branches ☑ Show text

Vocal tract acoustics

Radiation impedance
○ =0 (ideal soft opening)
○ Piston in sphere (Wakita, Fant)
○ Piston in wall
● Parallel circuit of R and L

Energy losses
☑ Boundary layer resistances
□ Heat conduction losses
☑ Absorbing/soft walls
□ Hagen-Poiseuille resistance

Additional options
☑ Nose sinuses
□ Piriform fossa
☑ Extra small signal resistances
☑ Lumped elements in T-sections

Sec. spectrum: 0
☑ Primary
□ Secondary
□ P.-Z.-Spec.
☑ Magnitude
□ Phase
5 kHz

Figure 2.18. Screenshot example of VocalTractLab v.1 from [42].

Chapter 3

2-D Digital Waveguide Mesh

In the first two chapters the general concept of speech synthesis with some successful examples was described. In this chapter the specific methodology used in this research, a physical wave model known as the digital waveguide mesh (DWM) is described.

3.1 Digital Waveguide Mesh (DWM)

The waveguide mesh is one of the simulation techniques that use the scattering principle to solve a set of time-dependent partial differential equations (PDEs) as described in [35]. It implements the finite-difference approximation to fit a numerical grid followed by recursion in a specific set of initial and boundary conditions with possible external excitations. Bilbao also comments on the great general benefit of the meshing system which is the network formulation that allows direct access to a measure of the system's energy carried by waves on a large network of lumped elements.

Twelve years before Bilbao, Julius O. Smith III described the digital waveguide method in [43, 44] as a way of avoiding a high computational cost because it does not need multiplication at each grid point in space. In the waveguide, each travelling wave component arises from solving the wave equation in a medium. For example, in a model for a stringed instrument, the travelling wave component travels along a string to the left or right at a speed c which is equal to

$$c = \sqrt{\frac{K}{\epsilon}} \qquad 3.1$$

when K is defined as string tension and ϵ is defined as linear mass density. The well-known d'Alembert solution for the wave equation (first published by Jean le Rond d'Alembert in 1747) was preferred to solve the travelling wave equation as

$$y(x,t) = y^r(x - ct) + y^l(x + ct) \qquad 3.2$$

for arbitrary functions y^r and $y^l$of a coordination *x* and *t* where *c* is a constant. The equation 3.2 is then the solution proven by twice differentiating with respect to *x* and *t.*

$$y_x = y^{r\prime}(x - ct) + y^{l\prime}(x + ct)$$

$$y_{xx} = y^{r\prime\prime}(x - ct) + y^{l\prime\prime}(x + ct)$$

and

$$y_t = -cy^{r\prime}(x - ct) + cy^{l\prime}(x + ct)$$

$$y_{tt} = c^2y^{r\prime\prime}(x - ct) + c^2y^{l\prime\prime}(x + ct)$$

hence

$$y_{tt} = c^2y_{xx}. \qquad 3.3$$

In the digital domain, the d'Alembert solution has to be considered on a sample-by-sample basis to represent the travelling wave. Amplitude at a time instant is sampled every T seconds. The sampling period T is related to the sampling frequency f by

$$f = \frac{1}{T} \qquad 3.4$$

This means that this digital system contains *f* samples per second. Meanwhile, in the view of a travelling object, for a temporal sampling interval *T*, sound could travel *X* metres.

$$X = cT \qquad 3.5$$

For example, in air at a room temperature of 20 degrees Celsius where the speed of sound *c* is

$$c = 331 + 0.6Tem \qquad 3.6$$

in which *Tem* is the temperature in degrees Celsius, we then get 343 m/s for sound components that are travelling in a space. In this case, if we consider a sample of sound for a system operating at a sampling rate of 44100 Hz, the spatial sampling interval will be equal to 343/44100 = 0.0077 meters or 7.77 millimetres or a spatial sampling rate at 128 samples per metre. That means we need 128 delays to simulate the travelling sound wave for a metre in a single one-way delay line. Therefore, a total of 256 delays is needed for the bi-directional delay lines.

In a simulation, the travelling waves are referred to as left and right according to their direction. Some texts use superscript “l” and “r” to distinguish them, while others use “-” and “+”. In this research I will use “-” and “+”. To get an output from a position *x* at time *t* in the delay line, both directions of the travelling waves are summed up as

$$y(x,t) = y^{+}\left(t - \frac{x}{c}\right) + \ y^{-}\left(t + \frac{x}{c}\right) \qquad 3.7$$

In a diagram view, the two delay lines can be drawn as upper and lower rails, as shown in Figure 3.1. Pairs of delay lines represent samples at a position x.

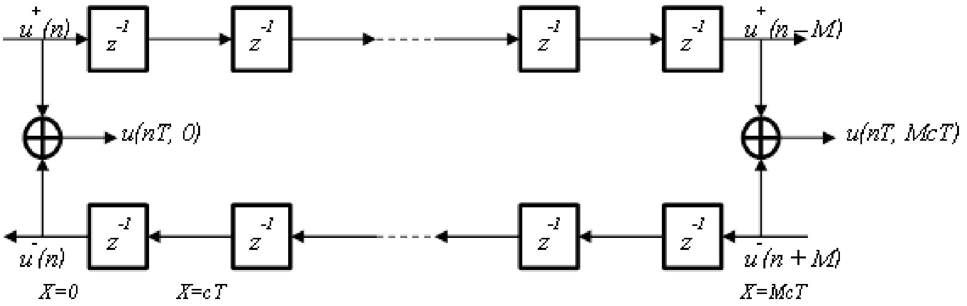


Figure 3.1. Bi-directional digital delay line diagram from [45].

The bi-directional delay line represents the ideal lossless one-dimensional waveguide. The simulation is band limited to half of the sampling frequency (Nyquist frequency).

From the point of view of tube modelling, each tube is *cT* metres long and *McT* metres in total. Each transmission pair of the travelling waves are considered as input and output to their neighbours. The Kelly-Lochbaum digital speech synthesis model [34] established the first successful one-dimensional acoustic modelling using this technique for considering variations of cross-sectional area parameters along the vocal tract shape. This idea of modelling is depicted at Figure 3.2.

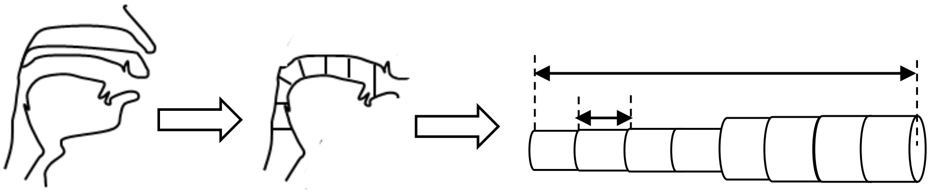


Figure 3.2. Example of concatenated tubes for Kelly-Lochbaum vocal tract modelling.

The acoustic behaviour in each individual tube is considered in terms of a volume velocity *u(x,t)* and pressure deviation *p(x,t)*. The cross-sectional area *A* is involved in the acoustic state as in equations 3.8 and 3.9.

$$\frac{\rho}{A}\frac{\partial u}{\partial t}+\frac{\partial p}{\partial x}=0 \qquad 3.8$$

$$\frac{A}{\rho c^2}\frac{\partial p}{\partial t}+\frac{\partial u}{\partial x}=0 \qquad 3.9$$

where *ρ* is defined as the air density. Here, the pressure scales the volume velocity by the tube admittance *Y* which is

$$Y\equiv\frac{A}{\rho c} \qquad 3.10$$

Therefore, the composition of pressure *p(x,t)* could be reconsidered as

$$p(x,t)=p^{+}\left(t-\frac{x}{c}\right)+p^{-}\left(t+\frac{x}{c}\right) \qquad 3.11$$

and those of the volume velocity as

$$u(x,t)=\ -Yp^{+}\left(t-\frac{x}{c}\right)+Yp^{-}\left(t+\frac{x}{c}\right) \qquad 3.12$$

Along the waveguide, the conservation of mass is applied at every junction to describe the continuity of the scattering equation. In the acoustic analogue view, Kirchhoff's laws for a parallel connection are applied from 1 to *M* element.

$$p_1=p_2=\cdots=p_M=p_J \qquad 3.13$$

$$u_1+u_2+\cdots+u_M=0 \qquad 3.14$$

3.2 Multi-Dimensional Digital Waveguide Mesh

In 1993, Van Duyne and Smith [46] modelled membranes and plates using the digital waveguide mesh. The model propagated the travelling wave through nodes (delay units) and scattered them in various directions depended on the number of concatenating ports or topology. The efficiency of interesting topologies has been discussed in [47]. In general, all travelling wave components are considered as incoming and outgoing components to and from a scattering node or junction. To notify the direction, superscript ”+” denotes the incoming and “-“ denotes the outgoing travelling components. At a junction the outgoing is calculated by

$$p_i^- = p_J - p_i^+ \qquad 3.15$$

where *i* is the concatenating port index and *J* is the node index. Meanwhile a junction pressure p_J is calculated by

$$p_J = \frac{2\sum_{i=1}^{n}\frac{p_i^+}{Z_i}}{\sum_{i=1}^{n}\frac{1}{Z_i}} \qquad 3.16$$

where *n* denotes the number of ports per node. The topologies have been named after the number of the concatenating port – for example, rectilinear (*n* = 4), hexagonal (*n* = 3) or triangular (*n* = 6).

In the time domain, the wave components travel through the mesh by time step. At every step *T* the outgoing components move towards their neighbour. The stepping also means that the components have been travelling for a distance. The inter-nodal distance or spatial sampling distance interval *d* is calculated by

$$d = \frac{c\sqrt{N}}{f} \qquad 3.17$$

where *N* is the number of dimensions and *f* is the sampling frequency which is equal to 1/*T*.

In 2-D, we construct a mesh with termination at a border that has to be pre-identified. The easiest design is to put single port nodes there; then we can have reflection coefficients to control the border conditions, such as an open or rigid end. The coefficient can then be identified by a ratio of the difference between the local admittance and the boundary; therefore it is bounded in between -1 and 1.

$$r = \frac{Y - Y_B}{Y + Y_B} \qquad 3.18$$

Increasing the number of dimensions makes the simulation lose computational efficiency, since it requires more multiplication and division at every node and also contains some dispersion error which strongly depends on meshing topology [47]. However, the rectilinear mesh is the most commonly used for 2-D modelling, since it is easy to fit into a space and the indexing is straightforward in practice; it is therefore used in this thesis. The boundary flexibility will be discussed in the following section.

3.3 2-D Digital Waveguide Mesh

This section describes the 2-D mesh in terms of its construction and modification. To construct the mesh in 2-D, topology has a direct effect on the model's efficiency but this research focuses only on the rectilinear. Therefore, all details from this point relate to the rectilinear meshing. The others are examined in [46, 47, 48].

3.3.1 Meshing and Scattering

The construction of a 2-D rectilinear mesh makes a four-port connection to a node. The ports are labelled north, south, east and west. Figure 3.3 shows the construction in which nodes or delay units align in rows and columns together with one-port terminator nodes at the borders.

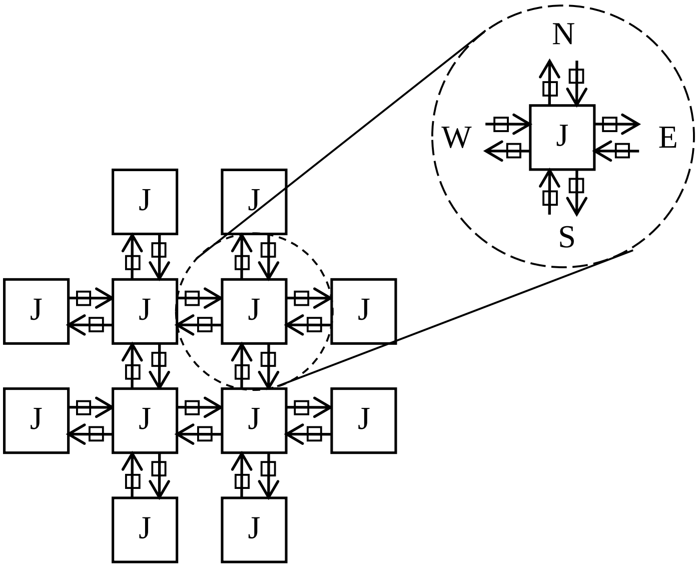


Figure 3.3. Construction of rectilinear 2-D DWM.

The wave propagation is simulated by equations 3.15 and 3.16. The scattering goes through nodes which are called junctions. The acoustic wave components scatter to and from the junctions. In the simulation that runs in a homogeneous medium the impedances can be set as a constant and the equation 3.16 can be simplified to

$$p_J = \frac{2}{n}\sum_{i=1}^{n} p_i^+ \qquad 3.19$$

This could be found in a room acoustic simulation or a single pipe organ. Computation efficiencies can be found in [47] for all topologies.

In terms of geometry, the impedance Z concerns the resistance between tension and mass displacement in the medium which is calculated from density ρ and Young's modulus of elasticity which can be calculated by the speed of sound and the density. Considering the tension and mass displacement in tubes or concatenated tubes for the vocal tract simulation, the impedance can be written as

$$Z_{tube} = 1/Y_{tube} = \rho c/A_{tube} \qquad 3.20$$

Equation 3.20 shows that Z is used in terms of its inverse – acoustic admittance Y – which adds another meaning to equation 3.10. According to our proposal, the waves travel through a variation of concatenated tubes. The cross-sectional area function has become the main concern.

Looking back to Figure 3.2, it depicts a variation of cross-sectional area functions. The diameter of each tube is represented in a function of the width $W(x)$ across the y-axis of the mesh which is a measure of cross-sectional area described by the 1-D area function; therefore the width of the tract is proportional to r, the radius of the equivalent cylinder in the 1-D model.

$$\begin{aligned} W(x) &= 2\sqrt{\frac{A(x)}{\pi}} \\ &= 2r \end{aligned} \qquad 3.21$$

From another point of view, the area function is cylindrically related to the power of r. Mullen, 2006 [49], considered area or sphere rather than radius or diameter. The following equation demonstrates his consideration of circle equation as the width; then the width is proportional to r^2.

$$W(x) = A(x) = \pi r^2 \qquad 3.22$$

Translating the area function into the distance across the mesh will always place restrictions on the minimum width allowed. In any narrow channel, its width must be at least two waveguides across, because in 2-D mesh a central line needs two attached boundary junctions on either side in the narrowest construction. The narrowest can also be a complete stop for plosives. The minimum two junctions can possibly work as a stop by increasing their admittance parameters (to stop the transmitting). Then an increase of those parameters can be given to release the burst. This practical dynamic control clarifies the obstructing ability in 2-D DWM (an example of synthesized sound can be found in [5]).

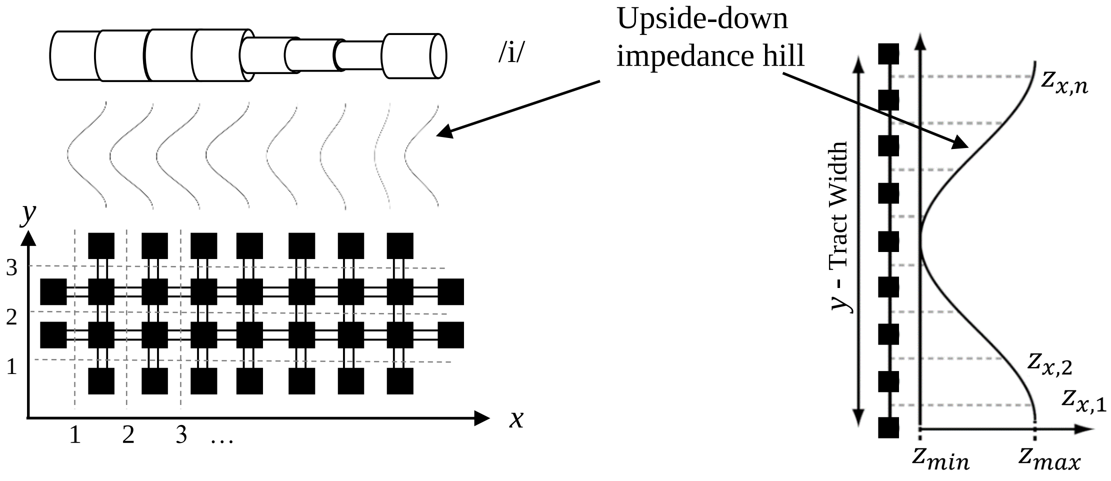


Figure 3.4. Raised impedance hills causing a constriction in a straight tube and plotted raised cosine impedance hills.

Figure 3.4 shows raised impedance on the edges of the straight tube. The upside-down bell-shape contours represent the impedance hill where Z_{max} is the maximum impedance value which is calculated from the cross-sectional area at *x*. The impedance value at the point of constriction *x* is separated into $Z_{x,1 \dots n}$ for different y positions. The maximum impedance value of a constriction *x*, Z_{max}, is at $Z_{x,1}$ and $Z_{x,n}$ and the minimum Z_{min} is at $Z_{x,n/2}$ which is called Z_{tube} in Mullen's work. The implementation of a cosine smoothing hill for *Z(x,y)* is then defined as

$$Z(x,y) = Z_x - \frac{(Z_x - Z_{min})}{2}\left[1 + \cos(2\pi(\frac{y}{w} - \frac{1}{2}))\right] \qquad 3.23$$

The highest impedance is called Z_{stop}. This is the impedance for a modelling of a complete cut-off of the air-flow [49].

3.3.2 Boundary Management in vocal tract modelling

Vocal tract widths are used to bound the 2-D mesh. The bounding in this sense is about local area function concerns, rather than reflection. A straight tube itself has a frequency related to the length (1-D), but the constricted tube has a frequency related to the size and position of the resonant cavities (2-D); therefore, the accuracy of the modelling of the constriction area affects the frequency response or the movement of formants.

Human vocal tract shapes are normally captured during speech production according to articulator movements. The capturing is done by image technologies such as MRI, CT or X-ray etc. Figure 3.5 is captured from [49] which perfectly illustrates an idea of the process of 2-D spatial sampling into a 2-D rectilinear waveguide mesh modelling from the 1-D area function of a static /i/ vowel shape.

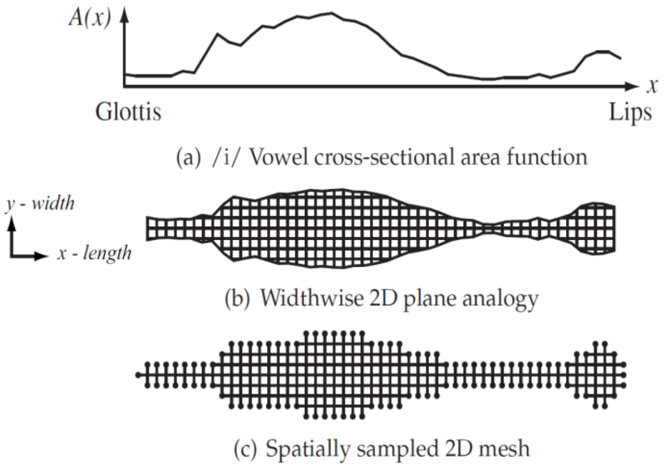

A(x)
x
Glottis
Lips
(a) /i/ Vowel cross-sectional area function
y - width
x - length
(b) Widthwise 2D plane analogy
(c) Spatially sampled 2D mesh

Figure 3.5. The 2-D widthwise /i/ vowel waveguide model from Mullen [49].

Figure 3.5 shows 2-D meshing in [49]. A rectilinear meshing with one-port junctions at boundaries is performed from a cross-sectional area function. The diffusion at the boundaries in 2-D DWM has been studied in [47, 48, 49, 50], yet these studies did not simulate for rough boundaries in a small space in a low sampling frequency. Here, in consideration of the simulation for air in small ducts, the diffusion has been considered as constants of reflection at the walls. In general, as mentioned in equation 3.18, the reflection has been considered as in a form of ratio between local and boundary admittance. For 2-D they were considered as walls, lips and glottal (r_{wall}, r_{lip} and $r_{glottal}$ respectively).

In a narrow view of human speech simulation, the wall reflection coefficient affects the formant bandwidth but the lip and glottis reflection coefficients do not [49]. In [51], experiments in wall reflection coefficients were conducted. The results support [49] the claim that the higher the wall reflection coefficient, the lower the resulting formant bandwidth. A small formant bandwidth leads to a greater potential distinction between different vowels by reducing the overlap between adjacent formants when they are close in frequency.

In detail, the wall or the area data move or change in order to understand the meaning of using a variety of the coefficients in movements. Hence, the effect of the wall reflection coefficient to English diphthongs was explored. The system is then set to test the effect of movement by simulating the wave propagation using white noise as the sound source to allow tracking of all the changing resonant frequencies. The reflection coefficients are varied as follows: 0.90, 0.92, 0.94, 0.96, 0.98 and 1.0. Figure 3.6 shows formant bandwidths of eight synthesized diphthongs overlaid using a Hamming window with the following parameters: length 0.049 seconds, 0.7 pre-emphasis, 0.01 second of a frame interval and LPC order 12. The data presented in the figure are the average of the formant bandwidths when the tract is changing its size.

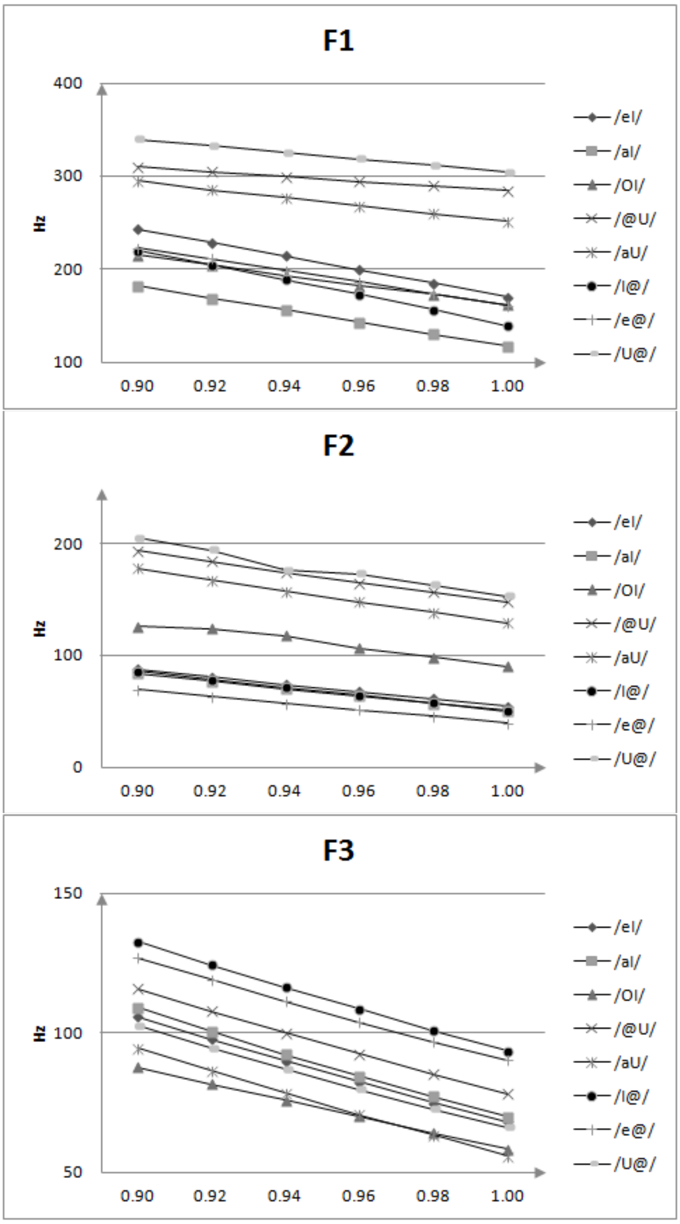
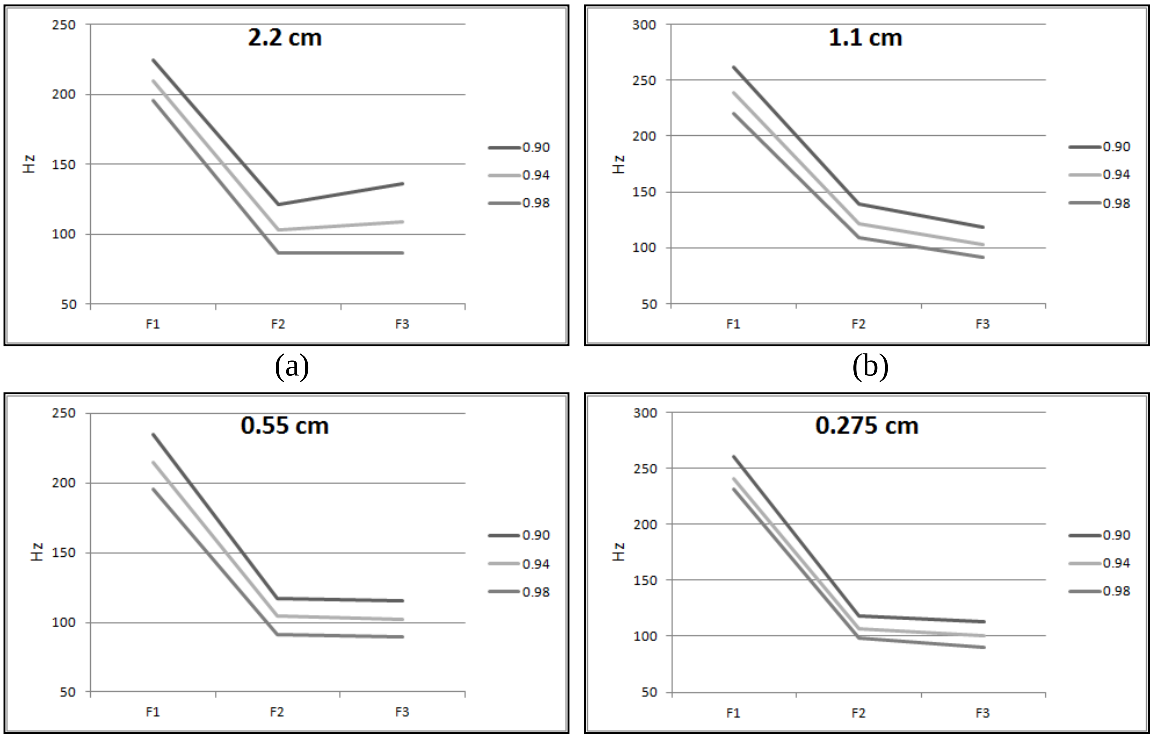


Figure 3.6. The formant bandwidth of eight synthesized English diphthongs using various wall reflection coefficients (0.90, 0.92, 0.94, 0.96, 0.98 and 1.0).

Figure 3.6 shows the results of speech analysis for the bandwidths of F1, F2 and F3. They are decreasing when the coefficient is getting slightly higher. This means that the damping in the time domain or the absorption of the sound energy by the moving boundaries could also be controlled by fixing the wall reflection coefficient. Meanwhile, the distinctiveness of each monophthong end-point component in each diphthong can be reduced by increasing the formant bandwidth or decreasing the coefficient.

The effect of changing waveguide size on the bandwidth was explored in order to understand the effect of using a more dense mesh. The results show the same trend of decreasing of the bandwidth when the reflection coefficient is increased in all waveguide sizes tested. Figure 5 shows the analysis results overlaid using 0.049 seconds of Hamming window length with 0.7 pre-emphasis, 0.01 second of a frame interval and LPC order 12 on the frequency analysis.



(a) (b)

(c) (d)

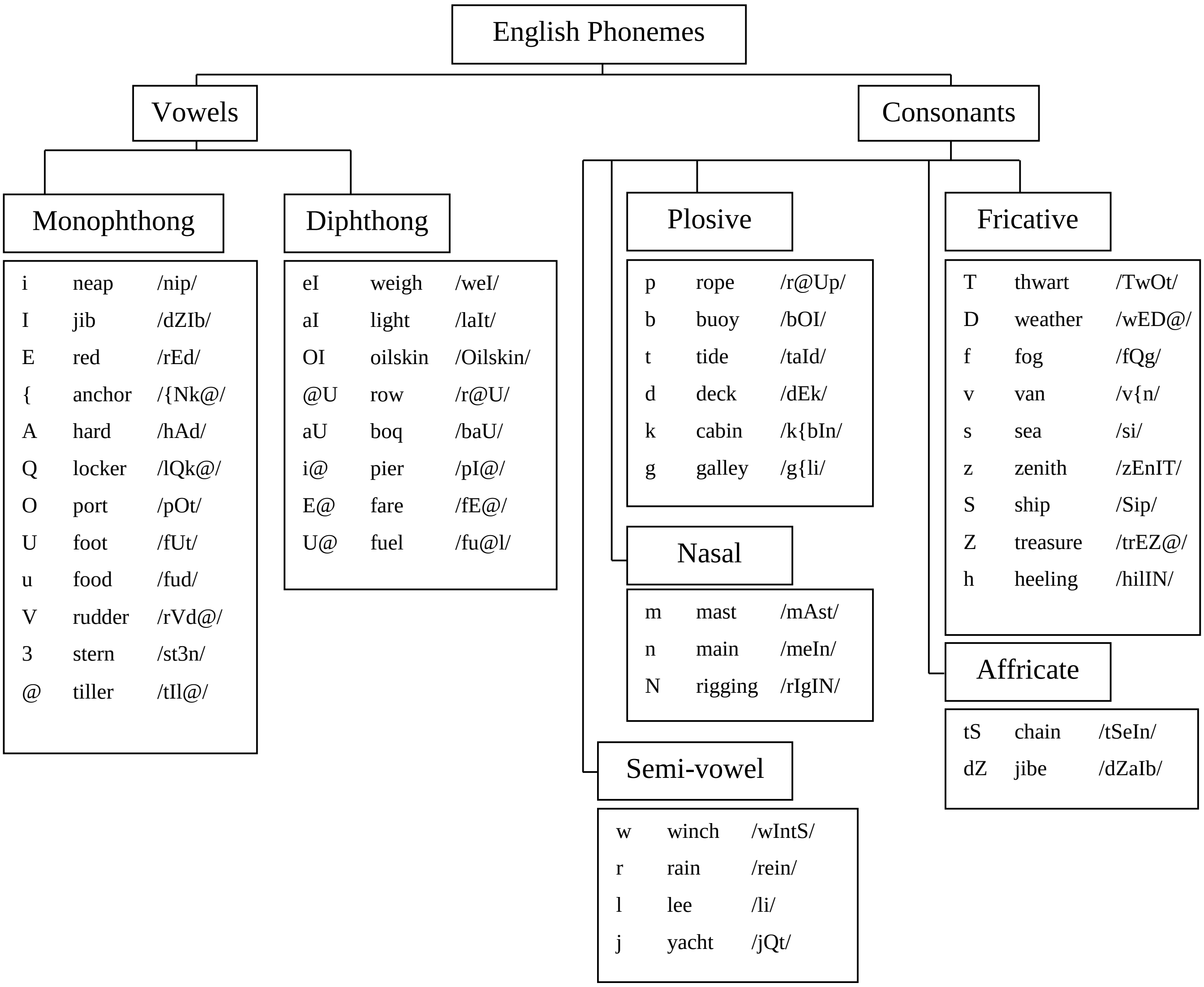
Figure 3.7. Formant bandwidth results from different waveguide sizes using wall reflection coefficients of 0.90, 0.94 and 0.98. (a) waveguide size = 2.2 cm, (b) waveguide size = 1.1 cm (c) waveguide size = 0.55 cm (d) waveguide size = 0.275 cm.

All in all, to manipulate a synthesizer when producing English using 2-D DWM, the junction pressure and reflection coefficient play the main roles. The vocal tract modelling for English consonants needs tract movement in articulation. This study will apply the rectilinear 2-D meshing with additional functions to study the movement of articulators in the synthesis.

Chapter 4
Acoustics of
English Consonants

The English, language is formed as mixed 24 consonants and 20 vowels sounds, concatenated and pronounced in spoken utterance. They are phonetically classified in terms of their voice, manner and place of articulation. This chapter gathers acoustic characteristics from linguistic and voice science literature to depict the desired target characteristics in the outputs, with some spectrograms visualizing those characteristics. The consonants are transcribed in the SAMPA (Speech Assessment Methodologies Phonetic Alphabet) transcription of John Wells, 1989. Figure 4.1 shows them in groups according to their manner of articulation. This research uses SAMPA rather than the IPA (International Phonetics Association) alphabet, since SAMPA uses ASCII characters and is therefore unambiguous as a computer font representation.

In each manner of articulation, phone members have similar acoustic characteristics which will be used as the target acoustic output in this research. For the sake of clarity, this chapter is divided into sections. Section 4.1 describes those acoustic characteristics which can be found in human speech sound, while section 4.2 describes the theory of perturbation which explains how the resonance frequencies change with articulation. Details of the vocal apparatus are given in section 4.3 and the acoustic characteristics that can be found in each manner of articulation are described in detail in section 4.4.



Monophthong

i	neap	/nip/
I	jib	/dZIb/
E	red	/rEd/
{	anchor	/{Nk@/
A	hard	/hAd/
Q	locker	/lQk@/
O	port	/pOt/
U	foot	/fUt/
u	food	/fud/
V	rudder	/rVd@/
3	stern	/st3n/
@	tiller	/tIl@/

Diphthong

eI	weigh	/weI/
aI	light	/laIt/
OI	oilskin	/Oilskin/
@U	row	/r@U/
aU	boq	/baU/
i@	pier	/pI@/
E@	fare	/fE@/
U@	fuel	/fu@l/

Plosive

p	rope	/r@Up/
b	buoy	/bOI/
t	tide	/taId/
d	deck	/dEk/
k	cabin	/k{bIn/
g	galley	/g{li/

Nasal

m	mast	/mAst/
n	main	/meIn/
N	rigging	/rIgIN/

Semi-vowel

w	winch	/wIntS/
r	rain	/rein/
l	lee	/li/
j	yacht	/jQt/

Fricative

T	thwart	/TwOt/
D	weather	/wED@/
f	fog	/fQg/
v	van	/v{n/
s	sea	/si/
z	zenith	/zEnIT/
S	ship	/Sip/
Z	treasure	/trEZ@/
h	heeling	/hilIN/

Affricate

tS	chain	/tSeIn/
dZ	jibe	/dZaIb/

Figure 4.1. English phonemes using the SAMPA transcription after [20].

4.1 Acoustic Representation in the Human Voice

The human voice is usually measured according to three main characteristics – loudness, frequencies and timbre [52]. Loudness is perceived from the size of pressure variation, pressure amplitude usually being calculated in decibels (dB). The frequencies are perceived in the ear at the cochlea which is about 3.2 cm in length but responds to a 20 – 20,000 Hz range of frequencies. The combination of frequencies can be graphically plotted in a spectrogram which also shows the relative amplitude of each frequency; we can then observe the overtone frequencies easily. In [52], Ladefoged called these overtones the formants. He also mentioned that the appearance of formants comprises the major acoustic components of speech which are counted in order from low to high frequency. Table 4.1 shows the correlation with their auditory correlate from his book [53].

Table 4.1. The correlation of the appearance of formants and their auditory correlates after [53].

Acoustic variable	**Auditory correlate**
Frequency of 1st formant	First natural mode of resonance of the vocal tract
Frequency of 2nd formant	Second natural mode of resonance of the vocal tract
Frequency of 3rd formant	Third natural mode of resonance of the vocal tract
Amplitude of 1st formant	Loudness of the first formant
Amplitude of 2nd formant	Loudness of the second formant
Amplitude of 3rd formant	Loudness of the third formant
Centre frequency of the semi-random noise	Pitch of the voiceless components
Amplitude of the semi-random noise	Loudness of the voiceless components
Fundamental frequency of voiced sounds	Lowest natural mode of the voice

The formants are sometimes known as resonances of the vocal tract. To put it simply, the longer the vocal tract the lower the resonant frequencies. Shaping the tract obstructs the node or antinode of resonance which then affects the formant frequencies. Perturbation theory describes the effects of the disturbance.

4.2 Perturbation Theory

Constriction or obstruction of any part of the vocal tract causes changes in the acoustic output, depending on where it occurs and its extent. These effects theoretically correspond to changes in the resonant frequencies in the vocal tract. The first three formants vary most markedly during vowels (higher formants exhibit little variation), thereby characterizing the acoustics of an individual vowel. The first and second resonances are in the same situation as the one-sided open tube in Figure 4.2, and the third resonance occurs in the same way following at 5/4 of the wavelength. For example, at room temperature the resonation of the first three formants in an average male vocal tract for a relaxed neutral vowel which has a length of approximately 17.5 cm are at

$$\mathrm{F1} = \frac{1}{4}\left(\frac{c}{L}\right) = 491 \text{ Hz}$$

$$\mathrm{F2} = \frac{3}{4}\left(\frac{c}{L}\right) = 1{,}474 \text{ Hz}$$

$$\mathrm{F3} = \frac{5}{4}\left(\frac{c}{L}\right) = 2{,}457 \text{ Hz}$$

These are the formants for the neutral schwa vowel /@/ that does not have any constriction or expansion along the vocal tract. In perturbation theory, nodes and antinodes of the waveform are considered. An antinode is the place where the pressure various between a maximum and minimum value, such as at the glottis, while a node is the place where the pressure does not vary. Places of constriction and expansion will affect where nodes and antinodes of each formant are, and the changes occur according to the following principles from [20]:

- A constriction near a pressure node decreases that formant’s frequency.
- A constriction near a pressure antinode increases that formant’s frequency.
- Lengthening the vocal tract decreases all formant frequencies.
- Shortening the vocal tract increases all formant frequencies.

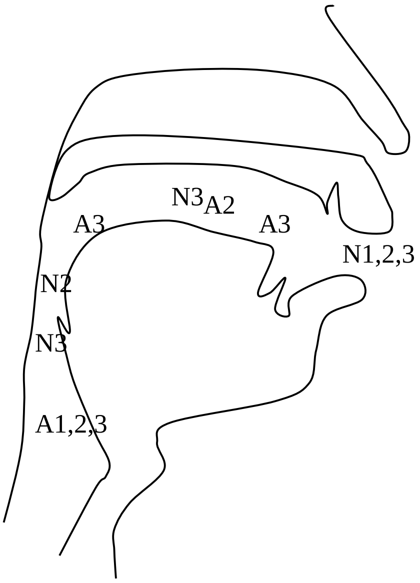


Figure 4.2. Approximate places of nodes and antinodes in the vocal tract from [20].

4.3 Vocal Tract Apparatus and Articulation

The human voice results from the vocal sound propagation system which involves several muscles around the lungs, trachea, glottis, pharynx, velum and articulators (jaw, lips, palate, tongue and teeth) in shaping each part of the vocal system. The figure below shows the approximate locations of the organs in cross section in the upper part of the human body (from Holmes [12]) and the lungs (from Howard [20]).

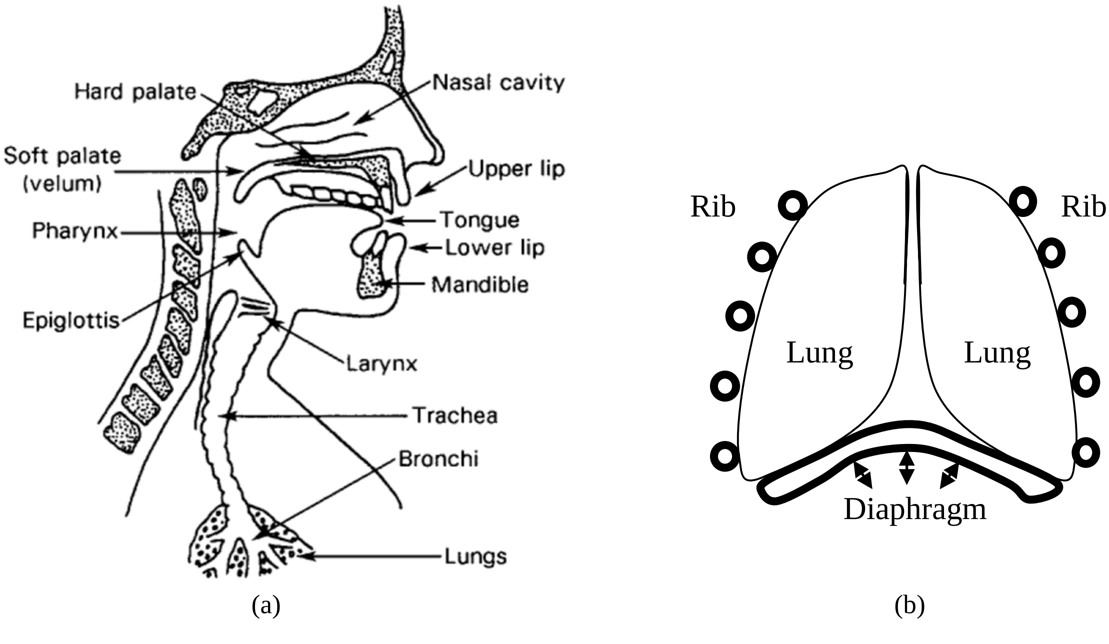


Figure 4.3. (a) Vocal organs from Holmes [12] (b) Lungs from Howard [20].

Air is sucked into the lungs by the contraction of muscles around the rib cage (intercostals) as well as the diaphragm. The contracting of the intercostals and diaphragm causes the lung volume to increase and, if the airway is open, additional air enters the lungs (breathing in). The resting after the contraction lifts the diaphragm, deflates the lungs and pushes the air back out. The air flows through a gap between vocal folds called the glottis. Adduction of the vocal folds decreases the size of the glottis and increases the speed of the trans-glottal air flow. The physical consequence of the increase in velocity caused by lowering of the pressure is called the Bernoulli effect [54]. This effect describes the physics of the vocal folds' movement which causes them to accelerate towards each other until they collide, causing a vocal fold closure, then come apart again by the air pressure pushing from the lungs. The folds can be seen as an oscillating pendulum. More details of these processes are described in [20]. However, if we consider the whole vocal system, the air stream starts flowing in/out through the trachea and larynx, at which time the muscles in the larynx are stretched and relaxed, lengthening/shortening the vocal folds and thereby altering the fundamental frequency. The vibrating of the vocal folds generates a periodic air stream called the voiced sound source, while the open vocal folds cause a non-uniform stream when the flow hits an obstacle or wall in the vocal tract which is called the voiceless source [20]. The sound source is then perturbed by articulators along the oral and/or nasal tract and emerges through the lips and/or nostrils.

In the mid-sagittal plane of the vocal tract, the articulators may be considered to consist of an upper and lower part. The upper consists of upper lip, upper teeth, alveolar ridge, hard palate, soft palate (velum) and uvula, as shown in Figure 4.3. The lower comprises the lower lip, tip, blade, front, centre, back and root of the tongue, and the epiglottis. These places have been used to specify the articulatory gestures in linguistic terms as the "place" of articulation [52].

4.4 Acoustic properties of English Consonants

Producing consonants involves shaping the vocal apparatus and/or rapidly moving its elements [53]. The tree in Figure 4.1 has already shown all of the English consonants in groups according to the manner of articulation. In each group, a similar gesture of specific articulators is involved in the production of the members of that group, as described in Table 4.2 from Ladefoged [53]. However, phoneticians have considered phones according to their place of articulation. Table 4.3 shows the associated place of articulation of each English consonant in relation to its each manner of articulation category.

Table 4.2. English manner of articulation and its details from [53]

Manner		**Description**
Nasal		Closure of the vocal tract and lowering of the velum such that air can go out through the nose, but not through the mouth.
Plosive		Complete closure of the vocal tract. Air is blocked from going out through the nose (velum is raised) and the mouth.
Fricative		Constriction of the vocal tract so that a noisy airstream is formed. Different fricatives have different places of articulation where the constriction occurs.
Affricate		A stop followed by a fricative made at the same place of articulation.
Semi-vowel	Approximant	Constriction of the vocal tract to a smaller extent than that required for a noisy airstream.
	Lateral	The tongue touching the roof of the mouth but without contacting the teeth at the sides.

Table 4.3. English consonants in a table of place and manner of articulation after [53]

		Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Palatal	Velar
Nasal		m			n			N
Plosive		p b			t d			k g
Fricative			f v	T D	s z	S Z		
Affricate						tS dZ		
Semi-vowel	Approximant	w			r		j	w
	Lateral				l			

Note that the consonant /w/ appears in two places of articulation because it involves lip rounding and constriction at velar at the same time. The following subsections describe the categories of consonants in more acoustic detail, in terms of the manner of their articulation.

4.4.1 Nasal

A nasal consonant involves a closure of the vocal tract such that air can go out through the nose, but not through the mouth; it then has a basic resonance due to the nasal cavity. The main acoustic property is therefore a nasal murmur which is within the 200 to 300 Hz range for males. Another property is formed by the anti-formants which come from the resonance in the oral cavity which acts as an acoustic side branch cavity. A range of frequencies from the oral cavity is cancelled in the acoustic output because the energy of these frequencies is absorbed in the resonant oral cavity. The frequencies of anti-formants in the spectrum depend on the length of the mouth cavity. For instance, if the mouth cavity in /m/ is about 8 cm long (as in the adult male), then its resonant frequencies are at 1,100 Hz and 3,300 Hz which we could therefore expect to see an anti-formant in the spectrum of /m/ at around 1,100 Hz. The figure below shows the positions of articulators for the bilabial stop /b/ and nasal /m/.

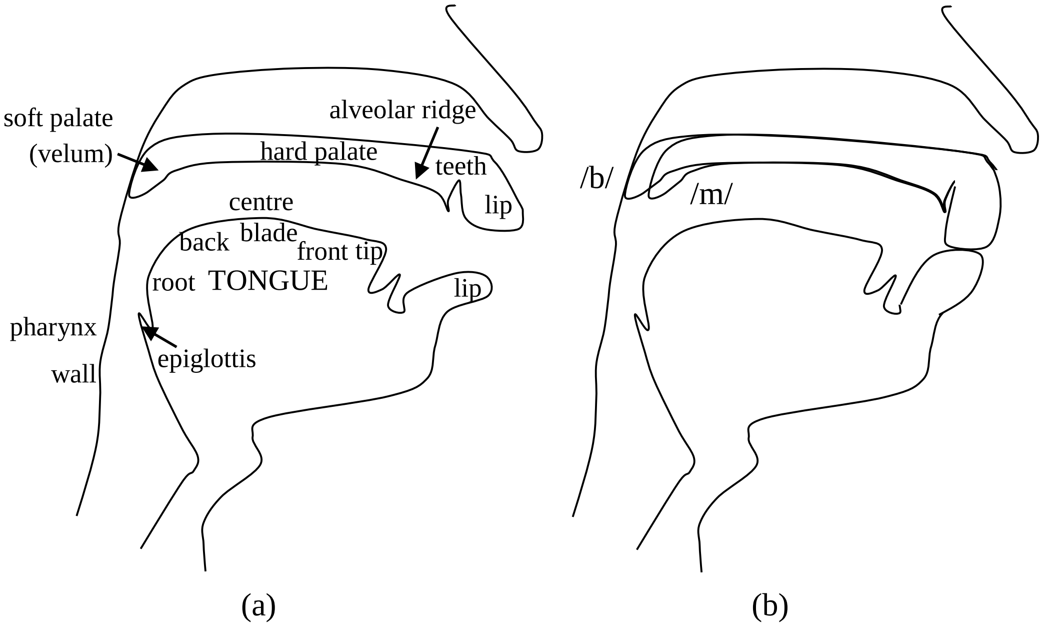


Figure 4.4. (a) The position of articulators in the vocal tract (b) The position of the articulators for the bilabial stop of /b/ and /m/.

Figure 4.4(b) shows the different positions of the velum for bilabial plosive and nasal. The bilabial involves the two lips. In plosives, the place affects the centre frequency of the burst and the formant transitions, while in nasal consonants the place affects the inverted formants or anti-resonance frequencies. For the English /m/ these are at approximately 1 kHz for a vocal tract of 17.5 cm in length, and approximately 3.5 kHz and 5 kHz for /n/ and /N/ [20].

4.4.2 Plosive

The plosive is a type of manner of articulation that involves complete closure of the oral tract, unlike the nasal consonants that involve the addition of the nasal cavity by lowering the velum to allow air to flow through the nasal cavity while the oral cavity is completely closed. After the complete closure the articulators suddenly come apart, the air flow being released as a small burst, such as /b/ in “by”, the lips being completely closed and stopping the air flow before the burst. There are three voiced plosives – /b/ as in “by”, /d/ as in “dye” and /g/ as in “guy” – and three unvoiced plosives – /p/ as in “pie”, /t/ as in “tie” and /k/ as in “key”. All plosives have three main acoustic properties: the centre frequency of the burst, the formant transition and voice onset time. The centre frequency of the burst changes with the place of articulation: it is at around 500 Hz to 1.5 kHz for bilabial (/b/ and /p/); 4 kHz for alveolar (/d/ and /t/); and around 1.5 kHz to 4 kHz for velars (/g/ and /k/) [20]. The formant transition refers to the way in which formants change as the tract shape changes. The change or transition is considered from the frequency of the formant during the hold stage of the plosive, which is called the locus frequency, to the slight move from the locus to the formants of the following vowel. The most prominent change is found in the F2 transition. The transition is rising, almost stable and falling for bilabial, alveolar and velar, respectively. The last acoustic property is the voice onset time (VOT) which refers to the period of time between the plosive burst and the onset of vocal fold vibration of the adjacent vowel. Figure 4.5 shows voicing for plosives from Ogden [55]. Specifically, it shows the vibration period of the vocal fold when the vocal tract is articulating for plosives, when C represents the closing, H the holding and R the releasing phase, and the dotted lines represent the vocal fold vibration period.

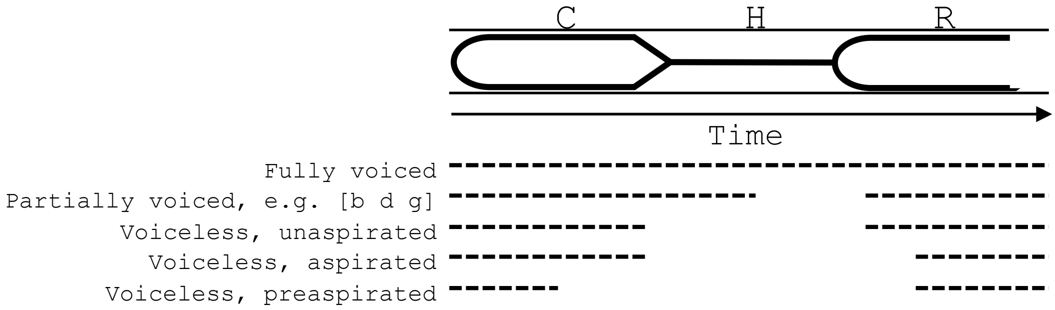
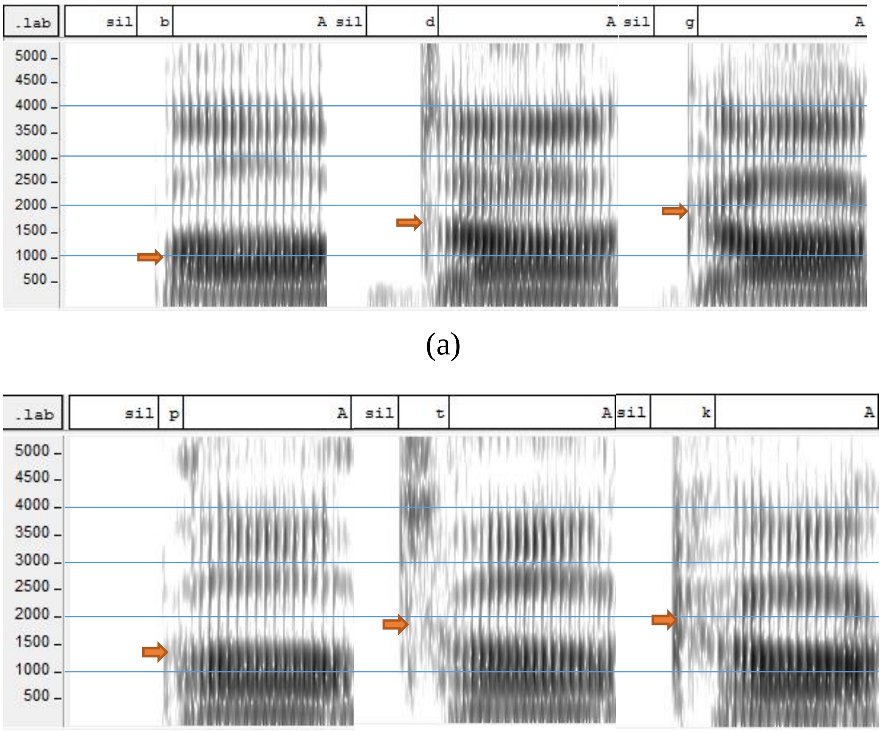


Figure 4.5. Voicing pattern in plosives from Ogden [55].

The voicing pattern in plosives in Figure 4.5 shows different voicing periods, with the dashes representing voicing time. The top graph shows the status of the closing articulator. The voicing period is considered after the stage of closing articulator, such that in voiceless, aspirated the voicing occurs only when the articulator is apart. The VOT strongly depends on the type of plosive. The time is marked with “+” if the voicing begins before the release stage and “-“ if it does on the other way round. All time interval details are included in Table 4.4.

Figure 4.6 shows these properties in recorded speech from [56, 57]. The centre frequencies of bursts are quite short but spread over a wide band. The locus of F2 shows various transitions depending on the place of articulation and the adjacent vowel, which in this example is /A/. The red arrows show the locus of F2. For the transition, F2 is rising in bilabial /b/ and /p/, slightly falling in alveolar /d/ and /t/, and steeply falling in palatal /g/ and /k/.



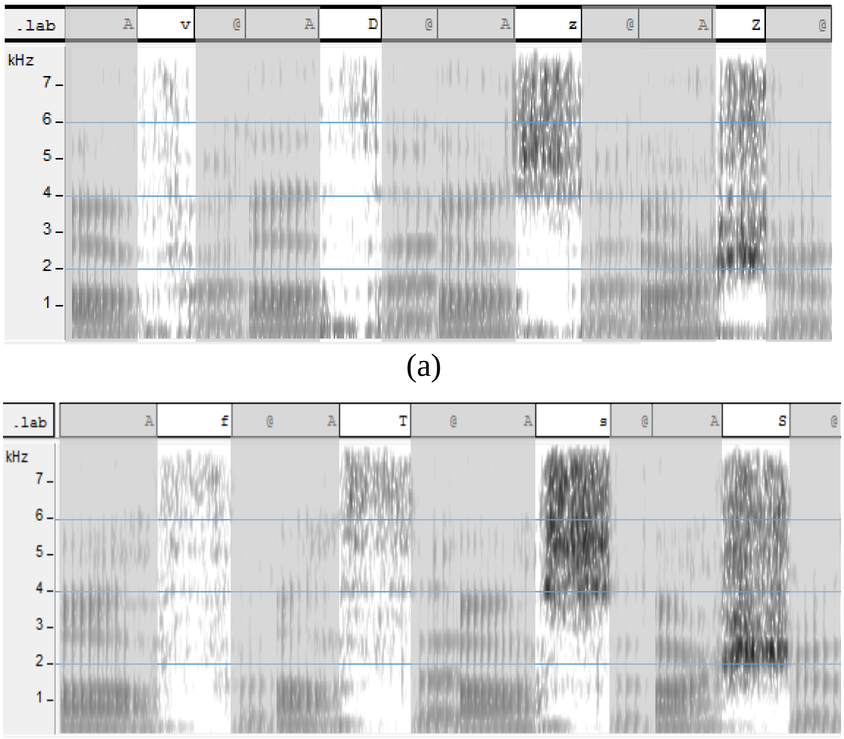
(a)

(b)

Figure 4.6. The spectrogram of (a) voiced plosive /b/, /d/ and /g/ and (b) voiceless plosive /p/, /t/ and /k/ from [56, 57]. All are analysed using Hamming windows laid over 512 points of FFT window length on Wavesurfer.

4.4.3 Fricatives

During the production of a fricative, there is a narrowed constriction in the vocal tract that causes turbulence when the fast-moving volume flow hits inert air at the end of the constriction. Shadle, 1991, mentioned in Johnson 2012 [58], states that the fricative sourcing can be classified into two types, obstacle and no-obstacle. The obstacle is the teeth in the production of /s/ and /S/ or the lips in /f/, while /h/ involves the wall source only and is therefore classified as no-obstacle. The flow that releases from the constriction forms a voiceless excitation source. It is then resonated by the volume of the acoustic cavity. The energy of frequencies of labiodental and dental fricatives (/f/, /v/, /T/ and /D/) is thin and lies over a wide range of high frequencies, whilst the alveolar and palatal have greater energy at above 3.5 kHz and 2.5 kHz, respectively. There are also voiced and voiceless fricatives. In addition to the turbulent noise in fricatives which is generated at the place of constriction, the vocal folds vibrate during voiced fricatives. To explore voiceless and voiced fricatives we can simply pronounce, with a finger touching the Adam's apple, /T/ in "thwart", /f/ in "fog", /s/ in "sea" and /S/ in "ship" (voiceless) and /D/ in "weather", /v/ in "van", /z/ in "zenith" and /Z/ in "treasure" (voiced), as four pairs of voiceless and voiced fricatives. The finger will feel vibration when pronouncing the voiced ones. Double sound sources in voiced fricatives effect a repeating on-and-off turbulence as the vocal folds open and close. Figure 4.7 shows the range of friction noise analysed from recorded speech, voiced on the top and voiceless on the bottom.



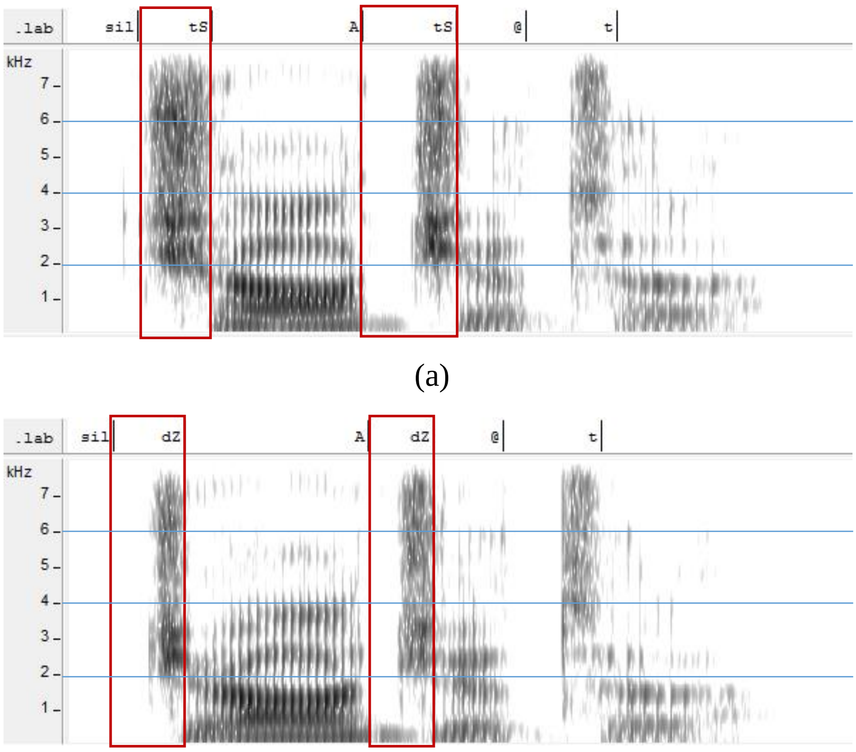
(a)

(b)

Figure 4.7. The spectrogram of (a) voiced fricative: /v/, /D/, /z/ and /Z/ and (b) voiceless fricative: /f/, /T/, /s/ and /S/ from CSTR diphone corpus [56, 57] analysed using Hamming windows overlaid on 512 FFT window points on Wavesurfer.

4.4.4 Affricates

Affricates are the fast concatenation of plosive and fricative gestures. Therefore their acoustic characteristics can be roughly identified sequentially as plosive, formant transitions and fricative, but they are shorter than when they are pronounced alone. There are two affricates in English: /tS/ as in "chain" and /dZ/ as in "jibe" [20, 53, 52]. Figure 4.8 shows examples of recorded /tS/ and /dZ/ in a carrier vowel /A/ and /@/.



(a)

(b)

Figure 4.8. Spectrogram of /tS/ (top) and /dZ/ (bottom) fricatives from CSTR diphone corpus [56, 57] analysed using Hamming windows overlaid on 512 FFT points on Wavesurfer.

4.4.5 Semi-vowels

Also known as "approximant", this manner of articulation does not involve any closure in the vocal tract or additional noise source or nasalization, but it does involve a narrowing in the tract though not as tight as in fricatives. The semi-vowels /w/, /j/, /r/ and /l/ are all voiced. The first three are central but /l/ is lateral.

As /w/ and /j/ are central approximants, /w/ involves a lip rounding and a raising of the back of the tongue. This articulation is similar to that for the vowel /u/. For /j/, the front of the tongue is raised but not sufficiently high to block the sound which is also like the articulation for pronouncing the vowel /i/. Hence, the major acoustic properties of this group of consonants are like their similar vowels. Figure 4.9 shows their formants which are vowel-like.

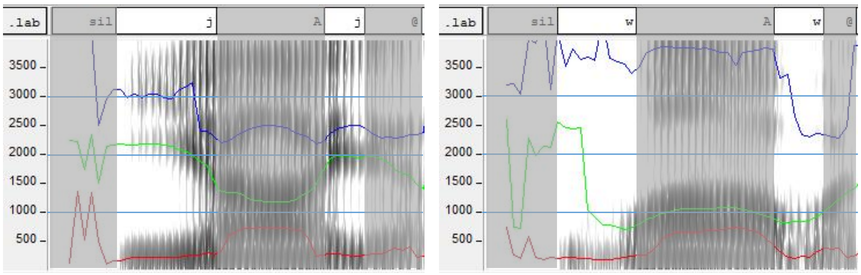


Figure 4.9. The formants of /j/ on the left and /w/ on the right from [56, 57] analysed on Wavesurfer using LPC order 12 overlaid on 0.049 s long Hamming windows.

In summary, this research attempts to synthesize the English consonants following these acoustic properties. Therefore, the centre frequency of burst, formant transition, voice onset time and anti-formants are summarized in Table 4.4, gathered from various works [20, 53, 59, 6, 32].

Table 4.4. Some characteristics of English consonants, based on [20, 53, 59, 6, 32].

	Centre frequency of burst	**Formant transition**	**Voice onset time**	**Anti-formant**
Plosive Bilabial	Approximately 500 Hz to 1.5 kHz	F2 frequency increase from stop release into following vowel	Relatively short; prevoicing likely for voiced bilabial plosives (V+ around 5 ms, V- around 75 ms)	-

	Centre frequency of burst	**Formant transition**	**Voice onset time**	**Anti-formant**
Plosive Alveolar	Approximately 4 kHz	F2 frequency decrease from stop release into following vowel except for the high-front vowels	Intermediate between bilabials and velars (V+ around 10 ms, V- around 85 ms)	-
Plosive Velar	Approximately 1.5 kHz to 4 kHz	F2 and F3 have a wedge-shaped pattern in which they are initially nearly fused but separate in frequency during the transition	Longest values across the 3 places of stop production; long lags likely for voiceless velars. (V+ around 15, V- around 90 ms)	-
Nasal Bilabial	-	F2 locus frequencies are around 1.2 kHz	-	Approximately 1 kHz
Nasal Alveolar	-	F2 locus frequencies are around 1.8 kHz	-	Approximately 3.5 kHz
Nasal Velar	-	F2 locus frequencies are around 2.1 kHz	-	Approximately 5 kHz
Fricative Labiodental	-	F2 locus frequencies are around 1 kHz	-	-

	Centre frequency of burst	**Formant transition**	**Voice onset time**	**Anti-formant**
Fricative Dental	-	F2 locus frequencies are around 1.4 kHz	-	-
Fricative Alveolar	-	Depends on context vowel. Significant noise energy around 4 kHz	-	-
Fricative Palatal	-	Depends on context vowel. Significant noise energy around 2.5 kHz	-	-
Glottal Fricative	-	No formant transitions	-	-
Affricate	-	Only small formant transition into the fricative	-	-
Semi-vowel	-	Formants are close to their associated vowels	-	-

Chapter 5
2-D Digital Waveguide Mesh for English Consonants

Digital waveguide mesh synthesis is one of the numerical methods for wave propagation simulation. It has been widely used in room acoustic studies and has also been shown to be capable of simulating the acoustics of a small cavity such as the vocal tract [5, 60]. In [61], Mullen described his 2-D DWM vocal tract modelling with a linear dynamic impedance function that could synthesize diphthongs successfully. In this research we extend its capability to consonant synthesis, using non-linear articulatory trajectories based on real measurements of human speech production. Tools, data and system configurations for consonant studies are described in this chapter.

5.1 Tools and data

5.1.1 Mullen's 2-D Digital Waveguide Mesh

The two-dimensional digital waveguide mesh (2-D DWM) was used for vocal tract tube simulation to model pressure propagation of sound in the tract. The implementation involved cosine function impedance modelling across the tract to manage the articulation [5].

The waveguide architecture, including the number of connecting ports at a junction, is considered according to the ease of area function mapping and capability to attach another lattice for the nasal branch; therefore, a rectilinear DWM is used in this research. Rectilinear meshing in this research is in 2-D and represents vocal tract width on the y-axis and length on the x-axis. The area function is discretized as W(x) which is a cross-sectional radius function that controls impedance which represents constriction in the tract. As mentioned in Chapter 4, consonants involve at least a constriction and/or articulation in the vocal tract; therefore this chapter describes the implementation of a system that can handle each manner of articulation using 2-D DWM.

The joining of two discontinuous tubes results in changes in impedance (or admittance) in the simulation. Using a 2-D model to represent the relationship in a static rectangular 2-D DWM, the changing of the area function would have a small effect in the cross-sectional plane. A constriction is applied by raising the impedance in order to encourage cross-tract reflection. The cross-sectional area is inversely proportional to the impedance value. Equation 5.1 shows the impedance function from [5] at a junction node at coordinate (*x*, *y*) where *w* is the width of the tract and *y* is the position of a corresponding waveguide junction node along the width and *x* is the position of a corresponding node along the length of the vocal tract.

$$Z^c(x,y) = Z_x^A - \frac{(Z_x^A - Z_{min}^A)}{2}[1 + \cos(2\pi\left(\frac{y}{w} - \frac{1}{2}\right))] \qquad 5.1$$

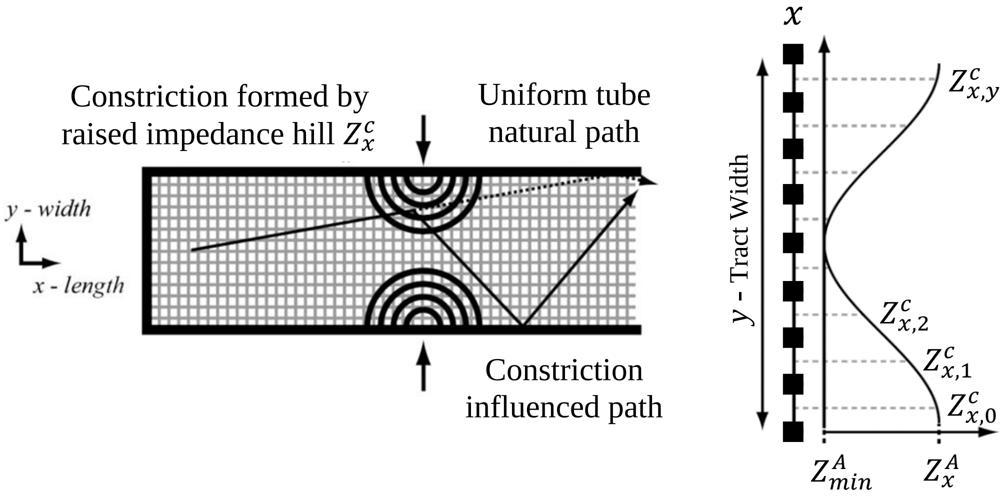
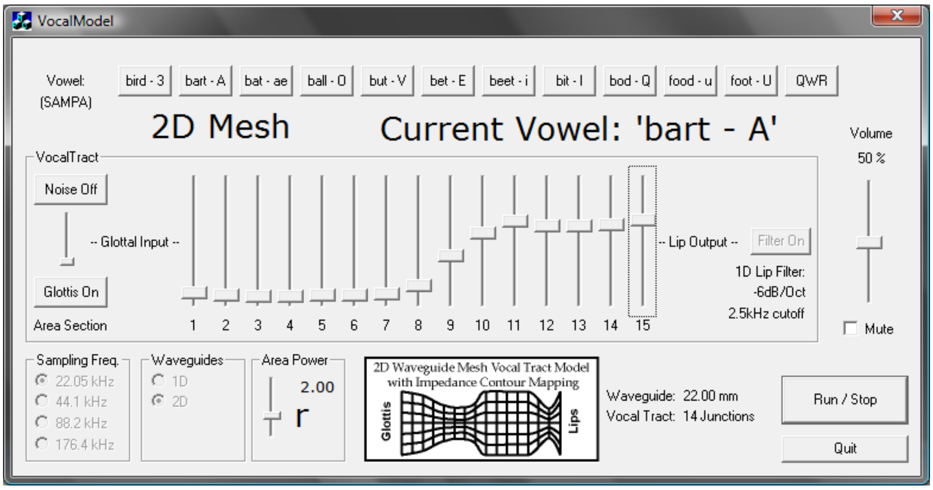

Constriction formed by raised impedance hill Z_x^c
Uniform tube natural path
y - width
x - length
Constriction influenced path
x
y - Tract Width
$Z_{x,y}^c$
$Z_{x,2}^c$
$Z_{x,1}^c$
$Z_{x,0}^c$
Z_{min}^A
Z_x^A

Figure 5.1. Raised impedance hills causing a constriction in a straight tube and plotted raised cosine impedance hills on either side of the constriction after [5].

Figure 5.1 shows raised impedance similar to those in Figure 3.4, but here superscripts indicate the calculating source of value. There are two superscriptions of impedance Z, c and A. They denote the type of impedance. *A* denotes one that is calculated from the *A*rea function and c denotes one that is already weighted by *C*osine impedance afterwards. The half circular contours represent the impedance hill $Z_{x,y}^{c}$ at the point of constriction *x*. $Z_{x,0\ldots n}^{c}$ represents cosine-weighted impedance values at different y positions and the maximum impedance value of a constriction x is $Z_{x,0}^{c}$ which equals Z_{x}^{A} and the minimum is $Z_{x,n/2}^{c}$ or Z_{min}^{c} which equals the smallest impedance Z_{min}^{A}, as the implementation of cosine smoothing the impedance hill. The highest impedance is called Z_{stop} which happens when the area function is 0. This is the impedance for the modelling of a complete cut-off of the air-flow.

Jack Mullen's 2-D DWM Synthesizer

Jack Mullen's articulatory synthesizer was written in C++. It is a dialog box application with MFC and the PortAudio – an open source audio I/O library. Mullen simulates the vocal tract using a 17.5 x 5 cm rectilinear (4 ports connected at node) and hexagonal (3 ports connected at node) DWM with 0.92 reflecting values at the walls, 0.97 at the glottis end and -0.9 at the lips. Some parameters are adjustable before the user starts the pressure propagation, such as the voiced:noise ratio, the power of area function, sampling rate and waveguide dimension (1-D/2-D), but the tract shape is able to be modified at run time. The most important parameter is the effect of changing the shape and it is also available for dynamic real-time synthesis. In common PC performance, users could perceive real-time continuous synthesized speech at a sampling rate of 22 kHz. The following figure is a screen capture of Mullen's application.



2D Mesh Current Vowel: 'bart - A'

1D Lip Filter: -6dB/Oct 2.5kHz cutoff

Volume 50 %

Sampling Freq.: ● 22.05 kHz ○ 44.1 kHz ○ 88.2 kHz ○ 176.4 kHz

Waveguides: ○ 1D ● 2D

Area Power: 2.00 r

Waveguide: 22.00 mm
Vocal Tract: 14 Junctions

Figure 5.2. A screen capture of Mullen's software.

The source code is written in Object-Oriented C. Each class works independently. VocalModelDlg is a class for the dialog box. It contains all handler modules as well as start and end points of the application data flow. As an overview of the application, the following class diagram shows data member association of some important classes of the application.

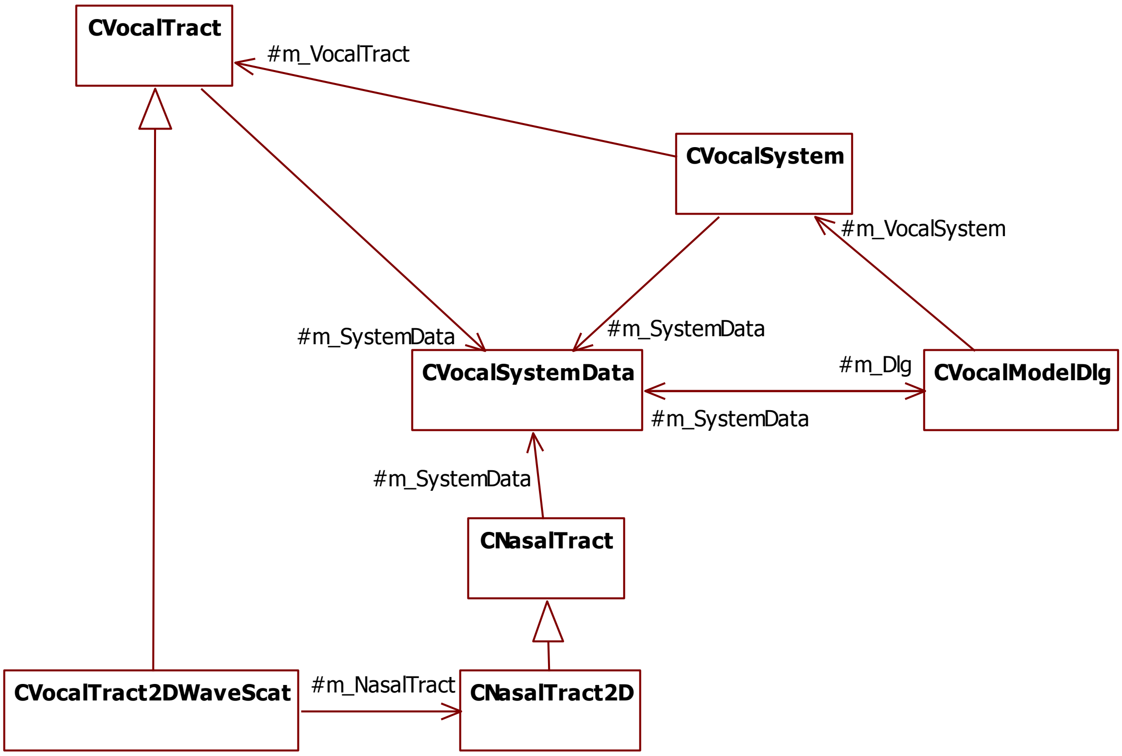


Figure 5.3. Class diagram of Jack Mullen's synthesizer.

CVocalSystem is the class that contains all vocal tract data of each vowel. CVocalSystemData is the class for the wave propagation management which constructs the DWM as user input (1-D/2-D, the sampling frequency and the power of area function). During propagation, the output out is sent via PortAudio's sound card buffer writer.

As mentioned before, the structure of Mullen's work is based on object-oriented programming. His 'VocalTract' class acts as a parent class from which all tract meshing classes have the same structure and inheritable properties. VocalTract2DWaveScat is a class that inherits from VocalTract which contains the implementation for 2-D rectilinear wave scattering. This class is implemented with functions to simulate wave propagation in the vocal tract under rectilinear scattering. The nasal tract class is the same. It inherits from VocalTract and also has rectilinear scattering function. The interconnection between them is implemented based on velopharyngeal port scattering. There was no need to implement any new classes; rather some part of their function was altered to enable the synthesis of consonants. Therefore the changes made were to connect them under the proposed 5 port scattering junction following equations 3.15 and 3.16 for nasal synthesis in Scattering(), some changes in Timestep() for noise source injection and some changes in getSample() for articulation control. All changes appear with comments in the source code in Appendix 1.

5.1.2 Voiced and unvoiced source

Due to the limitation of DWM, two types of sound sources are used to excite the system. In this research, the voice source is simulated from the Liljencrants-Fant (LF) model [62] with the fundamental frequency set to 130 Hz with male vocal tract shapes and articulatory trajectories that were captured from male subjects with small pitch variation. To set up the sound source, four main parameters for the LF model for modal voice were adopted from Mullen's work [5] of which t_c, t_p, t_e and t_a, are 1.000, 0.600, 0.780 and 0.028, respectively, where t_c is the parameter for the fundamental period, t_p is the parameter for the maximum glottal flow, t_e is the parameter for the abrupt glottal closure and t_a is the parameter for the effective duration of the return phase. In terms of voice quality, t_p, t_e and t_a can be adjusted to proper values; for example, for breathy voice the glottal flow (t_p) has to be longer (higher value than for modal) but the closure (t_e and t_a) has to be shorter. Figure 5.4 shows the fluctuation of glottal pressure source for modal voice from modelling.

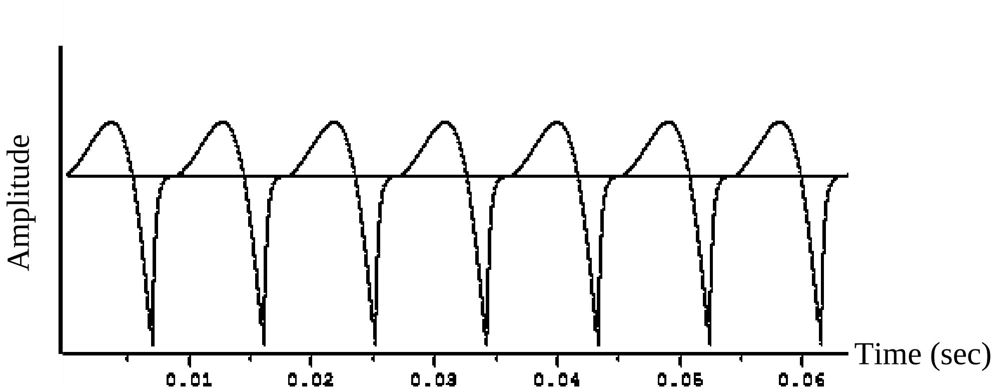


Figure 5.4. 88k voiced excitation source used in this research.

Another type of source is for unvoiced sounds which make use of frication and this uses white noise. It is generated using the random function in C seeded by a run-time value returned from function time() and is used as a monopole source only here.

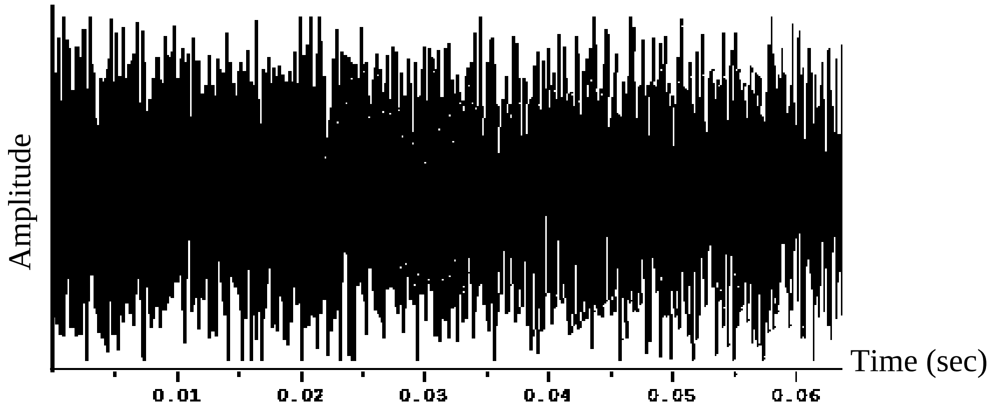


Figure 5.5. 88k white noise used in this research.

5.1.3 Vocal tract shape and articulatory trajectory corpus

In this research, the vocal tract shapes are extracted from Magnetic Resonance Image (MRI) and the articulation is done after the trajectory of articulators from Electro Magnetic Articulograph (EMA). The articulograph enables a speaker's articulator movements to be tracked when a speaker speaks a sentence or phrase. Figure 5.6 shows an image of a model in the Electro Magnetic Articulograph recording machine (Carstens AG500). Coils are attached to articulators inside and at his mouth. It records the position of coils when the speaker speaks a sentence/phrase. They can cause a speaker to feel uncomfortable when speaking so the speaker has to spend some time getting used to having the coils attached before starting a recording. The MRI detects those parts of the human body that contain hydrogen protons, which are those rich in water in their soft tissues, and hence bones and teeth are not detected which are two disadvantages of the method. However, there are pros and cons to all human body tissue imaging methods. One of the advantages of MRI is that it doesn't involve radiation exposure such as X-rays or CT [63], hence many researchers in this area use MRI.

The MRI data set that was recorded in Speed's work in 2012 [59] is used. There is only one set in his data that contains almost all the English consonants (except /l/ and /r/), and that is for the subject pseudo named Jack. This set was recorded from a male speaker [59], which coincidently and conveniently is the same gender of the speaker in the published articulograph data set called mngu0 distributed on-line by CSTR in the University of Edinburgh. Although these two data sets are for different male speakers, they are both British. Therefore the data sets are used in this attempt to synthesize English consonants by matching the four recorded trajectories, lower lip, tongue tip, tongue body and tongue dorsum are mapped to relevant cross-sectional areas manually.

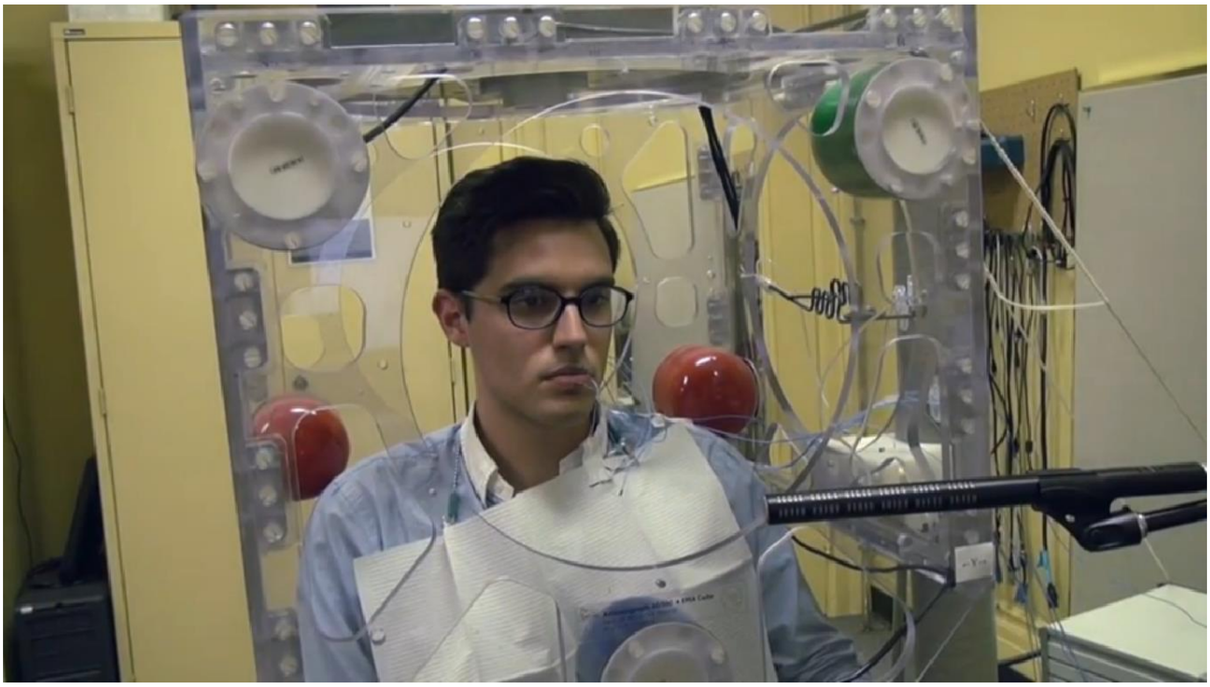


Figure 5.6. Image of a model in Carstens AG500 electromagnetic articulograph from [64].

Vocal tract MRI data from [60] are used in this research. The data set contains tract shape from five participants who are English speakers with a phonetic or professional singer background. Jack, Jill, Jasmine, Jim and Jeff were the names used for the speakers to hide real participant names and keep identities anonymous. The male subject Jack is mainly used in this work since it is the only subset that contains all English consonant images. They were scanned using a General Electric 3.0T HDx Excite MRI Scanner at the York Neuroimaging Centre (YNiC). The subjects were required to hold their articulators in a static position for a given sound for 16 seconds. The data was then processed using ITK-Snap [65] to extract the tract shape followed by the cross-sectional area data using VTK [66]. Figure 5.7 shows examples of cross-sectional images from the corpus. Figure 5.7 (a) shows Jack's vocal tract images when he holds his tract for pronouncing /b/, /d/ and /g/ respectively, and (b) shows Jim's images when he holds his apparatus for pronouncing /m/, /n/ and /N/. The vela of the participants also shows the use of the nasal tract in different manners of articulation.

To track the dynamic articulation of the tract shape, articulatory trajectories are adopted from an EMA corpus called *mngu0*. *mngu0* is an EMA-published articulatory corpus recorded from two speakers reading 460 British TIMIT utterances by CSTR, University of Edinburgh. It consists of more than 1,354 phonetically diverse utterances recorded at

Ludwig-Maximilians-Universität at München, using a Cartsens AG500 electromagnetic articulograph [67]. Using the machine, six coils are placed in the mid-sagittal plane, as shown in Figure 5.8. An extra one is placed on the subject's nose for head-movement correction. The AG500 tracks sensor coils in 3-D space with two angles of rotation which means five measurements per sensor coil pulse with two reliability indicators for each coil. The recorded data was manipulated and distributed in Matlab format.

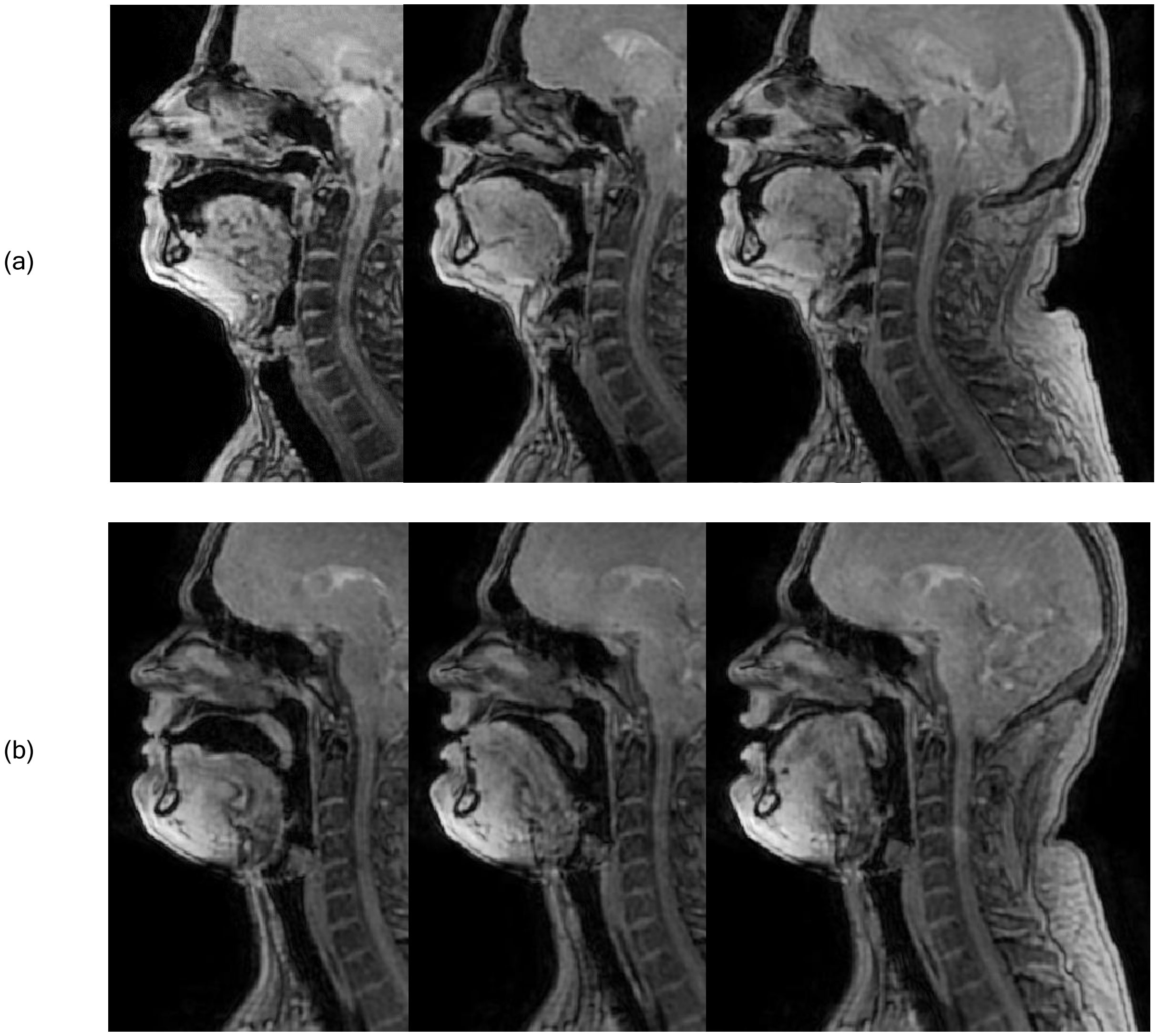
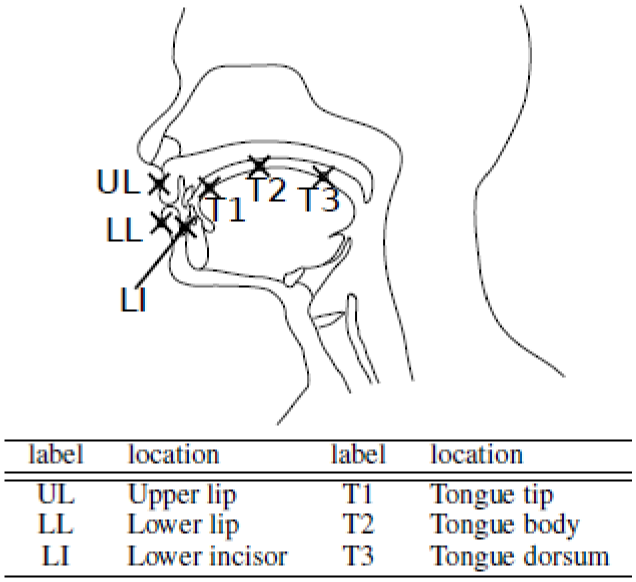


Figure 5.7. (a) Cross-sectional MRI from Jack pronouncing /b/(left), /d/(mid) and /g/(right) and (b) Cross-sectional MRI from Jim pronouncing /m/(left), /n/(mid) and /N/(right).



label	location	label	location
UL	Upper lip	T1	Tongue tip
LL	Lower lip	T2	Tongue body
LI	Lower incisor	T3	Tongue dorsum

Figure 5.8. Sensor coil locations from [67].

The data set contains approximately 1,715 distinct diphones of which 512 are paired with plosives. In our research we extract CV pairs from the corpus to obtain their trajectories. Table 5.1 shows the number of CV diphones found.

Table 5.1. Number of diphones found paired with corresponding phone.

Diphone paired with corresponding phone	Number found in *mngu0*
/p/	645
/b/	734
/t/	1707
/d/	896
/k/	791
/g/	325
/m/	856

Diphone paired with corresponding phone	Number found in *mngu0*
/f/	579
/v/	436
/s/	1006
/z/	487
/S/	217
/Z/	94
/h/	672

Diphone paired with corresponding phone	Number found in *mngu0*
/n/	939
/N/	91
/j/	198
/w/	972

Diphone paired with corresponding phone	Number found in *mngu0*
/tS/	210
/dZ/	266

Table 5.2. Number of distinct diphones found paired with corresponding phone (CV).

Distinct diphone pair (CV)	Number found in *mngu0*
/p/	19
/b/	19
/t/	19
/d/	17
/k/	19
/g/	19
/m/	18
/n/	19
/N/	17
/j/	17
/w/	17

Distinct diphone pair (CV)	Number found in *mngu0*
/f/	18
/v/	18
/s/	19
/z/	17
/S/	19
/Z/	12
/h/	18
/tS/	18
/dZ/	19

Diphone trajectories are used in the simulation for naturalness evaluation. Examples of trajectories are shown in Figure 5.9. Black, grey, black dash and grey dash represent normalized trajectories of lower lip, tongue tip, tongue body and tongue dorsum, respectively. These trajectories will be used to control coarticulation between phone and adjacent vowel in this research. Figure 5.9 shows an example of normalized articulatory trajectories.

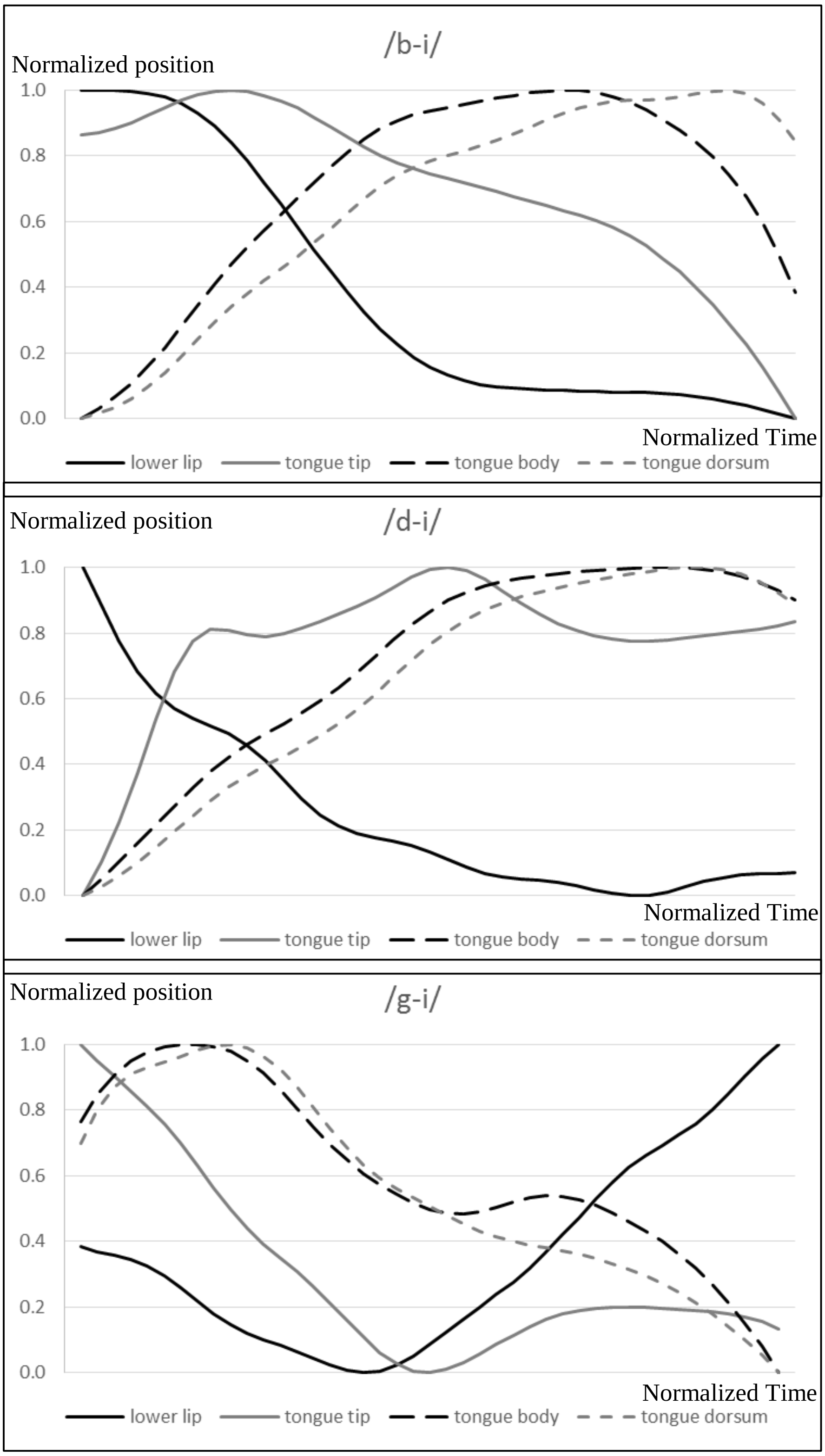


Figure 5.9. Examples of articulatory trajectories from /b/ to /i/, /d/ to /i/ and /g/ to /i/ in which the black line is normalized lower lip trajectory, grey line is normalized tongue tip trajectory, black dash line is normalized tongue body trajectory, and grey dash line is normalized tongue dorsum trajectory.

The movement of trajectories (as shown in Figure 5.9) demonstrate that articulators move non-linearly. For example, the movement of the lower lip (black line) and the tongue tip (grey line) in the left graph shows a different glide from their source position for phone /b/ as compared to phone /i/. The lower lip starts moving down steeply before the tongue tip even reaches the high position and slowly drops to the position for the vowel /i/. These show a non-linear changing of the tract shape. Therefore, in this research we called the adoption of the changing a non-linear articulation of the simulation.

The main use of recorded trajectories rather than linear interpolation between source and target phone tract shape is to support the idea of improving the naturalness of the movement of areas at four places of recorded articulation. It has been proved in [61] that different tract shapes have different acoustic resonances. Therefore changing the tract shape as a function of articulation can produce resonances similar to those in recorded speech for appropriate articulatory gestures. These are shown and compared in the next chapter.

5.2 The simulation

This section describes the setting of the simulation for consonants using the 2-D digital waveguide mesh. Vowel settings have been evaluated in [45]. This research modifies the system to control the articulation of the non-vowel sounds. Specific requirements and settings for each phonetic manner of articulation are described in the following subsections.

5.2.1 Semi-vowels

The semi-vowel manner of articulation is similar to that for vowels (the details are described in Chapter 4). The semi-vowels in English are /y/, /w/, /r/ and /l/. To form them, two articulators have to be articulated close to each other but not so close as to generate friction noise or a closure. In the lateral /l/ and rhotic /r/, the constriction is formed by raising the tongue tip against the alveolar ridge. This creates two separated air spaces; we can feel them on the left and right when we hold the tongue for /l/, and above and under when we hold the tongue for /r/. As there is no MRI of /r/ and /l/ in our MRI data set, this research focuses only on the semi-vowels /j/ and /w/. As mentioned in Chapter 4, in the articulation for /j/ the vocal tract needs to be shaped as for the vowel /i/ but the articulation needs to be faster than for a diphthong that contains /i/. Similarly for /w/, the tract needs to be shaped as for /u/ but articulated faster than a diphthong that contains /u/. The constrictions are located at high-front position and high-back, respectively.

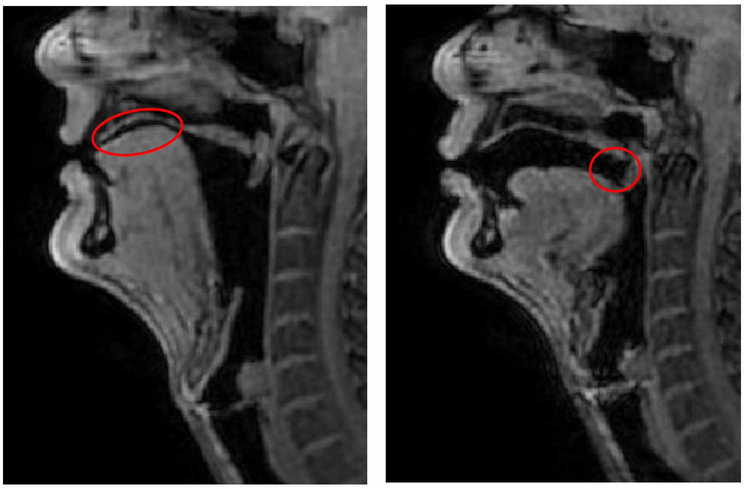


Figure 5.10. Constriction of /i/ (left) and /u/ (right) in MRI from [60] for synthesizing /j/ and /w/, respectively.

The articulatory trajectories for the semi-vowels are used along with cross-sectional area data extracted from the MRI corpus. A series of cross-sectional areas are extracted from the MRI. Each tract area in the series is paired with the area of a target phone of the same location in the vocal tract for the articulation process. The trajectories then work correspondingly within the gap between those areas to guide the change. Figure 5.11 (a) shows the changing cross-sectional area for synthesizing /jA/ and /wA/. Each line shows the trajectory at each waveguide junction for one articulator, which is mapped to cross-sectional area data during synthesis. Four solid lines show cross-sectional areas that change according to the position of recorded articulators in *mngu0*, lower lips, tongue tip, tongue body and tongue dorsum. Dashed lines show examples of linear change on an area trajectory where there is no articulatory guide line. Note that the articulatory trajectory data are recorded in the oral space only, and therefore there is no trajectory guide line in the pharyngeal space (only lower lip, tongue tip, tongue body and tongue dorsum positions recorded); hence, linear interpolation is applied from the position of velar to glottal.

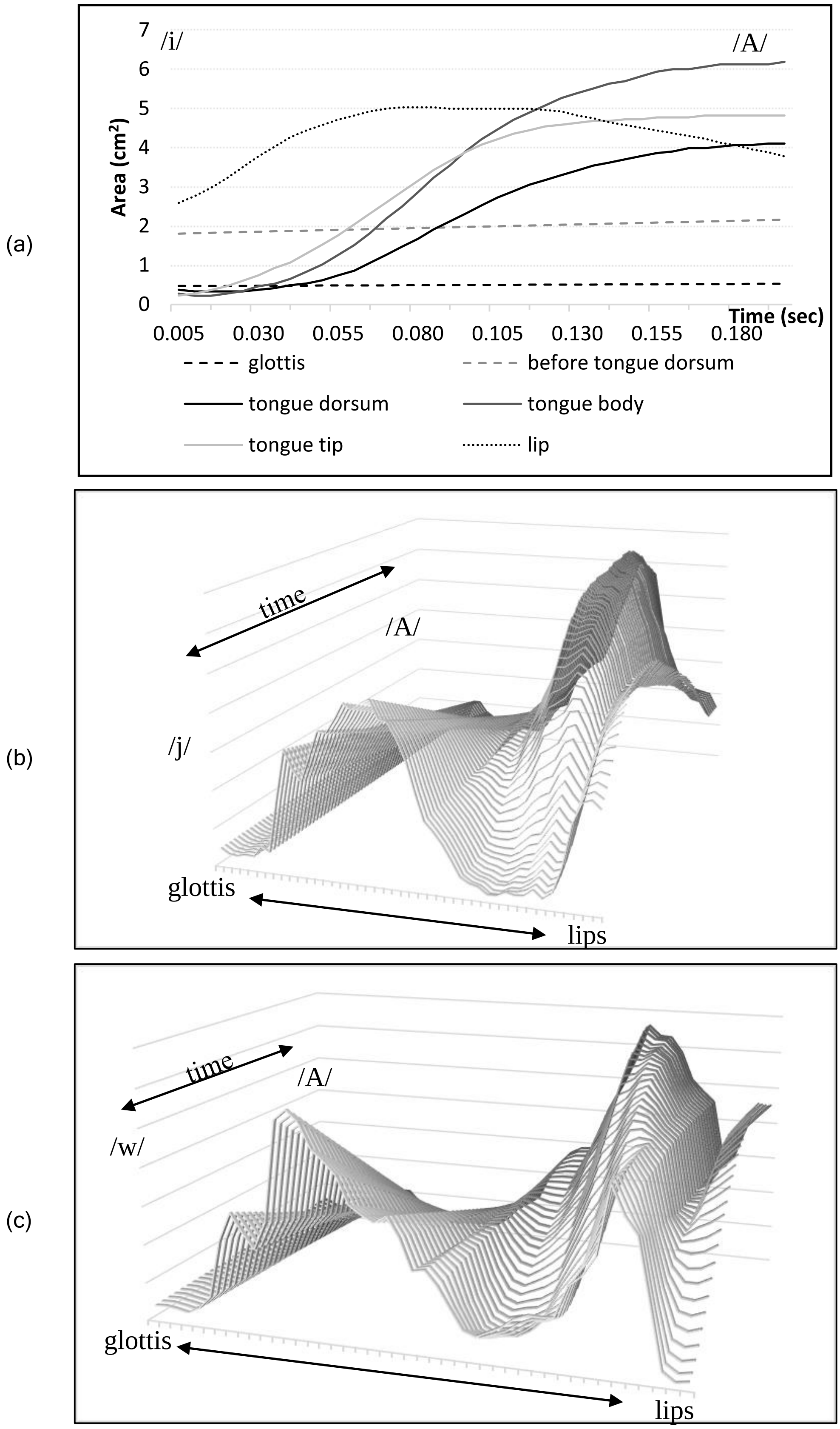


Figure 5.11. (a) Area function changing after trajectory adoption in /jA/. (b) Vocal tract shape ladder from /j/ to /A/. (c) Vocal tract shape ladder from /w/ to /A/.

The middle and bottom plots in Figure 5.11 (b and c) show the articulation from another perspective. They were generated by implementing the trajectories from the corresponding areas extracted from MRI data. Plot (b) shows the ladder of articulation from /j/ to /A/ in 0.005 s steps as in plot (c), which also shows another set of the articulatory trajectories from /w/ to /A/.

To control articulation in the tract, all trajectories are mapped one-to-one to area functions. Imagine that the vocal tract is mid-sagittally filled in by a rectilinear 2-D mesh, the x-axis representing vocal tract length and the y-axis representing vocal tract width. Then node junctions in the rectilinear mesh are strung together vertically and horizontally by waveguides. Here, the impedance values are equal in each waveguide in the vertical string but vary along the horizontal direction when the vertical waveguide string represents waveguides that are connected together vertically at position x in the vocal tract. The cosine function acts as an impedance weight hill which is mapped on to waveguides differently in each vertical waveguide string, highly weighted at the boundaries and lightly at the middle, as shown in Figure 5.1.

5.2.2 Nasals

For nasals, an additional branch is needed to model the nasal cavity. Nasal tract wave propagation simulation is done by adding a nasal tract branch to let the pressures propagate through the velopharyngeal port. In the simulation, the pressures have to be scattered in both directions from the vocal tract to the nasal branch and vice versa. The port is set to be 10.0 cm away from the glottis with a 5.0 cm width for a connection that it has to make with the vocal tract. The 10 cm distance from the glottis is averaged from the MRI data set of /m/, /n/ and /N/ which are slightly different, varying in the degree of the lowering velum (less than 0.52 cm difference in the data set). For the ease of computation with slight effects from using waveguide sizes 2.2, 1.1 and 0.55 cm in the synthesis, which are larger than the variation of the port location, the distance is approximated and fixed at 10.0 cm for all nasalization. The actual port size is controlled by the impedance function of the first waveguides that attach/connect to the vocal tract junctions (shown in a red circle in Figure 5.12). Here, the system consists of two tube models – one for the vocal tract and another for the nasal. The model for the vocal tract has one end attached to the sound source (glottis) and another end for giving output speech (mouth). The addition of the nasal tract requires another model with one end attached to the vocal tract at the velopharyngeal port and another end as nostrils. For 1-D digital waveguide simulation, the attachment is simply simulated by adding a nasal waveguide chain to the vocal main line at the junctions of the velopharyngeal port, while in 2-D a construction, the attachment has to be considered in the X or Y axis. Figure 5.12 shows a 2-D example of attaching a nasal model in the Y axis which is 5 cm wide; therefore the attachment is done by a conversion of 4-port to 5-port junctions.

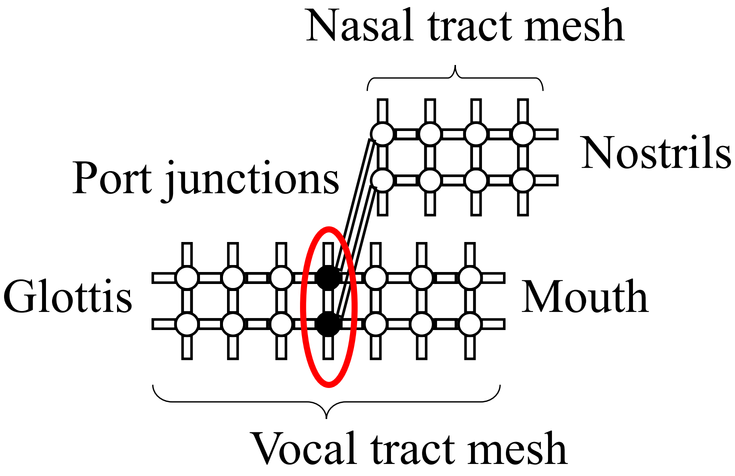


Figure 5.12. Tract structures when running the rectilinear 2-D mesh on 22 kHz.

The size and number of waveguides came from a frequency dependent dispersion equation from Savioja and Lokki, 2001.

$$f_s = \frac{c\sqrt{N}}{d} \tag{5.2}$$

where *c* is the speed of sound, *N* is a number of dimensions of the waveguide mesh, *d* is the waveguide length and f_s is the sampling frequency. In our simulation, f_s is 22 kHz, *c* is 343 m/s, *N* is 2, then *d* is 0.022 m (2.2 cm). The vocal tract is 17.5 x 5 cm^2 so the simulation needs 8 x 3 waveguides which contains 14 junctions (as shown in Figure 5.12). The nasal tract is 11 x 5 cm^2 and therefore it needs 5 x 2 waveguides (8 junctions).

To evaluate the proposed system nasalization performance, synthesizing English nasals was done within two settings. The first was done by making a simple closure at a point in the oral tract together with opening the velopharyngeal port to let pressure propagate through the nasal tract. At the closure, the cross-sectional area is set to 0 by pulling a slider on the application interface down to the minimum. Then the other cross-sectional areas are set to values of tract shape for the vowel /3/. The second was done by setting the vocal tract area according to real nasal cross-sectional area data. Neither settings involve articulation; therefore the velopharyngeal port parameter is set open by using 1.0 as a covariance. The results will be shown and discussed in the next chapter.

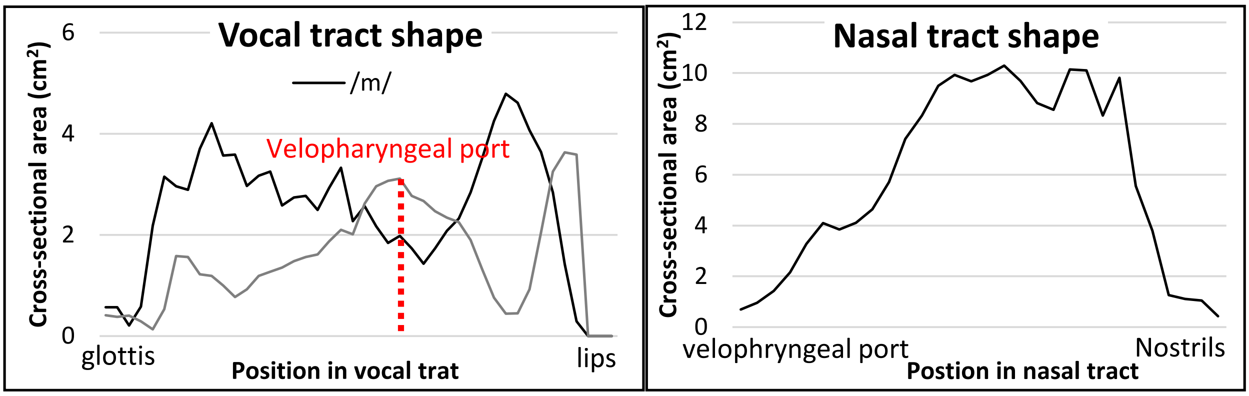


Figure 5.13. Vocal tract shape (left) and nasal tract shape (right).

The two settings are used to work with cross-sectional area data that is extracted from the MRI corpus [60] as shown in Figure 5.7. The extracted area data is shown in Figure 5.13. The left-hand figure shows tract shape differences. The black and grey lines show tract shape extracted from the MRI but the shape in grey was adjusted from the original shape of the vowel /3/ by putting a closure at the lips. The resonances from these two shapes are shown and discussed in the next chapter. The right graph shows extracted nasal tract shape depicted using extracted cross-sectional area through the tract from the MRI. All area data in this research are obtained by

- filling the tract by 3-D mesh using an automatic segmentation tool in itk-SNAP v. 2.2 [68],

- storing the extracted tract shape mesh in Visualization Toolkit (VTK) file format [69],

- and extracting cross-sectional areas using a slicing tool (written by Matthew Speed, 2012) [60].

In the filling process, the complicated shape of the soft tissues, turbinates and cartilages was ignored and counted as nasal space. This nasal tract shape is used throughout the experiments in this research.

5.2.3 Plosives

For plosives, there is no air escaping through the nose, so the oral tract is the only space we are concerned with. The name “plosive” describes the bursting noise resulting from the plosive release (e.g. lip closure release in /b/ or /p/). A plosive has three main phases of articulation: closing, hold and release [55]. During the hold phase, the pressure behind the closure builds up, then a sudden release causes the burst of noise, or plosion, on release. The DWM modelling in this research has a limitation of simulating pressure propagation only and not the flow; therefore it cannot simulate the bursting noise. Hence, this research uses white noise as the bursting noise source to verify the output acoustic resonance instead. In addition, further details and more research in plosive noise source generation and simulation can be found in [70, 71].

This research mimics the burst using an extra noise source injection at the place of constriction which is located at

$$x_{min} = \min A(x) \tag{5.3}$$

where x is junction index along the tract length, A is the cross-sectional area function and x_{min} is the index that points to the minimum area.

Figure 5.14 shows an example of projection of tract simulation under cosine admittance control for the phone /d/, where the red arrow locates the closure.

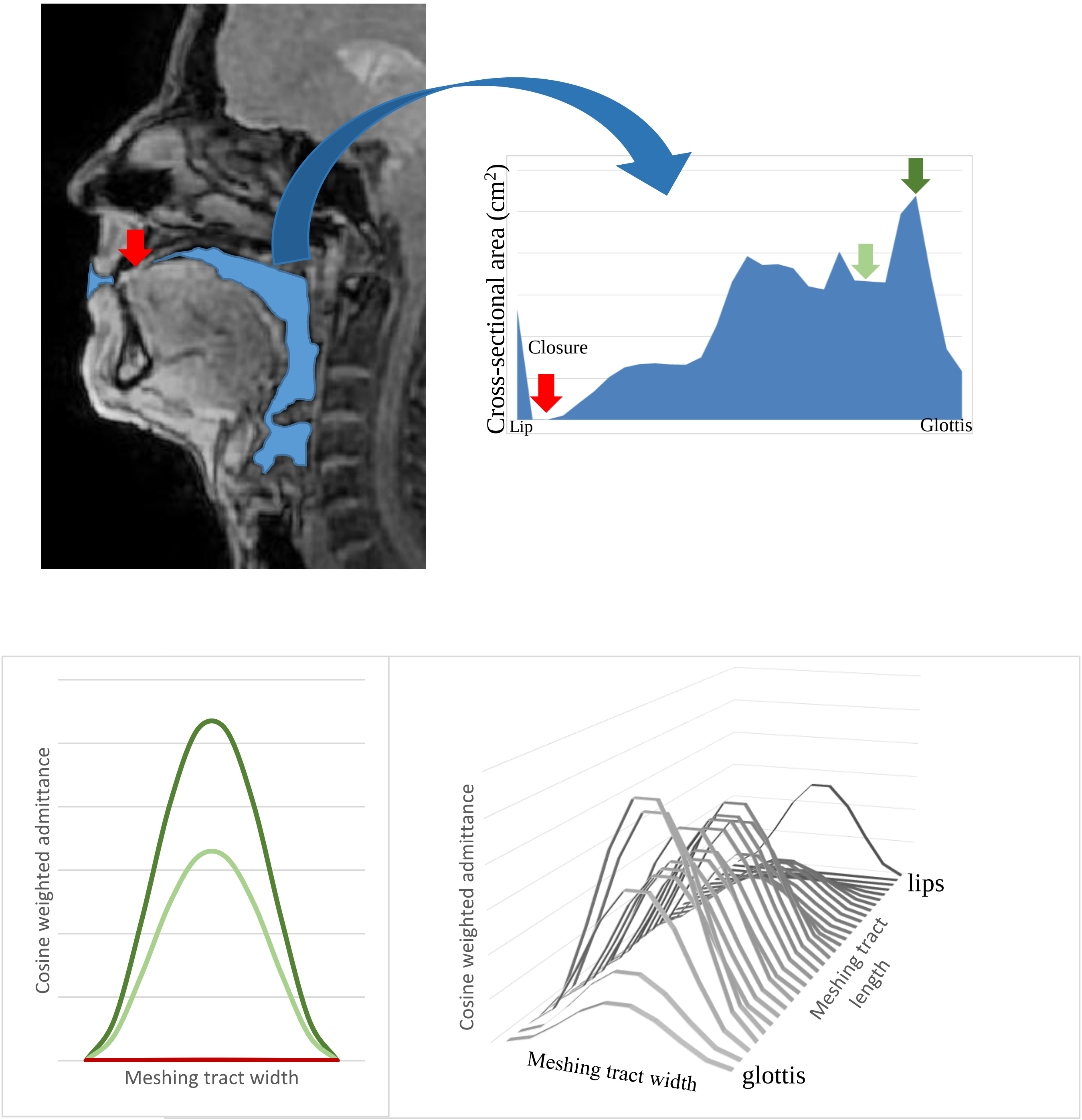


Figure 5.14. An example of impedance function mapping from MRI vocal tract image to cosine impedance hills in plosive simulation.

At closure, the lowest admittance/highest impedance is involved. The area function is 0, so the impedance reaches its maximum, $1/\text{admittance}_{\text{min}}$. Note that this research set the minimum admittance as 0.0001 kg-^{1}m^{4}s to avoid division by zero; therefore, the maximum impedance is 10000 kg m^{-4}s^{-1} for attempts to block the air flow at the closure.

Moving from the physical configuration to the dynamic setting and looking back to the three main acoustic properties for the plosive manner of articulation – the centre frequency of the burst, the VOT and formant transitions – the frequency of the burst is a result of the burst being modified by the local front and back cavities which relates to the place of articulation. It then depends on the physical setting in the modelling, while VOT is the result of setting the voice source injection at the glottis. The formant transition shows the frequency response of the coarticulation between plosive and adjacent phone. Considering one of the examples of articulatory trajectories from *mngu0* for /b/ with following vowel /A/, /u/ and /i/ in Figure 5.15, the example shows normalized trajectories of four places of articulation – lower lip, tongue tip, tongue body and tongue dorsum – as described in section 5.15. The coarticulation in the figure shows a rapid change of area function after plosion, then a short period of almost steady shape for the vowel, then change again for the coming phone. The movement after the plosion results in the formant transitions.

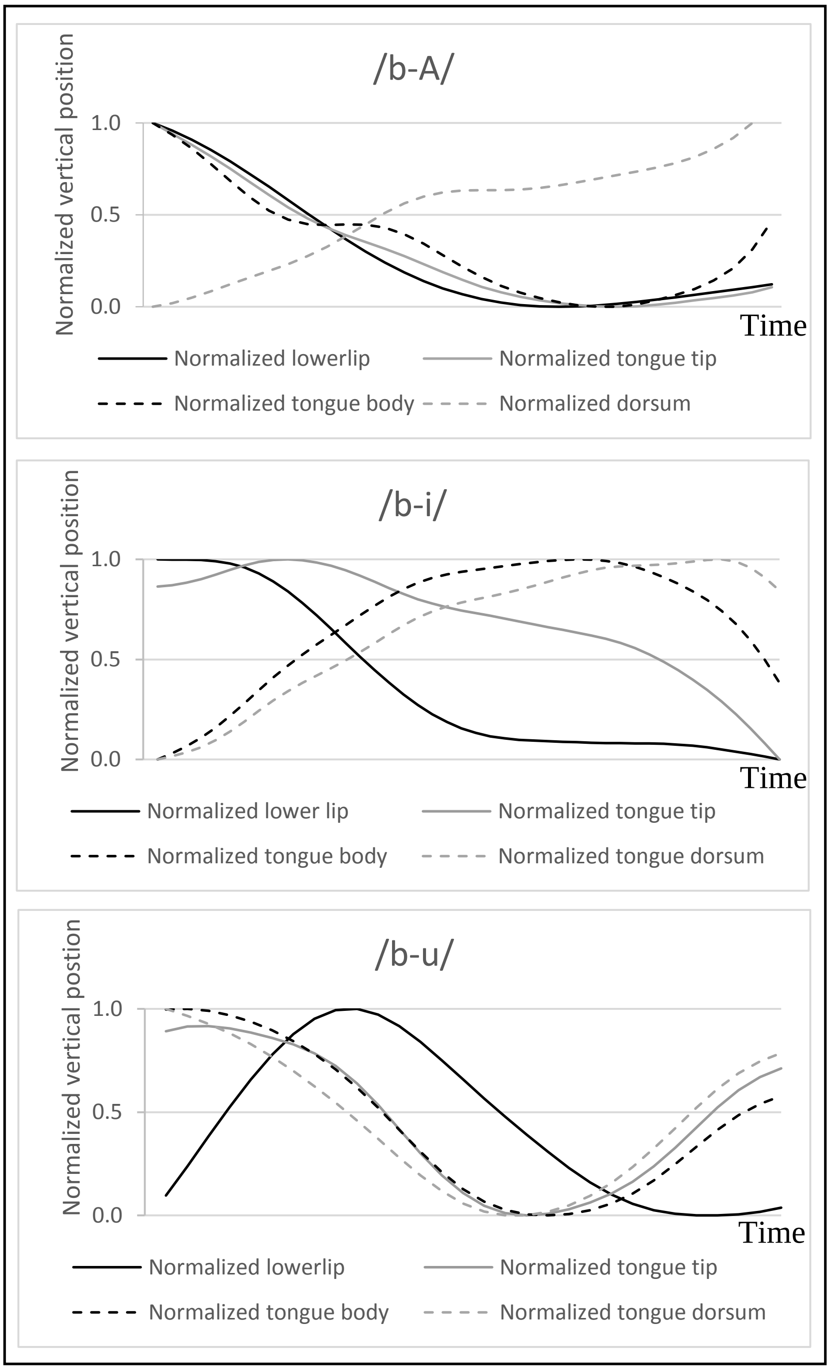


Figure 5.15. Normalized articulatory trajectories for /b-A/ (top), /b-i/ (mid) and /b-u/ (bottom).

5.2.4 Fricatives

The manner of articulation for fricatives involves holding the constriction of articulators long enough to create friction noise. However, as mentioned in the discussion on plosives, concerning the limitation of DWM’s inability to simulate the sound source, this research uses an extra noise source injected into the system at constriction for fricative production instead. Shadle modelled the physics of frication in her research in 1985 [70]. The two types of friction production for fricatives caused after obstruction in the tract are named obstacle and no-obstacle in her report. To model them, she used a dipole source in her transmission-line model for obstacle involved and just a source-tract interaction for no-obstacle involved. Her results show significant results for obstacle which contains /f/, /v/, /s/ and /z/. Later, Narayanan and Alwan claimed that source types have to be broken into monopoles, dipoles and quadrupoles, and injected them in different places, such as, dipole at the teeth and monopole at the constriction exit [71]. In 2000, Jackson used inverse filtered (coloured) white noise based on Shadle's regression curve [70] as the frication-noise source for his fricative study [71]. However, in this work, white noise is the only noise source used.

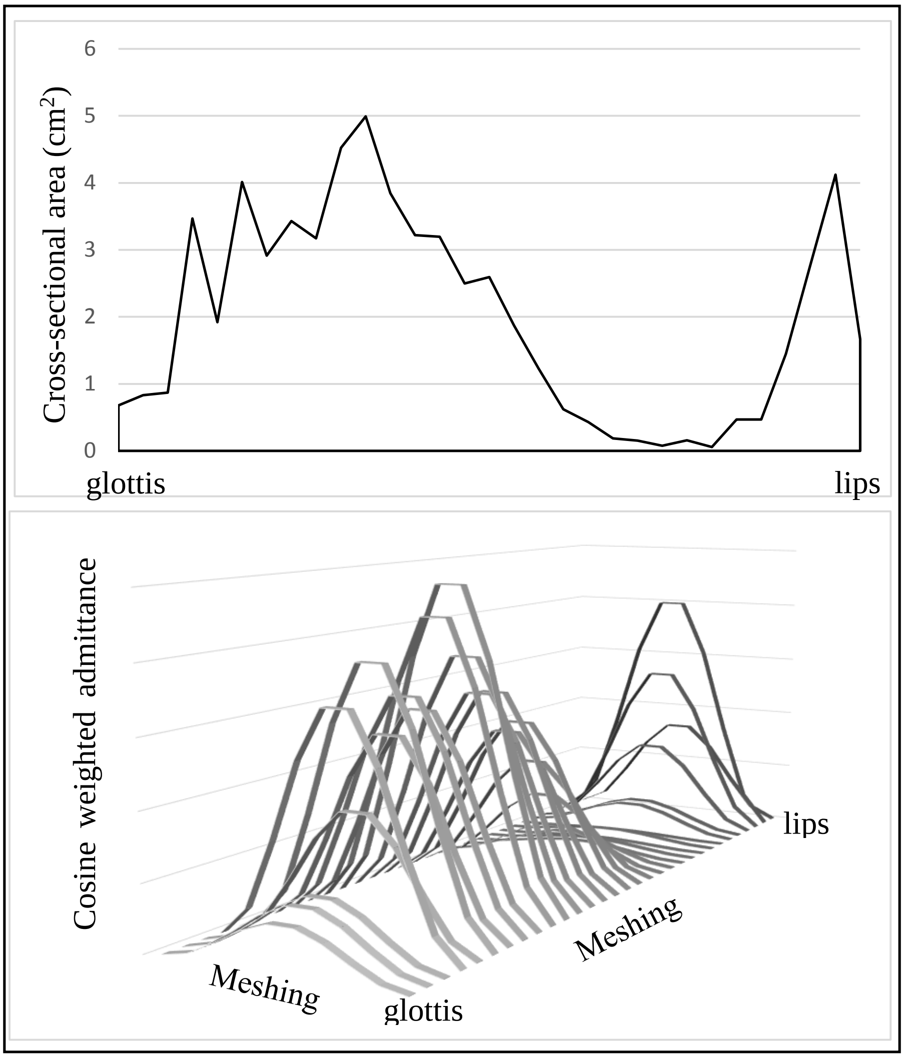


Figure 5.16. Vocal tract shape for /S/ in 2-D (top) and cosine weighed impedance converted from area function (bottom).

Meshing and cosine impedance control are also used for fricative synthesis. Figure 5.16 shows an example of the tract shape for /S/ (top) and its cosine admittance mapping (bottom). The coarticulation between fricative and carrier affects formant transition as usual (more details can be found in [72]). Therefore, articulatory trajectories from *mngu0* are also used to project the coarticulation in time frames.

Another important setting is the reflection coefficient at boundaries. Voiced and voiceless fricatives cause a different termination at the glottis. When the glottis is held open for a voiceless fricative, the reflection coefficient at the glottis is set at the same value as for open end at lips, -0.9. On the other hand, for a voiced fricative the vocal folds vibrate periodically to generate the voice source; the model assumes that it is essentially a rigid end at the glottis and therefore the reflection coefficient is set to 0.97. These values were verified and adopted from Mullen, 2006 [49].This causes different tract resonance properties, which is discussed in the next chapter. Moreover, as we did not focus on source generation but on filtering in the vocal tract, the naturalness of our results is poor for this manner of articulation. More details on how specific the source is, such as position-dependence gain, low-pass filtering etc., can be found in Steven, 1998 and Birkholz, 2014 [73, 74].

5.2.5 Affricates

As described in many books on linguistics, for example [72, 58, 20], affricates can be seen as a combination of a plosive and a fricative which is articulated quickly enough to produce a short burst followed by a short period of holding for frication; therefore the coarticulation between them plays the most important part. In this research, the friction behaviour is the main acoustic characteristic of this manner is which is not convincing enough because we did not use a proper friction source but white noise only (as discussed in 5.2.4). Consequently, the error can be accumulated from both plosive and fricative parts and so we have to leave the evaluation for naturalness of synthesized affricates and focus instead on the range of resonance frequencies. Their spectrogram is shown and discussed in the next chapter.

Resonances of friction from the tract are studied in this work rather than naturalness from affricate characteristics in *mngu0*. The spectrogram of recorded WAV files in *mngu0* shows roughly segmentable timing for burst and friction, and then the simulation starts in the same way as the other manners.

The cosine impedance area function works along the tract. The reflection coefficient reflects the pressure wave differently at four boundaries (glottis, lips and two more tube boundaries along the 2-D mesh). The articulatory trajectories create steps for tract change for articulation. The ladder between plosive and fricative is short but practical because the tract is just open for bursting and then keeps the constriction for slightly longer for frication before continuing the articulation to the adjacent vowel. Figure 5.17 shows interpolation between plosive and fricative parts and continues to the adjacent vowel.

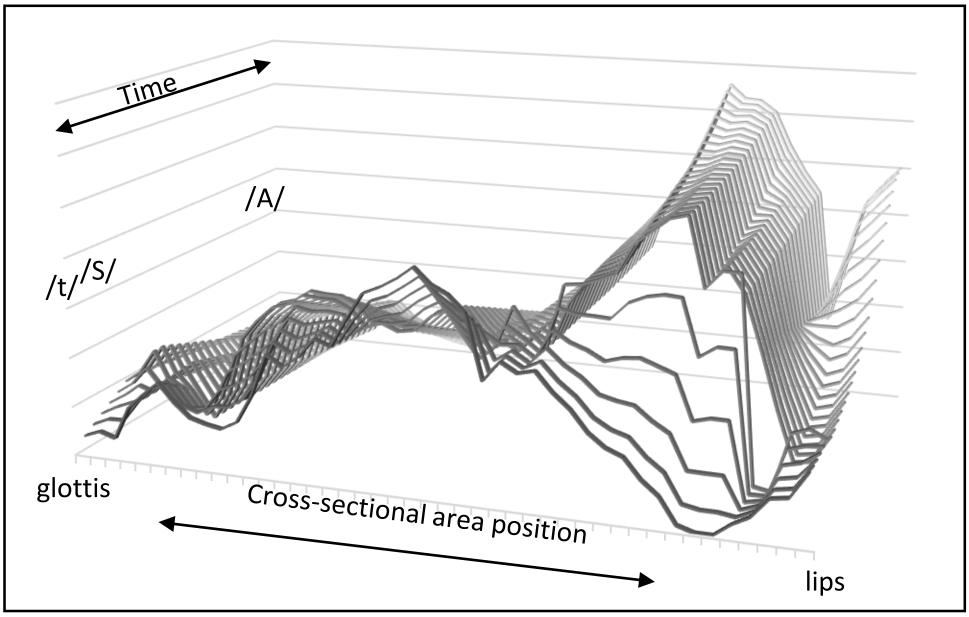


Figure 5.17. Tract shape interpolation when synthesizing /tS/ from plosive (very front) and fricative parts and continuing to the adjacent vowel /A/ (very back).

In summary, this chapter has described the tools, data, configurations and settings needed in each type of consonant synthesis experiment using 2-D DWM. Mullen's 2-D DWM is used as the wave propagation modelling tool. The LF model and white noise are used as sound sources. MRI from [60] is used as the tract shape image source. The articulatory trajectories from *mngu0* are used to guide the articulation. Images of changing tract shape in steps are plotted to show the articulation in the simulation. The synthesized sounds and their spectrograms will be shown and discussed in the next chapter.

Chapter 6
Evaluation and Results

The evaluation of the performance of the proposed system was conducted from two perspectives, objective and subjective. To evaluate and show the objective accuracy, the frequency response of the outputs from the proposed system are considered, comparing the acoustic properties of the synthetic sounds with those from natural speech. The subjective test gathers perceptual responses from audio engineers and phoneticians.

6.1 Objective evaluations

Objective evaluation experiments have been carried out separately following the phonetic manners of articulation for English: nasals, semi-vowels, plosives, fricatives and affricates.

6.1.1 Nasals

The two main characteristics of nasals, which are anti-formants (zeros) and the nasal murmur, were introduced in Chapter 4. This chapter looks at those characteristics in the synthesized speech output from the 2-D DWM. In this experiment on English nasals, acoustic pressure variations are propagated through both nasal and oral cavities, but are blocked from emerging from the mouth, due to a complete closure at the appropriate place of articulation.

This blocking causes absorption at the resonant frequencies of the oral cavity, which are the nasal anti-formants. The size of the oral cavity depends on the place of constriction, which affects the frequencies of the anti-formants. Therefore, evaluation of the absorption is achieved by making closures in the oral tract. Figure 6.1 shows absorption results from making a constriction at the bilabial, alveolar and velar place of articulation by closing the slider No. 15, 13 and 11 which can be seen as the first set of results in attempting to simulate /m/, /n/ and /N/, respectively. This experiment is initialized by setting the tract shape as for producing /3/ and then recording nasalized /3/, nasalized /3/ with a bilabial closure, nasalized /3/ with an alveolar closure and nasalized /3/ with a velar closure.

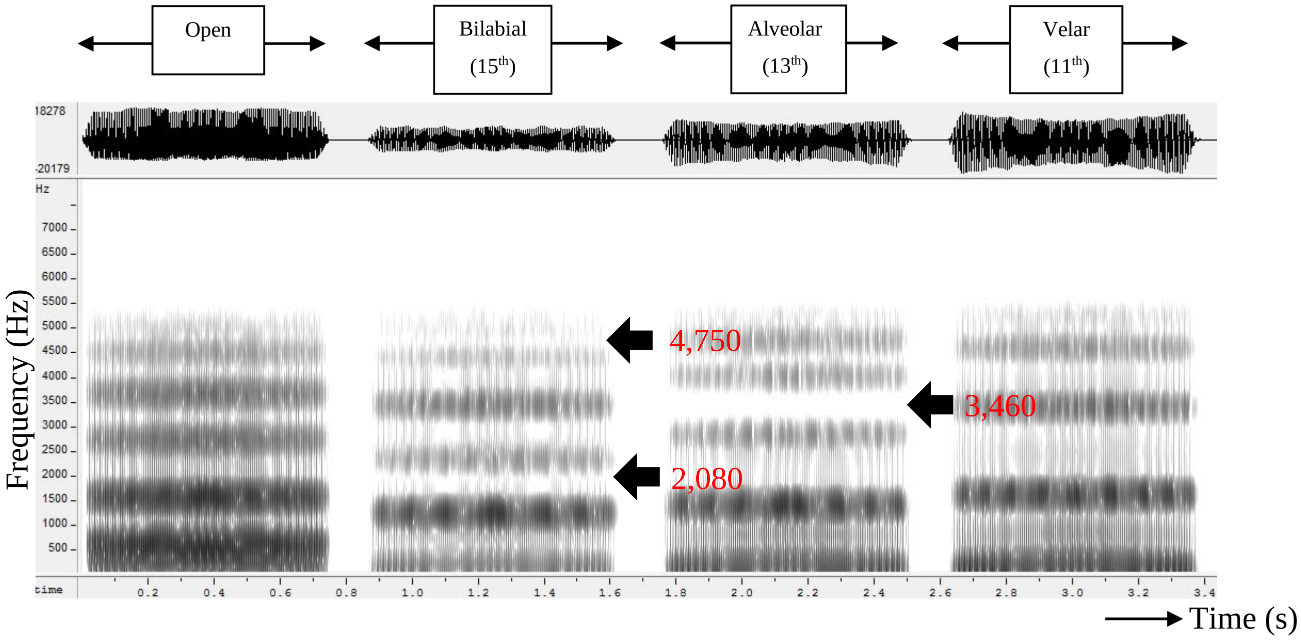


Figure 6.1. A spectrogram of four examples of synthesized speech of nasalized /3/ and nasalized /3/ tract shape with bilabial, alveolar and velar closure. From left to right: the first is from the open mouth; the second is from adding a closure at slider No. 15 for bilabial; the third is from closing at slider No. 13 for alveolar; and the last is from closing at slider No. 11 for velar.

In Figure 6.1, three black arrows indicate anti-formants on a narrow-band spectrogram. Note that the velar one does not show any anti-formant because the branch tubes are too small and the anti-formant is shifted to be at a high frequency. The anti-formants can be viewed more clearly in the long-term average spectrum (LTAS) analysis shown in Figure 6.2, which is based on the use of a Hamming window and 0.98 pre-emphasis. The figure shows three anti-formants indicated by the black arrows, which support the results in Figure 6.1. For the murmur, Figure 6.2 shows the first peak at around 130 Hz which is the fundamental frequency (F0) of the voice source; the second peak at around 260 Hz (which is the second harmonic) corresponds approximately to the expected nasal murmur (around 300 Hz) as described in [20]. Therefore, the murmur of synthesized speech in this experiment is firstly assumed to be the peak that we can see at 260 Hz, because this is the harmonic in the output which is closest in frequency to the expected 300 Hz nasal murmur.

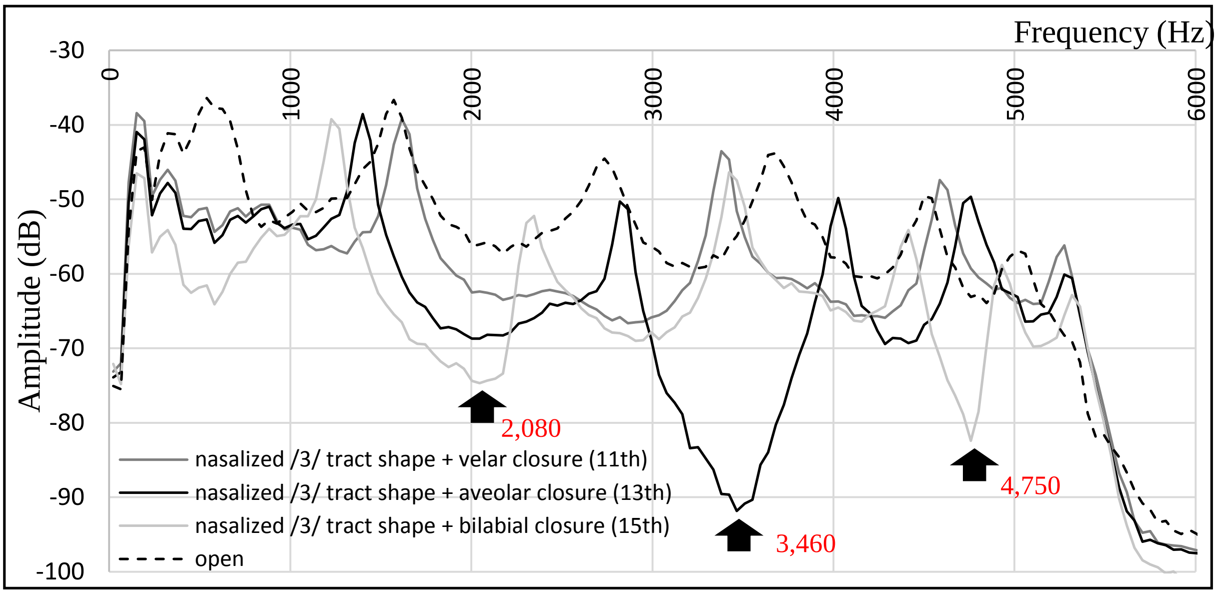


Figure 6.2. FFT long-term analysis of the synthesized speech in Figure 6.1.

The effects of nasalization in the proposed system are shown again here, a set of four selected vowels (/A/, /3/, /i/, /u/) having been synthesized in nasalized and non-nasalized conditions with neither additional constriction in the oral tract nor extra noise excitation source. Figure 6.3 shows their frequency analysis results.

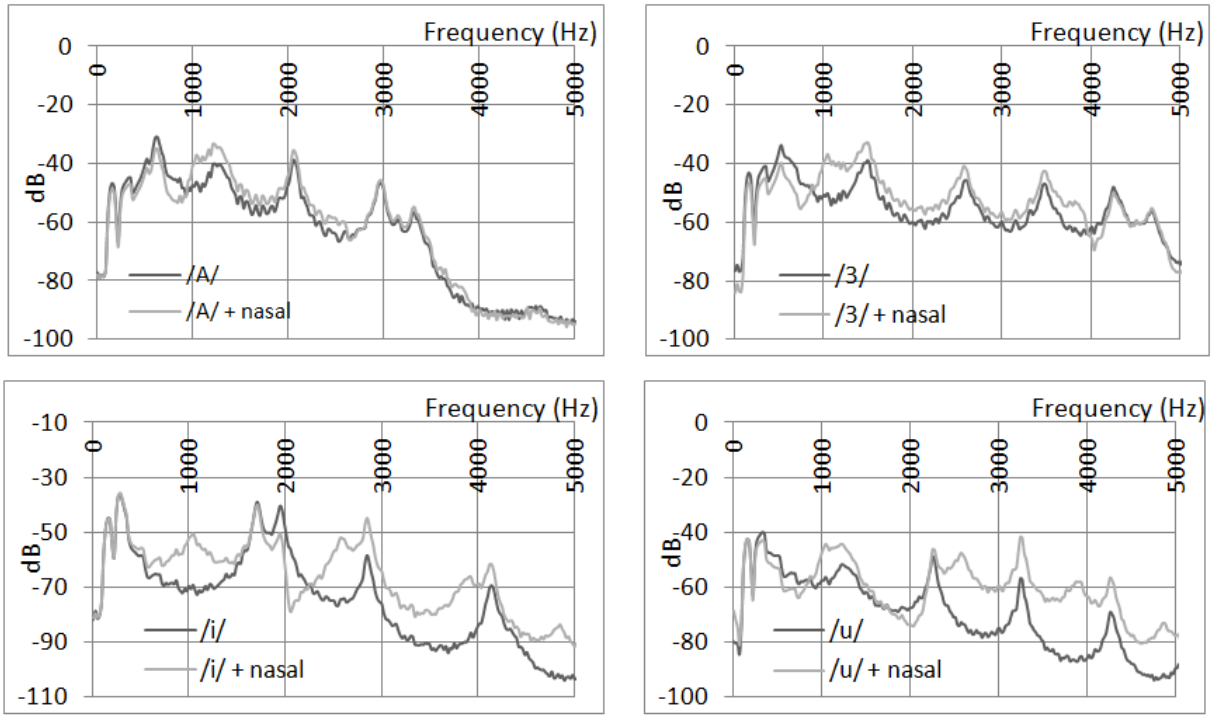


Figure 6.3. FFT of four nasalized and non-nasalized 22 kHz synthesized vowels: /A/ on the top left, /3/ on the top right, /i/ on the bottom left and /u/ on the bottom right.

The results show that peaks at 260 Hz appear in both nasal and non-nasalized vowels. Based on the nasal murmur characteristic, which is nasal formants relating to the nasal plus pharyngeal cavity, the F1 of nasal murmur should have higher amplitude followed by less amplitude caused by damping [58]. The damping of the resonance is caused by absorption in the complex nasal wall in the nasal cavity. Cox explained the results of damping as follows: "The most general effect of adding nasal resonance to oral resonance is an overall loss of power." in [75]. Kent and Read drew the idealized damped amplitude of nasal vowels compared to non-nasalized vowels in Figure 5-34 in [32]. However, 2-D DWM is used to propagate pressure only, not absorption. The managing of the wall reflection coefficient as in [76] can cause a slightly different bandwidth but not enough for nasal damping; hence, the energy of nasal murmur still exists in the system and is carried out through the end of the nasal tract. Therefore, we can see the second formant of nasal murmur at around 1 kHz in all results which can be predicted by (3 x 343) / (4 x 0.22), 3c/4L, which equals 1,169 Hz.

In detail, these first two sets of results were synthesized using a 22 kHz sampling rate, which means the waveguide length is 0.022 m. The oral cavity is 0.075 m in length in these vocal tract simulations, which means that it is simulated by the last three waveguides (0.022 x

3 = 0.066 m) (Note that width is not discussed here, because resonation considers the length of the tract only.) Of these three, the last represents the pressure propagation at the lips, the penultimate represents the alveolar, and the one before that represents the palatal approximately. These three places were the places of closure mentioned in the first set of results in Figure 6.1. The LTAS of the first attempt (shown in Figure 6.2) shows three anti-formants, two of which are from bilabial closure (2,080 Hz and 4,750 Hz) and the third from alveolar closure (3,460 Hz). They are the results of absorption in the oral cavity. When closing slide No. 15 for bilabial closure, the oral tract is two waveguides long which equals to 0.044 m; therefore, for this quantization, we can expect to see anti-formants at (1 x 343)/(4 x 0.044) = 1,559 Hz and (3 x 343)/(4 x 0.044) = 4,677 Hz, and we get 2,080 Hz and 4,750 Hz from the effects of using /3/ tract shape rather than a uniform tube. In the same way, closing slide No. 13 for an alveolar closure, the oral tract is 0.044 m in length; hence, we can expect to see an anti-formant at (1 x 343)/(4 x 0.022) = 3,897 Hz, for which we get 3,460 Hz.

We next evaluate the proposed system in a more dense mesh. This next experiment is another attempt to synthesize English nasals with a more dense mesh, by changing the sampling rate from 44.1 kHz to 88.2 kHz and 176.4 kHz. The results from the 88.2 kHz version are selected to be shown here. The resonances from using real tract shapes of /m/, /n/ and /N/ are shown in Figure 6.4. The long-term spectrum shows anti-formants at around 1,300 Hz in /m/, 3,000 Hz in /n/ and 6,400 Hz in /N/.

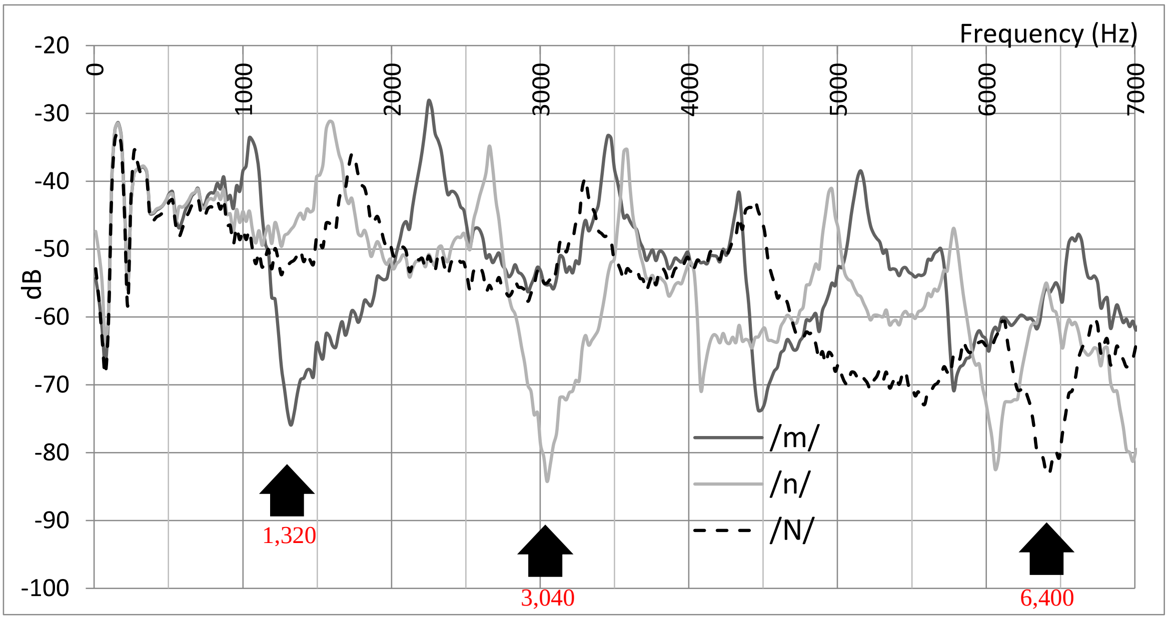


Figure 6.4. Long-term analysis of synthesized /m/, /n/ and /N/ from 88 kHz of sampling frequency.

The next figure shows comparisons between LTAS results and those from the Praat articulatory synthesizer [38]. Usually Praat is used as a speech tool for studying, analysing and modifying human speech; however, it also has a function to synthesize human speech using the articulatory synthesis method. The synthesizer has flexible walls for forcing and passing pressures and velocities with a GUI for controlling tensions of vocal tract muscle apparatus. Here, Boersma labels his methodology as functional phonology [39]. LTAS of three English nasals from Praat are shown together with results from our synthesizer (Mullen). The abbreviations f_m, $af_{/m/}$, $af_{/n/}$ and $af_{/N/}$ stand for theoretical murmur frequency, anti-formant of /m/, anti-formant of /n/ and anti-formant of /N/, respectively [20].

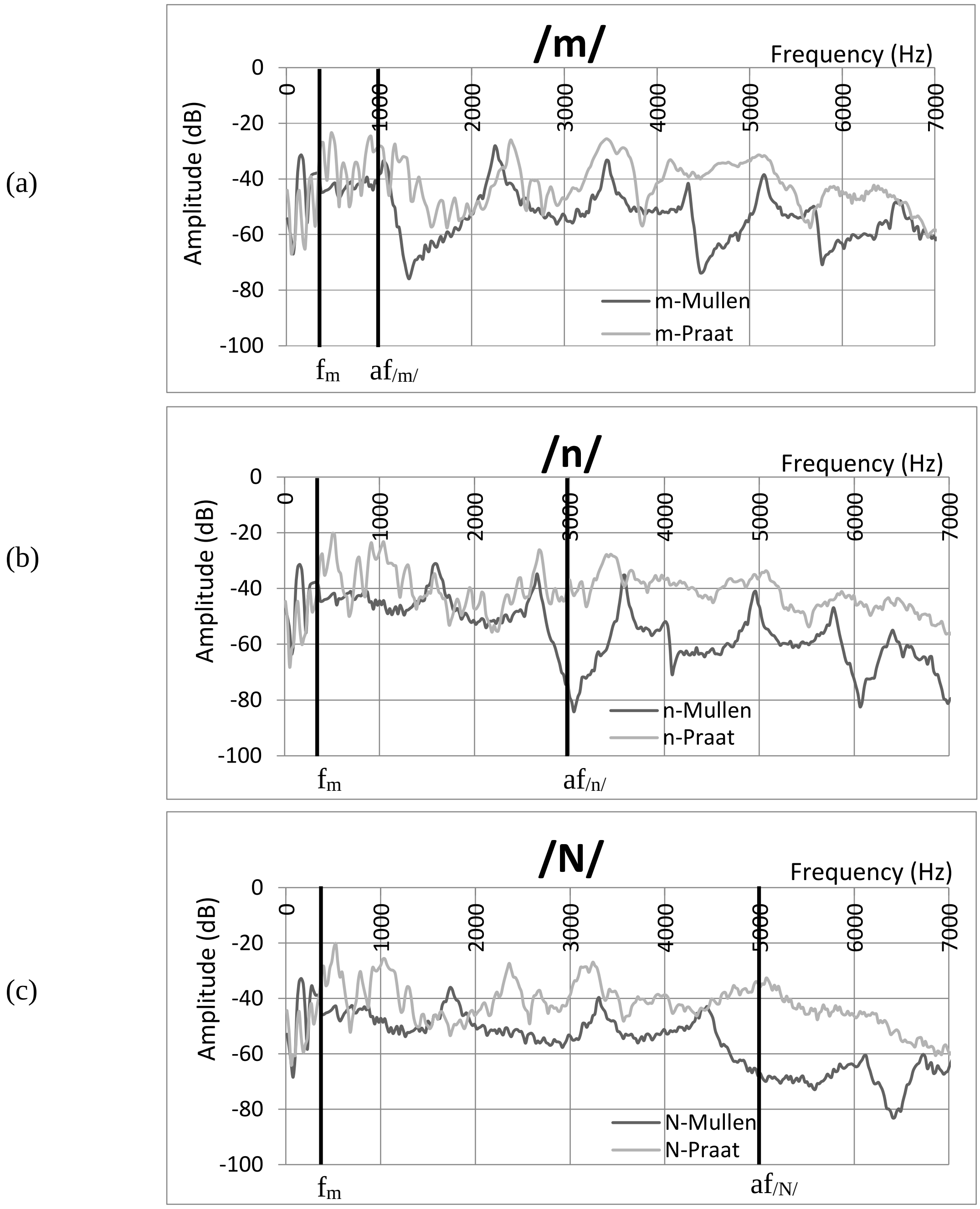


Figure 6.5. FFT shapes of noisy nasalized synthesized speech from articulatory synthesizer in Praat: (a) for /m/, (b) for/n/ and (c) for /N/.

6.1.2 Semi-vowels

The evaluation of semi-vowel synthesis was considered by setting up the system according to the configurations in section 5.2.1 and then recording the output for acoustic analysis. One characteristic of a semi-vowel is that it is similar to a diphthong but with a quicker articulatory transition; the acoustic characteristics of this set are shown in terms of resonance change caused by movement of the articulators, or change in area function (formant transition). Two semi-vowels are considered in this work, /j/ and /w/; therefore, the formant transitions between /i/ and an adjacent vowel and /w/ and an adjacent vowel are discussed here.

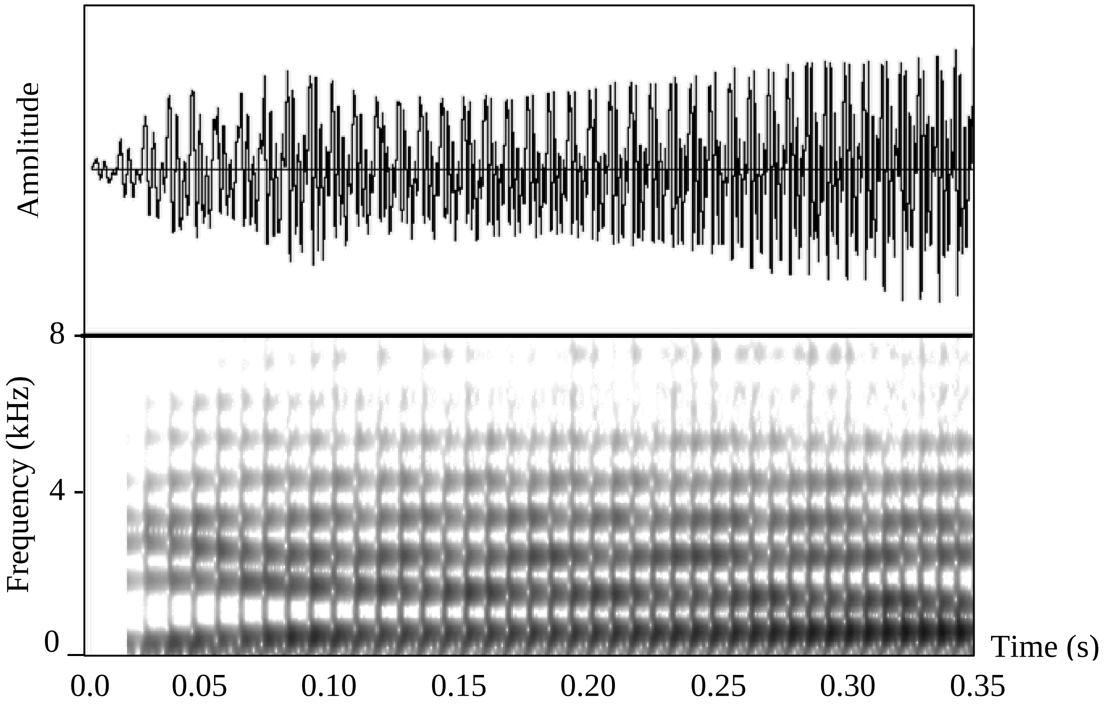


Figure 6.6. Formant transitions in synthetic /jA/ using Wavesurfer wide band spectrogram analysis.

The spectrogram in Figure 6.6 came from adopting the articulation according to articulatory trajectories from *mngu0* [77], so the duration of /j/ depends on trajectory data. For synthesizing /jA/, it uses 0.334 sec for the trajectories. Figure 6.7 shows the spectra of the beginning and end of the synthesized /j-A/.

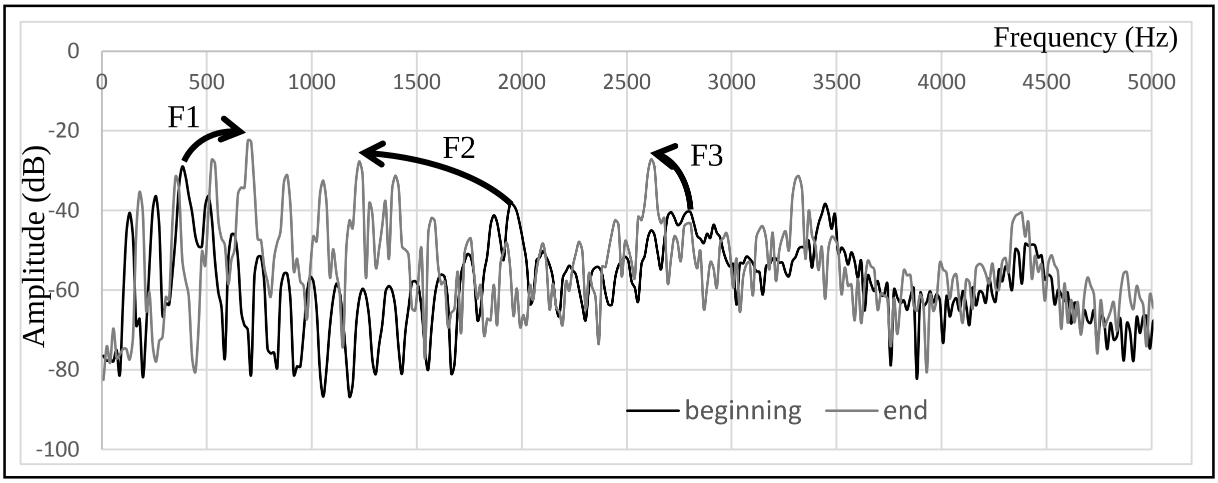


Figure 6.7. Spectra of the beginning (black) and end (grey) part of the synthesized /j-A/.

We can see that, at the beginning of the spectrogram, the first formant starts at 344 Hz and then rises to 648 Hz for /A/ while the second starts at 1,944 Hz and then falls to 1,316 Hz for /A/, similarly to the third formant which starts at 2,875 Hz and then falls down to 2,632 Hz. The figures are shown in Table 6.1 in a comparison to synthesized sound from holding the area function steady for individual /i/ and /A/, their first, second and third formants respectively being 344, 1,944 and 2,855 Hz for /i/ and 688, 1,316 and 2,632 for /A/ which are the same as those at the beginning and end of the synthesized sound.

Table 6.1. Formant frequencies comparison between those from vowel and semi-vowel synthesis for /jA/.

	Formant frequency (Hz)					
	Beginning of synthetic /jA/	End of synthetic /jA/	Synthetic /i/	Synthetic /A/	Theoretical /i/ for men from [20]	Theoretical /A/ for men from [20]
F3	2,875	2,632	2,855	2,632	2,550	2,450
F2	1,944	1,316	1,944	1,316	2,000	1,100
F1	344	648	344	688	400	730

The transition between /i/ and /A/ plays the most important part here. As discussed in [20], theoretically we can consider the two semi-vowels /j/ and /w/ as their closest vowels (/i/ and /u/ respectively but articulated more quickly, so these transition durations are seen as the duration of formant transitions in the spectrogram and they show up as smooth transitions due to the dynamic articulatory changes in the 2-D DWM. Hence, these results confirm that the system could support a synthesis of /j/ by including appropriate articulatory trajectories and changing the resonance frequencies relevant to the area at the corresponding position smoothly (see perturbation theory in Chapter 4 for details of the corresponding position). More details about the perception of the acoustic behaviour of the synthesized sound will be shown and discussed in the subjective evaluation section later.

The spectrogram of two studied semi-vowels are plotted in Figure 6.8. It presents the spectrogram comparison between recorded speech (left) and synthetic (right). The top two are the spectrograms of /jA/ and the bottom those of /wA/; the synthetic in each pair adopted the trajectories that recorded synchronizing with the recorded ones.

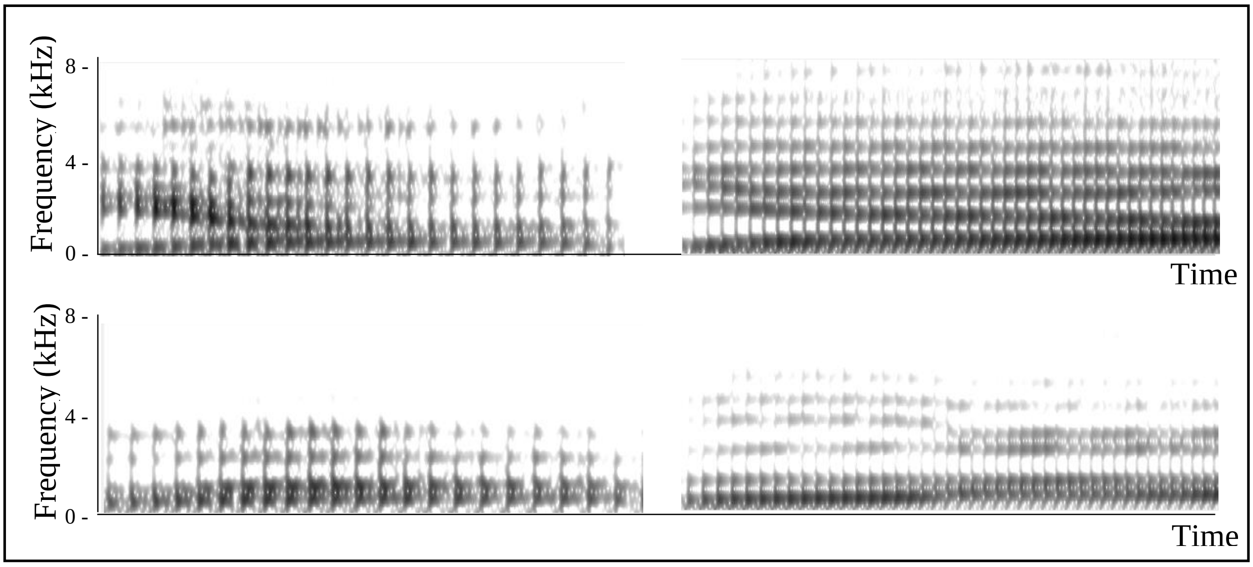
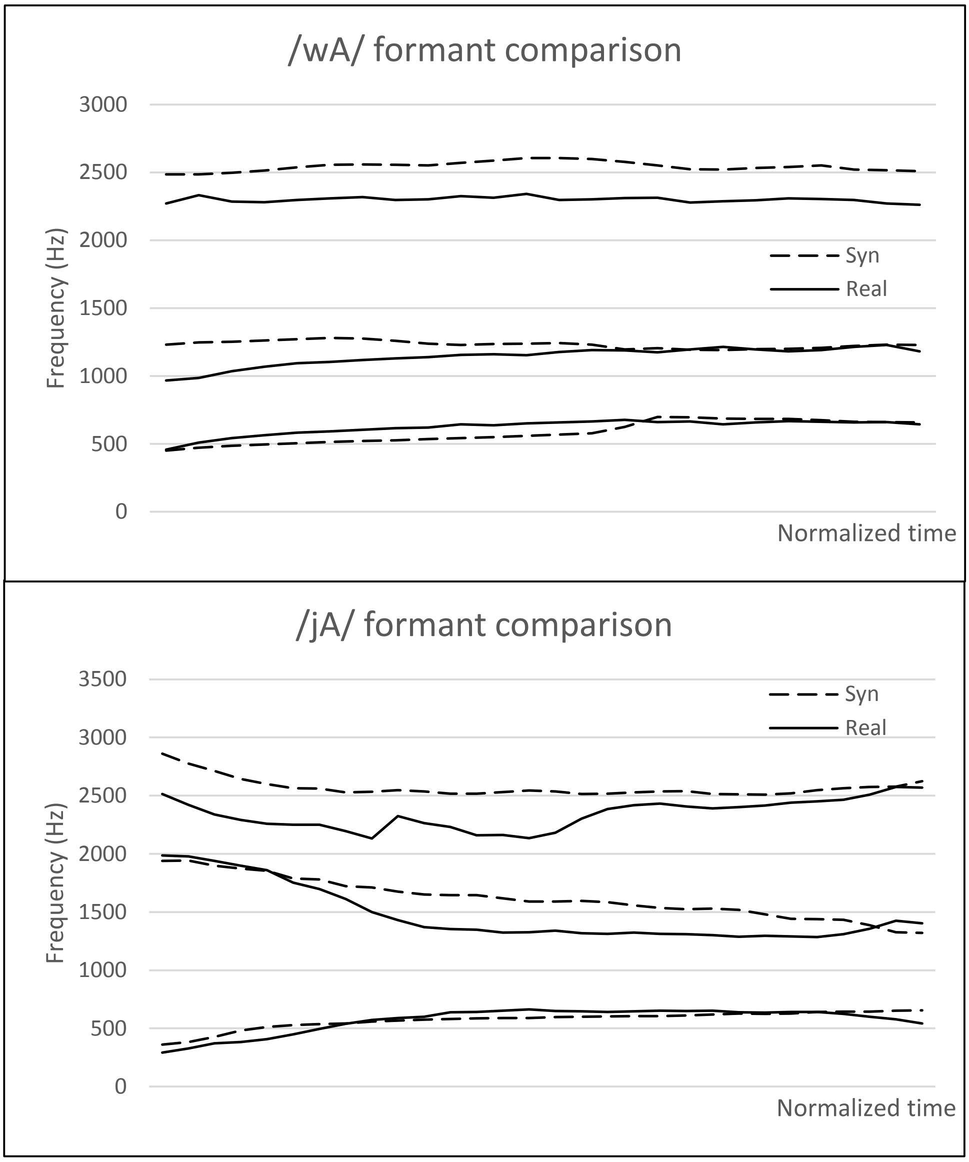


Figure 6.8. Spectrogram of recorded /j/ and /w/ (left) and synthesized /j/ and /w/ (right).

The figure shows the similarities of formant transitions, which relate to vocal tract shape variations. From close comparisons, Figure 6.9 pairs the formants from synthesized and recorded sound together. According to the adoption of trajectories of one to a different subject who gave his vocal tract shape scanned in MRI, all tract shapes are not the same but similar in the process of articulating for the same phone; therefore, the comparison will show the similarity of transitions only, not the formants. The left shows those of /wA/ and the right shows those of /jA/. All formant transitions in /jA/ are similar: F1 moves up, F2 moves down at a slower rate than in real speech, and F3 moves down at the same rate. In /wA/, F1 and F3 move slightly upwards, while F2 is slightly different which is a result of the different tract shape at the beginning in that the proposed system did not put an extension for round lips which makes it unrounded (F2 in an unrounded vowel is higher than in a rounded [78]).


/wA/ formant comparison

/jA/ formant comparison

Figure 6.9. Formant transition comparison between that of synthesized (dashed line) and real (solid line).

In addition, the adoption of articulatory trajectories is done in discrete timing control. The trajectory was adopted in length/duration as it was recorded. For more information about phone duration in the *mngu0* corpus that was used as the articulatory trajectory guideline, we examine the duration behaviour more closely. There are 198 and 972 diphones which pair with /j/ and /w/ respectively. In these there are 5 pairs of /j/ followed by /A/, 3 followed by /i/ and 284 followed by /u/. Figure 6.10 shows their durations.

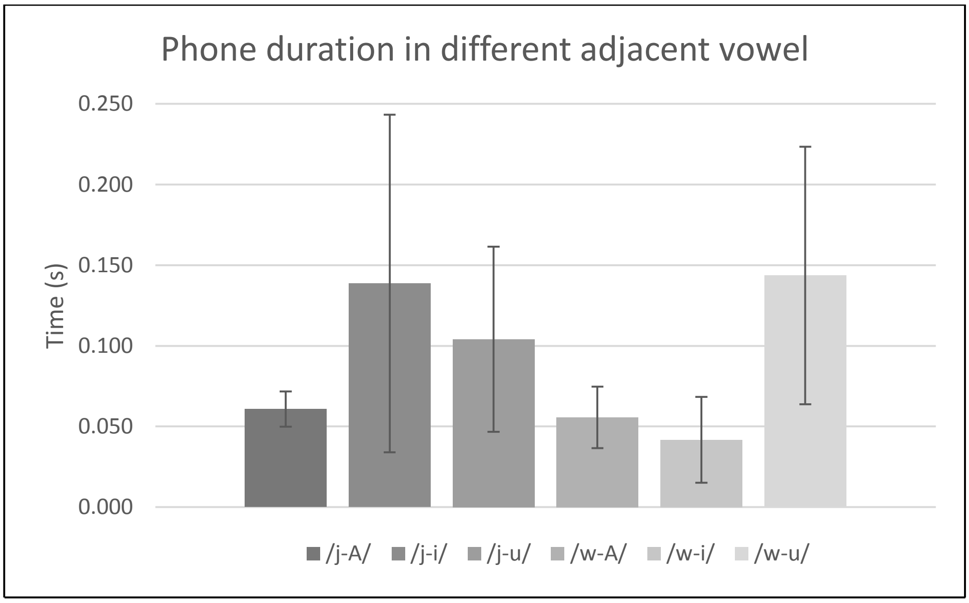


Figure 6.10. Average recorded duration in mngu0 of /j/ followed by /A/, /i/ and /u/ are labelled as /j-A/, /j-i/ and /j-u/ while those of /w/ are labelled as /w-A/, /w-i/ and /w-u/.

The figure shows the variation in phone durations when it pairs with different vowels. The average duration for /j/ and /w/ in the corpus is calculated but not used, because the real articulation will have to be modified for the average duration and we want to move our area function after the real articulation to evaluate our system’s performance. However, all these attempts to use articulatory trajectories recorded from one subject to guide articulation simulation for another subject’s vocal tract shape show successful results for semi-vowel simulation.

6.1.3 Plosives

Three main acoustic properties – centre frequency of the burst, formant transition and voice onset time – are evaluated separately because they relate to different stages in plosive production. The burst length is controlled by the release stage, formant transitions are controlled by articulatory trajectories after the release and/or closing stage, and the VOT is controlled by timing the voice source onset from the glottis.

Although 2-D DWM does not simulate flow which makes it unable to simulate/generate the sound source, the effect from quickly opening the tract in our simulation with a pressure source injection at the glottis could generate a small burst. The first set of results arises from an attempt to try to see the burst.

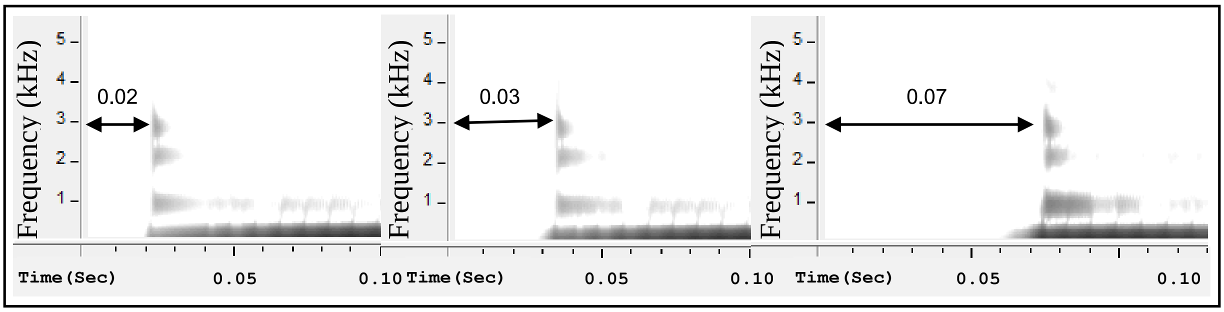


Figure 6.11. Spectrograms of bursts from different lengths of hold stage, 0.02, 0.03 and 0.07 seconds (indicated by arrows), analysed using Wavesurfer [79].

Figure 6.11 shows the small or transient bursts resulting from different lengths of hold stage, 0.02, 0.03 and 0.07 seconds, respectively. The pressure source is accumulated behind the closure while the cosine impedance function allows a small leakage but still blocks the pressure flow. The results show that the longer the duration of closure, the stronger the burst's energy. It also depends on other factors, such as amplitude of the source. However, Ogden [55] mentioned that in real speech frication is caused by narrow constriction after sudden release. Limitations on frication in 2-D DWM are discussed in section 6.1.4 (fricative). Therefore, in plosive an extra noise injection is needed when the system is set to mimic a recorded plosive from real speech, the additional noise source being used to hold the energy slightly longer, as it appears in the spectrogram of real speech.

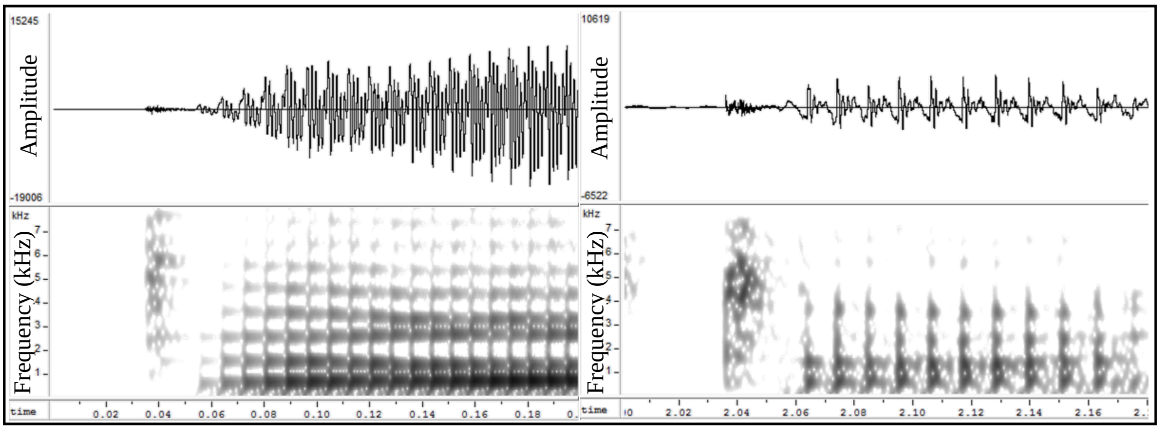


Figure 6.12. Spectrogram of a recorded chunk of /tA/ from *mngu0* sentence No. 0455 (right). Synthetic waveform of /tA/ with additional noise source injected at the place of burst and its spectrum (left) analysed using Wavesurfer.

The result in Figure 6.12 comes from an attempt to carry the frication energy from an extra noise source injection. The system is set to mimic the spectral frequency of the burst from recorded /tA/ (right). White noise was injected after the closure position, according to [80]'s work on frication production, immediately after the tract was opened. The result shows a similar range of frequency, which resulted from front cavity resonances. The dissimilarity of the range of frequency could come from different tract shape and position of articulation from a different source (speaker).

In this section, the range of frequencies plays the most important role and it is resonated from the front cavity. The next set of results then shows resonances from different places of articulation which affect the frontal cavity size according to the recorded MRI tract shape for /p/, /t/ and /k/. Figure 6.13 shows the frequency range results.

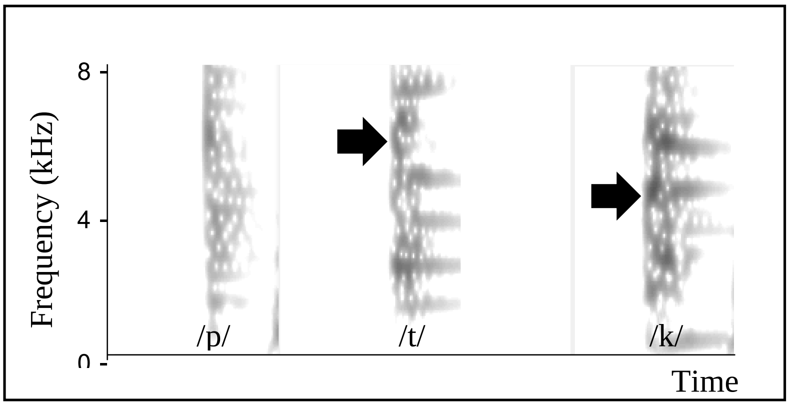


Figure 6.13. Spectrogram of the burst from different places of articulation: /p/ (left), /t/ (mid) and /k/ (right).

The resonance spread over almost all frequencies for /p/, about 6 kHz for /t/ and 4.5 kHz for /k/. This is a slightly unexpected range according to [20] who suggests that the centre frequency of burst for an alveolar (/t/) should be a little above 4 kHz and for /k/ should be between 1.5 kHz and 4 kHz. However, the results conform with the discussion on vocal tract filter functions in Johnson, 2012. He mentioned that the spectrum of bursts from labial stops has no formant peaks and that energy is spread diffusely, and that of alveolar has higher frequency peaks than that of palatal because of the shorter front cavity.

Another characteristic of plosives, formant transitions, was investigated with reference to Johnson’s F1 and F2 transition chart of plosive formant transitions in Figure 6.14. Three sets of results are shown in Figure 6.15. Those making use of /i/, /A/ and /u/ as the following vowel show comparative results. The results were analysed using a narrow-band spectrogram. Those for /b/ all show a small rise at the start for all formants. Those for /d/ show a small rise for the first formant and a rise in a transition to /i/ but falling in transition to /A/ and /u/. The first formant transitions for /g/ are all rising while the second are all falling. These are similar to the ideal plot of formant transition in [20].

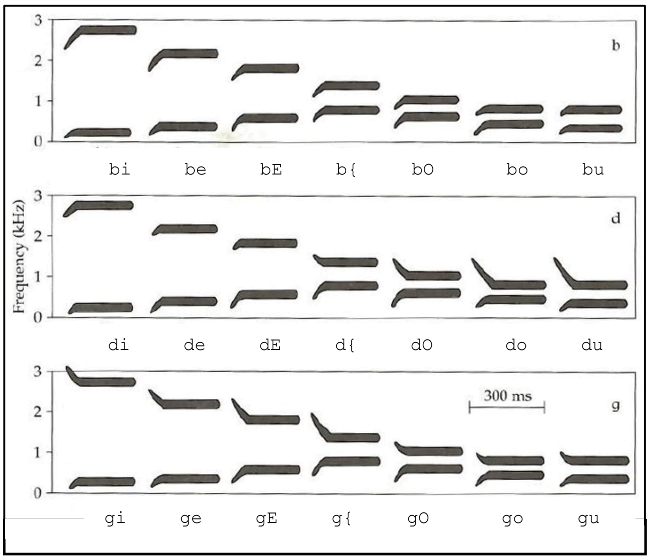


Figure 6.14. Johnson’s F1 and F2 transition patterns adapted from Delattre et al., 1955 [81] (adapted from [58]).

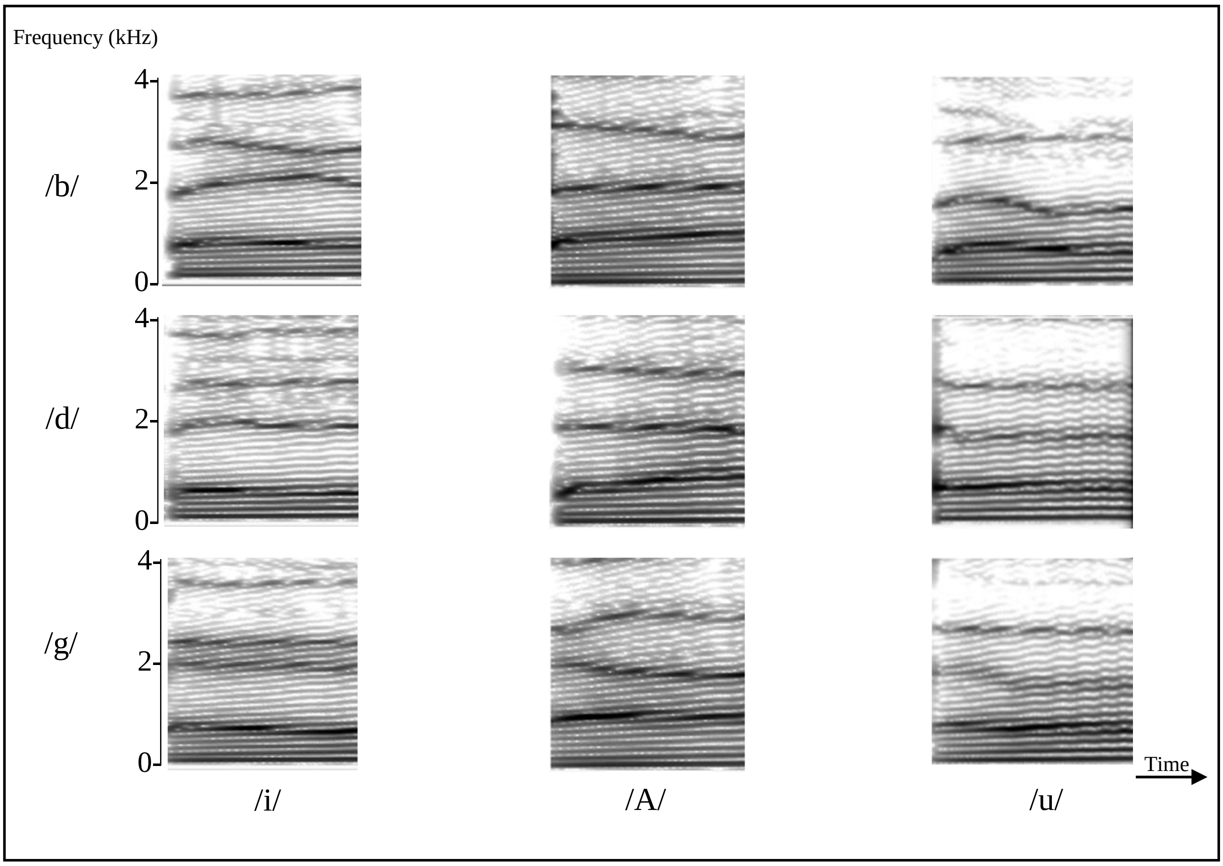


Figure 6.15. Examples of formant transition from synthesizing /b-i/, /b-A/, /b-u/, /d-i/, /d-A/, /d-u/, /g-i/, /g-A/ and /g-u/.

The ideal plot shown in Figure 6.16 [20] depicts the effects of changing places of articulation that change the loci of the second formant. Our results also show that bilabial /b/ second formant's loci are lower than their formant frequency, while those of alveolar /d/ are either lower for a front vowel or higher for a back vowel, and those of velar are higher than their formant frequency whether they are followed by a front or a back vowel.

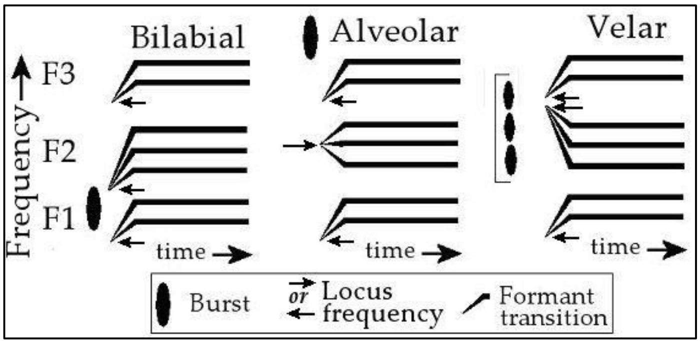


Figure 6.16. Ideal plot of formant transition from Howard, 2008 [20].

The comparisons between formant transitions from real and synthesized speech are plotted in Figure 6.17. Dashed lines represent those of synthesized speech while solid lines represent those of natural speech. All F1 and F2 transitions are similar to real speech, but F3 varies in the synthesized sounds where its transitions are not as fluctuated as in natural versions. F2 in transitions with the vowel /u/ are more different from real ones because of the unrounding effect.

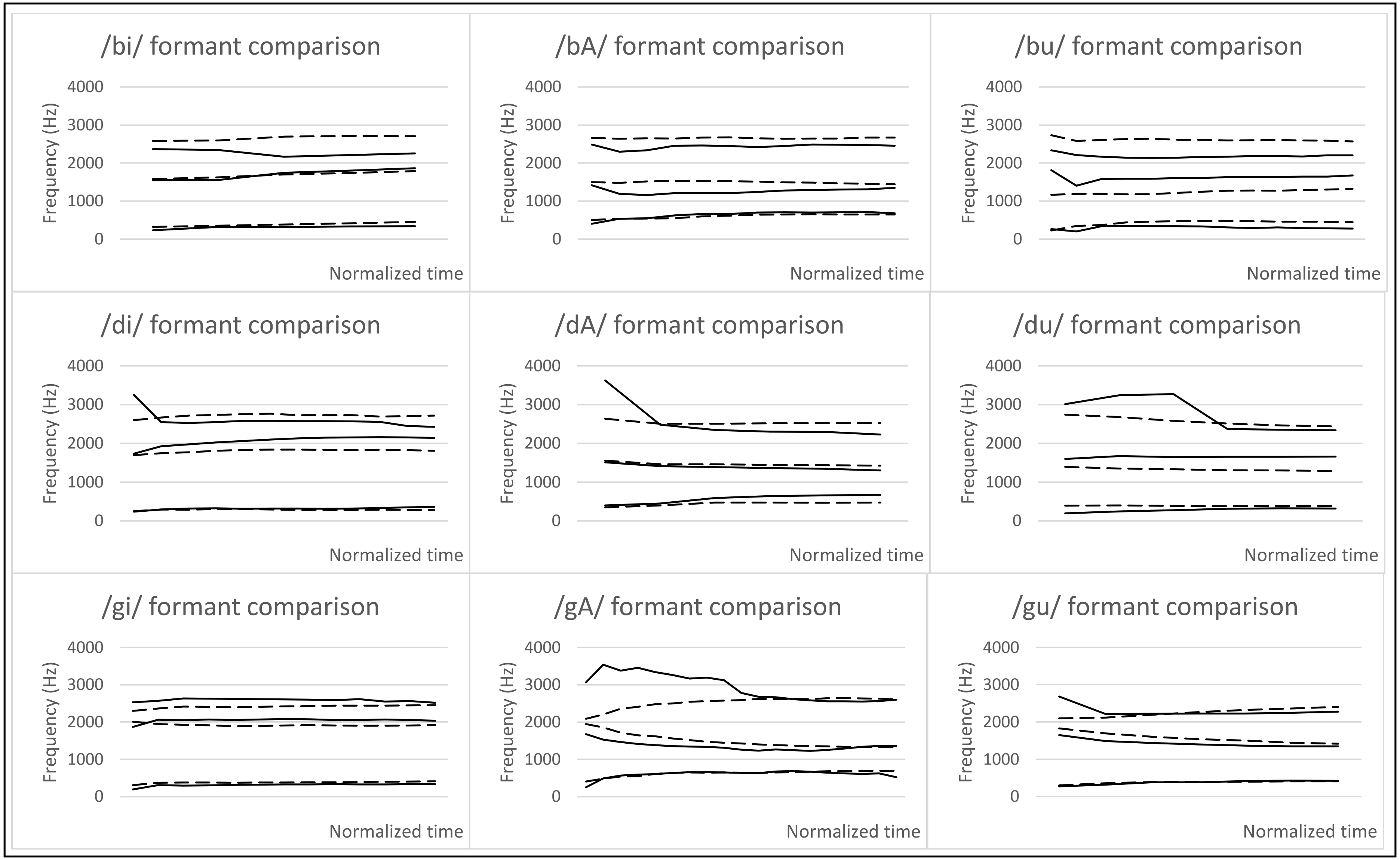


Figure 6.17. Formant transitions comparison between those from real speech (solid line) and those from synthesized sound (dashed line).

6.1.4 Fricatives

To pronounce a fricative, the glottis has to be held open, letting air flow through the tract and being constricted at place(s) in the mouth, except in the case of /h/. Turbulent airflow is produced at the constriction and then filtered by the vocal tract. The vocal folds are open for voiceless fricatives but vibrate for voiced fricatives. This means that the glottis is represented as closed or open end in the simulation, as appropriate.

As mentioned in Chapter 5, this work does not implement any friction source simulation (details on frication simulation can be found in Badin, 1989 [82]; Shaddle, 1991 [80]; Jackson, 2000 [71]; Birkholz, 2014 [74]). Therefore, frication is simulated using white noise injection only, without filtering or band pass before injection. The resonances primarily came from the front cavity only, because the coupling between front and back cavities is weak. Comparing resonance frequency from our system with the ideal from Stevens, 1989 mentioned in (Johnson, 2012 [58]), yields the results shown in Figure 6.18.

The comparison between Figure 7.4 after Stevens, 1989 in Johnson, 2012 to our resonance frequencies in Figure 6.18 shows a similar frequency response to that which appeared in Stevens. A constriction is set at a time at different distances from the mouth: 0.0 cm, 1.5 cm, 2.2 cm, 3.5 cm and 5.9 cm. The response is a peak at 6.5 kHz from constriction at 1.5 cm, a peak at 4 kHz from constriction at 2.2 cm, a peak at 2.65 kHz from constriction at 3.5 cm and two peaks at 1.8 kHz and 5 kHz from constriction at 5.9 cm.

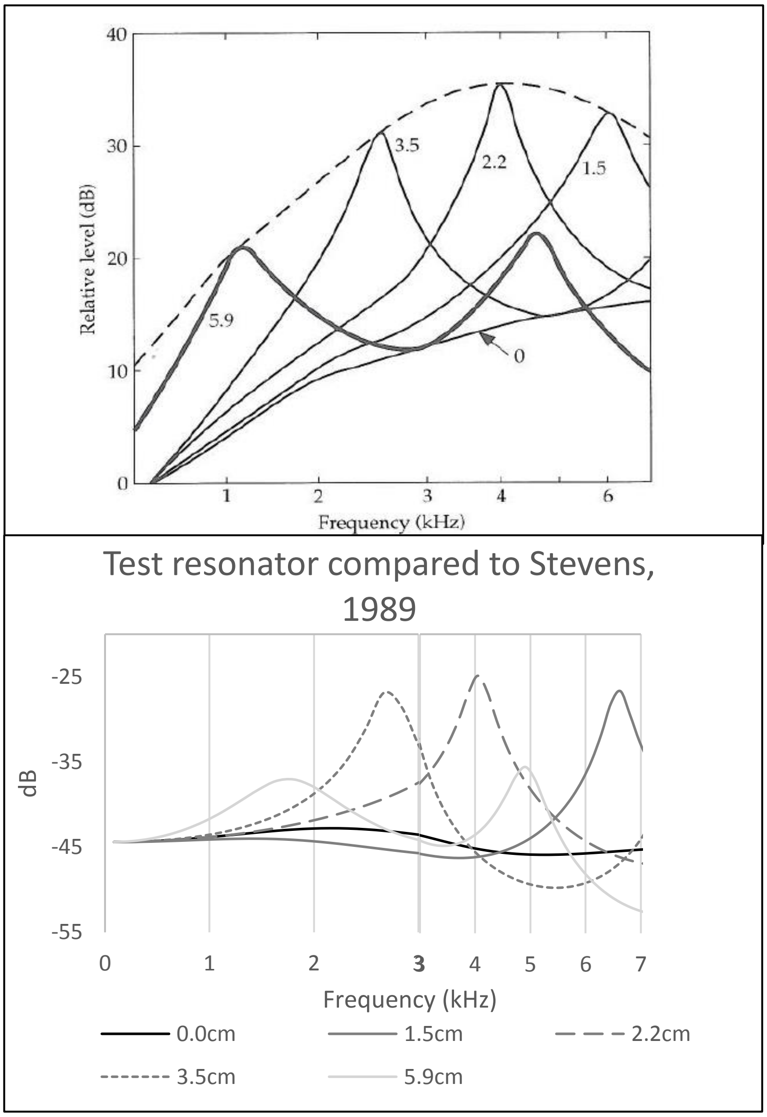


Figure 6.18. Fricative spectra from Stevens, 1989 mentioned in Johnson, 2012 (top) and fricative spectra from the proposed system (bottom). Note that the vertical scale is slightly different between the two graphs.

Another impulse response test from a constriction was set according to real tract shape. Figure 6.19 shows the responses when the tract was set as for /f/, /T/, /s/, /S/ and /h/.

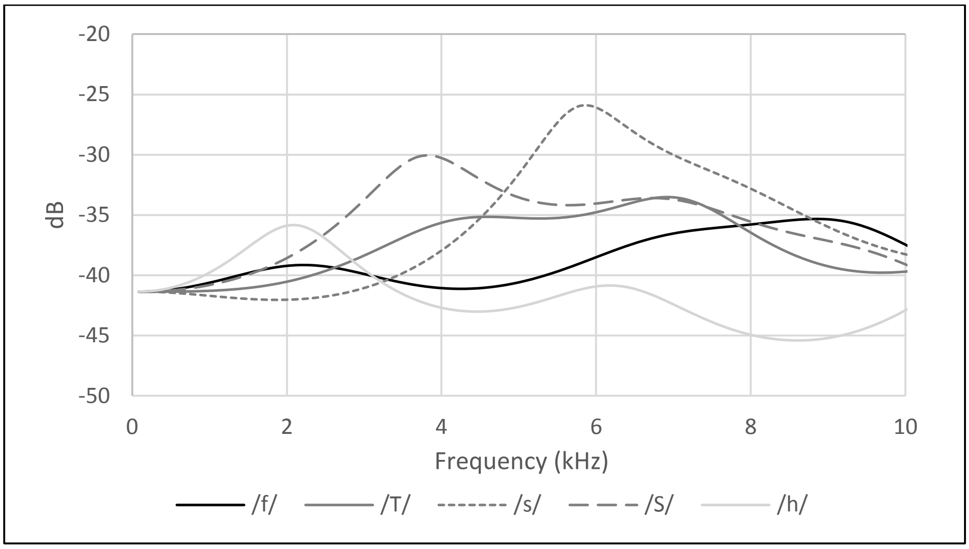


Figure 6.19. Fricative spectra from the proposed system using tract shape from recorded MRI data.

Comparing the results in Figure 6.19 with the theoretical ones in [20], the alveolar /s/ in the figure exhibits a peak of energy at around 5.7 kHz and the palatal /S/ at around 3.7 kHz while Howard mentioned that the alveolar should be around 4 kHz and the palatal at above 2.5 kHz. Moreover, comparison of the results with the real human speech shows similarities of a prominent band of frequencies in /s/, /z/ and /S/, /Z/. Those of /s/ and /z/ have a prominent range of frequency at 4.7-7 kHz, and 1.9-5 kHz in /S/ and /Z/ which is shown in Figure 6.20.

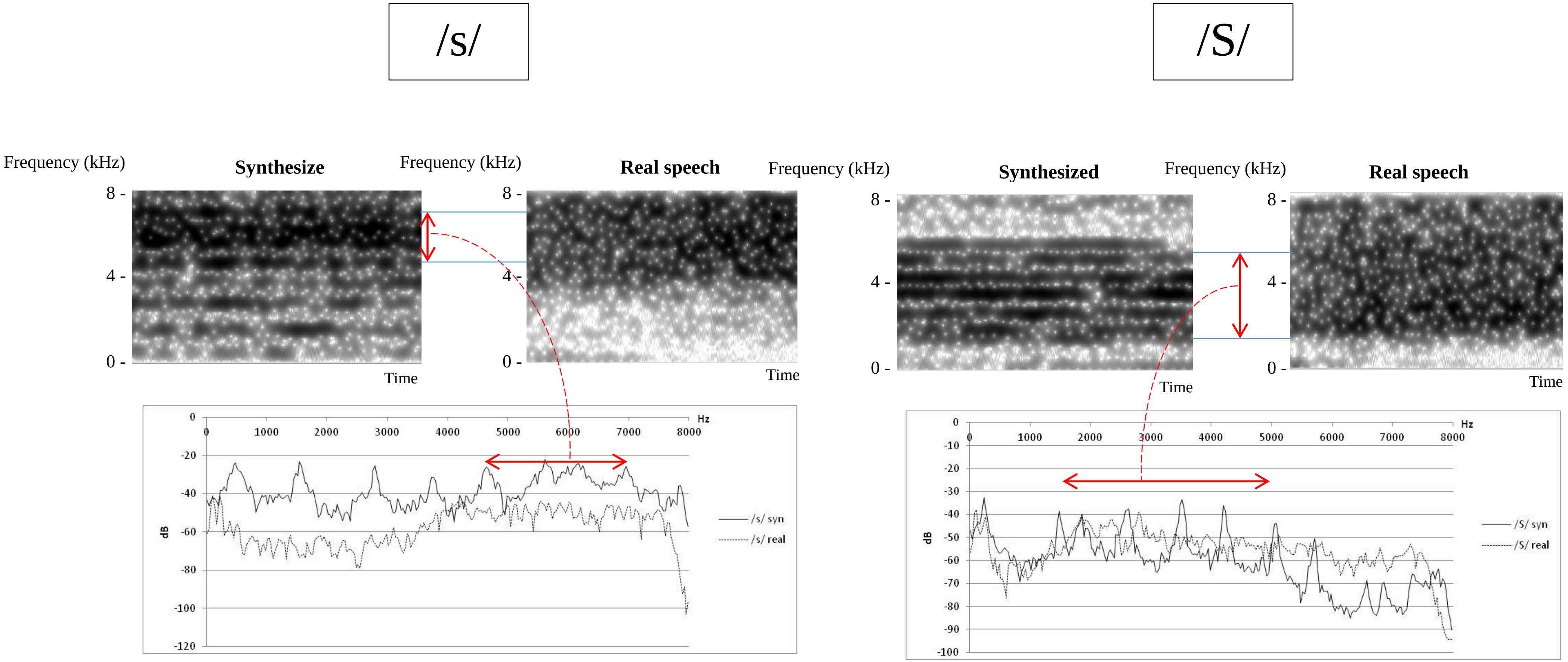


Figure 6.20. Spectrogram comparison of /s/ and /S/ from real and synthesized speech.

6.1.5 Affricates

The affricate is a combination of a stop followed by a fricative, according to Ladeforged in [53]. The plosive part has acoustic characteristics similar to those in /t/ and /d/ [20] but the place of articulation is actually at post-alveolar [55] or further back to palatal [32]. The frication part is shorter than when fricatives are pronounced on their own [20], [53]. Comparing the frication part to that of the fricative pronounced alone, the rise time, which refers to the rising amplitude of the frication part, is clearly perceived [58], [32].

In our experiment, we attempted to control the frication amplitude with the tract shape for /t/ articulated to /A/, according to articulatory trajectories from *mngu0*. The results show similar spectra for the frication parts but not for the plosive parts, as shown in Figure 6.21. This is a result of the limitation of the model, which is unable to simulate pre-voicing before the release stage. The pre-voicing was recorded while the tract was shut completely, but sound was transmitted through the tract walls. Perceptual test results are shown in the next section.

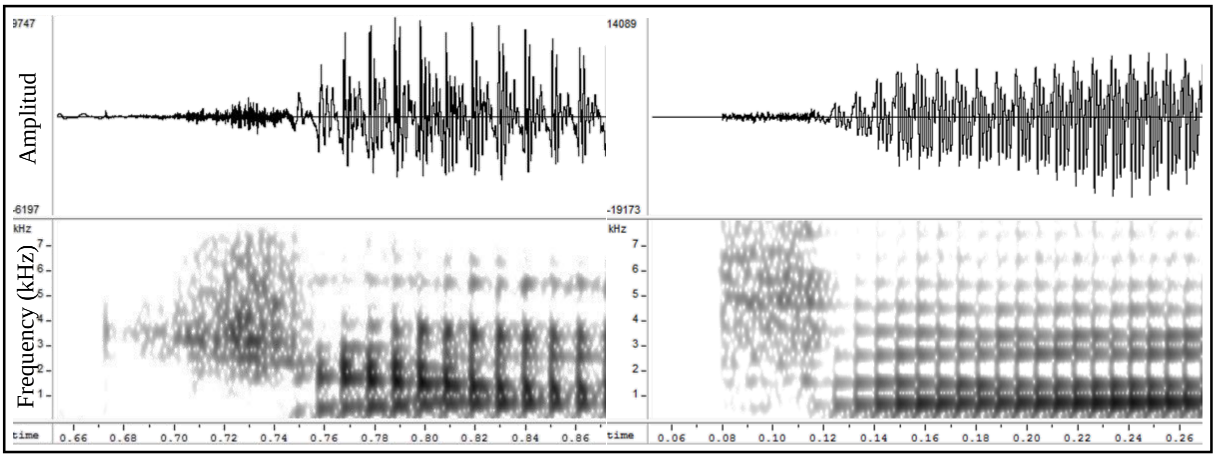


Figure 6.21. Spectrogram of affricate /dZ/ with real speech on the left and synthesized from MRI + *mngu0* on the right.

6.2 Subjective test results

The subjective test was designed to evaluate the accuracy of the output from the system in regard to human perception. The test was designed to:

- evaluate the accuracy of the resonator in synthesizing English consonants;
- evaluate participants' perception of pairs of consonant-vowel (CV) English diphones;
- evaluate the accuracy of applying articulation using discrete time configuration; and
- evaluate the naturalness of the proposed system.

Two experiments were conducted separately: one to evaluate human perception of the synthesized English consonants and the other to evaluate that of synthesized consonant-vowels. In the first experiment participants were asked to score the synthesized consonants by perceiving synthesized sounds one by one and selecting the consonant they thought the sound resembled. All synthesized English consonants were involved, with the exception of /tS/.

This test is for consonants only. It was conducted online. The link and instruction were sent via email as follows.

===========

Instructions

===========

- There will be 21 short sounds you will perceive. Please evaluate them if they sound similar to any English consonant.

- In each page you will see the sound player on the top and a list of consonants under it.

- Please play the sounds as many times as you like then choose the consonant(s) you think it sounds similar to then click submit and then click the Next at the bottom of the page.

The web pages are shown in Appendix 2. All participants were asked to use their personal headphones. All of them could use headphones except one who listened to the sounds from a laptop speaker. The environment was his/her office, hotel room or home. The test begins with filling participant's personal information and then the participants were given 6 short sample sounds without an answer to give them idea how short the test sounds are and then the test starts.

Eight native English acoustic engineers and two forensic phoneticians participated in this experiment. They were five males and five females, aged 30 years on average. The experiment was conducted online. The selected consonant for each sound is used as the score, so 10 is the maximum and 0 the minimum score for the question of what does this sound like. Figure 6.22 shows the scores grouped by manner of articulation.

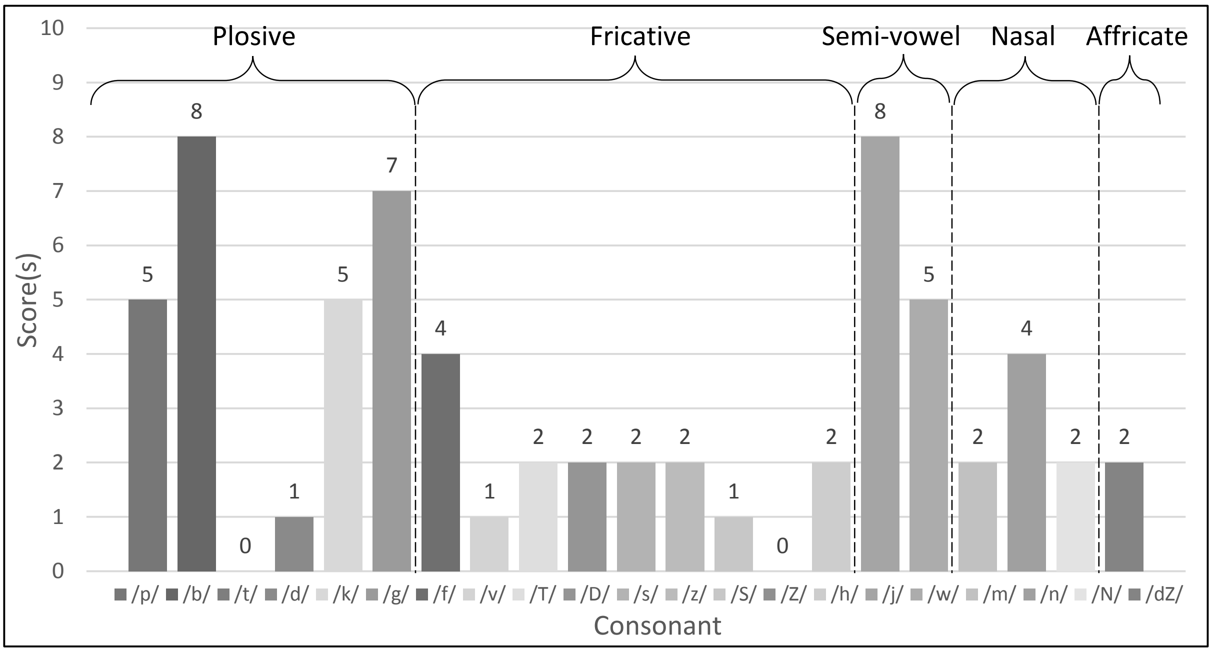


Figure 6.22. Scores of each synthesized consonant, grouped by manner of articulation.

The results show that almost all the synthesized consonants, with the exception of /t/ and /Z/, were perceived. Comparing between manners, semi-vowel obtained the highest average score which is 6.5 out of 10, followed by plosive 4.3, nasal 2.6, affricate 2.0 and fricative 1.7 on average. This indicates that synthesizing English semi-vowels was the most successful while those consonants that involve plosion (plosives and affricates) were better perceived when voiced and less well when voiceless. The voiceless plosives involved white noise injection to carry the friction energy in the frication part of the manner of articulation. White noise is used here and, as mentioned in the previous section, white noise is not good enough to be a frication sound source which relates to the lower scores in voiceless fricatives [70, 74]. Moreover, the low scores for /t/ and /d/ are the result of the adoption of the trajectory from the most extreme position of the tongue tip for alveolar plosives which led to a score of 0 for /t/ and 1 for /d/. In addition, /t/ was mostly marked as /f/ while /d/ was mostly taken as /w/.

White noise is also used for the frication sound source in fricative and affricate synthesis. The use of white noise is believed to be part of the reason for the low scores in all fricatives, except /f/, which has a wide range of frequency (some suggestions for the improvement of the noise sourcing are given in the future work section). Here, in another set of results for nasals, the anti-formant, which is the main acoustic characteristic in nasals, works best for /n/; set to around 3 kHz in this synthesized sound, it was perceived by four participants (the biggest score for nasals) while /m/ and /N/ were picked up by two participants only. These are the results of the complication in having too similar an anti-formant energy in /m/ and too high an anti-formant in /N/ (the anti-formants are shown in Figure 6.4). Even so, within the nasal set, the most participants could perceive was that there were anti-formants in the sounds; they then marked them as nasal but were confused about the place of articulation.

In summary, Table 6.2 shows all perception results grouped by their manner of articulation. The scores are from the responses by subjects. Each column represents each synthesized sound while each row represents scores from the answers. Correct identification is shown alone the X=-Y diagonal of this table, while the other cells in the confusion matrix show non-correct responses or the confusions. The average scores of these results will be shown in percentage and compared to another set of results in Table 6.2 summery of all results of the listening tests.

Table 6.2. Perception results confusion matrix grouped by manner of articulation.

	Plosive						Fricative									Affricate	Nasal			Semi-vowel	
	p	b	t	d	k	g	T	D	f	v	s	z	S	Z	h	dZ	m	n	Ng	w	j
/p/ (as in pet)	5	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
/b/ (as in bat)	3	8	2	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
/t/ (as in time)	2	0	0	0	0	0	1	0	0	0	2	0	0	0	1	1	0	0	0	0	0
/d/ (as in dine)	4	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0
/k/ (as in kind)	3	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/g/ (as in game)	2	0	0	2	3	7	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
/th/ (as in thin)	0	0	1	0	0	0	2	0	0	0	1	1	3	0	0	1	0	0	0	0	0
/dh/ (as in thine)	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	2	0	1	0	0	0
/f/ (as in fog)	0	0	3	0	0	0	2	0	4	1	3	0	1	0	3	0	0	0	0	0	0
/v/ (as in van)	0	0	0	2	0	0	0	0	1	1	0	0	0	1	0	1	2	1	1	3	0
/s/ (as in sea)	0	0	0	0	0	0	1	0	2	0	2	0	2	0	1	0	0	0	0	0	0
/z/ (as in zoo)	0	0	0	0	0	0	1	1	1	0	1	2	2	1	0	1	0	0	0	0	0
/sh/ (as in ship)	0	0	0	0	0	0	0	0	1	0	1	0	4	0	1	0	0	0	0	0	0
/zh/ (as in treasure)	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
/h/ (as in heel)	0	0	0	0	0	0	0	0	2	1	0	0	0	0	2	0	0	0	0	0	0
/dZ/ (as in jibe)	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	2	0	0	0	0	0
/m/ (as in mast)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	0	0
/n/ (as in none)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4	2	0	0
/ng/ (as in ring)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	0	0
/w/ (as in wall)	0	3	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	5	0
/y/ (as in yacht)	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0	0	1	1	0	8
None of them	0	0	1	0	2	0	4	6	3	4	2	7	2	4	4	2	3	3	1	3	2

However, in much phonetic research, the perception of consonants also depends on the context and coarticulation [55, 83]; therefore, another experiment was conducted to examine how participants responded to different coarticulation.

Another perception test for consonant-vowel was conducted under office environment in a laboratory office or quiet zone in a library to convey the real environment of real conversation. Participants perceived the sounds via SONY stereo headphones MDR-XD200 through CX20672 HD Audio Codec audio output supporting sample rates of up to 96 kHz, 16-24-bit resolution [83].

Three experimental sound lists were randomly generated from 70 sounds. Each list contained 18 sounds from (3 nasal + 3 voiced-plosive) x 3 vowels, 14 sounds from (9 fricative + 3 unvoiced-plosive + 2 semi-vowel) x 1 vowel, and 3 individual vowels in 2 sets of synthesized and recorded versions. Fricatives and unvoiced-plosives enable an evaluation of how participants perceive them, even though the noise source is just unfiltered white noise. The recorded sounds were cut from mngu0 WAV files while the synthesized ones adopted coarticulation from it. Participants were then asked to label sounds in SAMPA transcription according to the SAMPA table from [20] or in IPA according to the IPA chart (2005) [82]. The participants were also asked to award naturalness scores (0 for totally robotic to 10 for natural-like human speech) to each sound.

In this test participants were asked to label the sounds they perceived after listening to a pair of synthesized and recorded diphones. The phone pairs in this experiment were consonant-vowel (CV) for testing the resonance of the articulation of the proposed model. An adjacent vowel works as a carrier which makes the overall sound more like real syllables in natural speech. The sounds are pairs of all studied English phones with the three vowels /u/, /i/ and /A/. These three vowels are located at corners of the cardinal vowel chart, which is an ideal oral articulation chart for all possible vowels in the world’s languages, and represents extremes of articulation in two perspectives – front-back and open-close, for which /u/ is front and narrow, /i/ back and narrow and /A/ back and open. The consonant set contains nasals /m, n, N/, plosives /p, b, t, d, k, g/, fricatives /f, z, T, D, s, z, S, Z, h/, semi-vowels /j, w/ but not /r, l/ or /tS, dZ/ as described in the previous section.

Twelve subjects participated in this experiment; some of whom had also done the previous experiment but some of whom had not. They were eight males and four females, aged between 23 and 34 years (mean: 27.75 years). Four had a strong phonetic background and three had taken courses in phonetics, while the rest had no phonetic experience but were British with a musical or audio engineering background.

The first set of analysed results shows the subjects' perception of vowels in CV chunks taken from *mngu0*. In a comparison of the subjects' perception to phone label provided in the *mngu0* corpus, it shows that 75% of the participants perceived the recorded vowel /A/ and labelled it as /A/, the same as it is labelled in the .lab file, while 100% perceived /i/ and put it as /i/, and 50% perceived /u/ and put it as /u/. It shows that the chunks of vowel used in this research can be perceived confusingly when they are presented alone. However, their trajectories together with their leading part (from studying consonants) are used in this research. In addition, Figure 6.23 shows the percentage of participants who labelled chunks of sound cut from *mngu0* as they were labelled in the database.

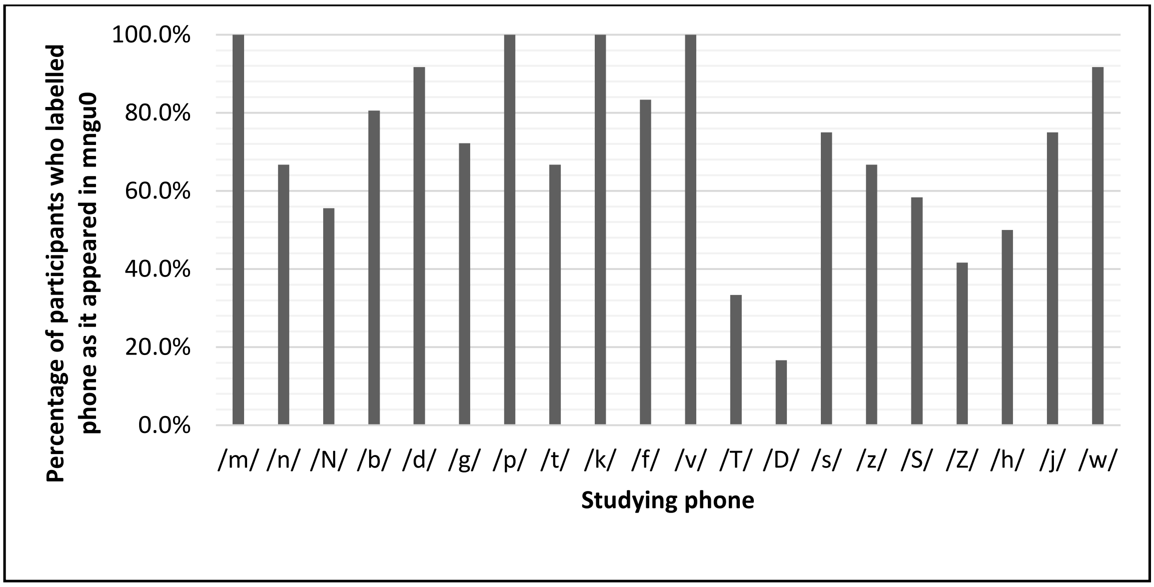


Figure 6.23. Number of participants who labelled phones from recorded speech correctly.

This chart shows the ability of participants to perceive each recorded English phone from hearing chunks of recorded speech.

Despite the fact that consonants and vowels in the *mngu0* corpus were not perceived 100% correctly according to their labels, their recorded articulatory trajectories were used as articulatory guidelines in this research. As described in the previous chapter, vocal tract shape and articulatory trajectories are mixed from two sources, vocal tract shape from [60] and articulatory trajectories from *mngu0*. All the studied phones, as well as /A/, /i/ and /u/ MRI data from the subject named Jack in [60], are used. In Speed's work, he claimed that, using them in his 3-D DWM vowel synthesizer to reproduce the vowels, his subjective accuracy was that 20 out of 20 participants perceived and labelled synthesized /A/ correctly as /A/, while 14 of them labelled /i/ as /i/ and 10 labelled /u/ as /u/. In this experiment, the same vocal tract shape extracted from the same MRI data is used to resynthesize the phones, and 4 out of 12 participants perceived /A/ as /A/, while 7 out of 12 perceived /i/ as /i/ and 6 out of 12 perceived /u/ as/ u/, using 2-D DWM.

Moving from the analysis of the participants' perception of vowels to that of consonants, the synthesized consonants were put in this listening test with a carrier, an adjacent vowel, in CV format. /A/, /i/ and /u/ are used in nasals and voiced-plosive synthesis, and only /A/ is used for unvoiced plosives, fricatives and semi-vowels. Even though many phonetics researchers have mentioned that using the carrier can strongly affect the perception of consonants, the consonants are not usually pronounced alone in real speech. Therefore, the vowel carrier under the implementation of articulatory simulation is involved in this perception test. Figure 6.23 shows the first set of percentages of participants who perceived and labelled synthesized nasal phones as they were produced by nasal models for /m/, /n/ and /N/.

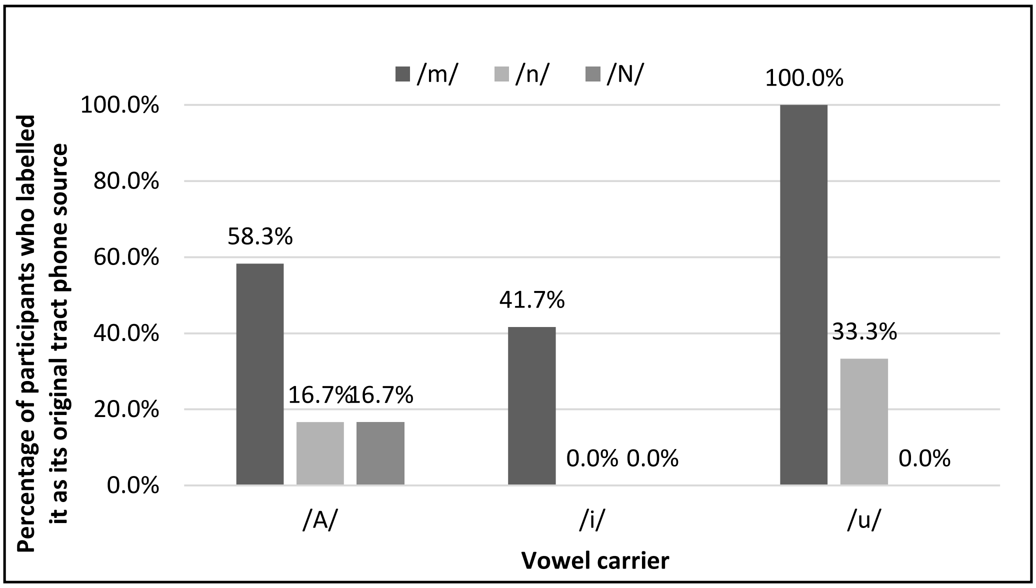


Figure 6.24. Percentage of participants who were able to label synthesized nasal in different vowel carrier.

The results in the figure show that /m/ was perceived when it was produced with the vowels /A/, /i/ and /u/, but /n/ was perceived only when it was synthesized with the back vowels /A/ and /u/ while /N/ was perceived only when it was produced with the open-back vowel /A/. With these different vowel carriers, these results indicate the accuracy of the proposed model in resonating formant transition when the nasal tract is attached. Kent and Read [32] describe how the transitions in nasals also affect the listener's perception and have similar characteristics as those in plosives [32]. The results then not only show how participants perceive nasal consonants themselves, but also how well the nasal characteristics work with adjacent vowel carriers. The place of articulation is then discussed here. From the results, participants perceived /m/ from the model with bilabial closure correctly when it was followed by the close-back vowel /u/; the next most accurate was for the open-back /A/ at 58.2% while the least accurate was for the close-front /i/ at 41.7%. Then the alveolar for /n/ and velar for /N/ work only for back vowels. These indicate that the anti-formants in synthesized /n/ and /N/ (shown in Figure 6.4) cannot be perceived and/or F2 transition to close vowels /i/ and /u/ are not realistic enough to be perceived. These results also add up more accuracy to /m/ and /N/ which have low scores in the previous experiment; this indicates that, when they are synthesized with the proper coarticulation, they can be perceived more accurately.

Without nasalization, from slowly releasing the blockage to suddenly releasing the phone type plosive, Table 6.3 shows how accurate the model is when it is used in producing English plosives with an adjacent vowel carrier. The table shows the percentage of participants who perceived them with loci and formant transition for voiced plosives /b/, /d/ and /g/ and unvoiced plosives /p/, /t/ and /k/, and then labelled them correctly. Voiced plosives were produced with /A/, /i/ and /u/, but unvoiced were produced only with /A/. This was because the unvoiced was not the focus of this perception test as the noise source is not focused on this research. Instead resonance and then white noise is used for the plosion excitation and is definitely not sufficiently accurate for a perception test. However, as testing how participants perceive them is still of interest, /A/ is used as a carrier for unvoiced plosives.

Table 6.3. Percentage of participants who perceived each plosive grouped by carrier vowel.

	/b/	/d/	/g/	/p/	/t/	/k/
/A/	0.0%	8.3%	91.7%	8.3%	33.3%	50.0%
/i/	58.3%	0.0%	83.3%	-	-	-
/u/	75.0%	0.0%	75.0%	-	-	-

From the overall view of the results in this table, we can see percentages of perception in all plosives which means that all synthesized plosives are perceivable but differently so, depending on the vowel carrier. The table shows that /g/ had the highest number of participants who could perceive it accurately, giving the best response when it is synthesized with /A/ then /i/ then /u/, while /b/ was identified best when it was produced with close vowels /i/ and /u/ only (not /A/), and /d/ was best identified when it was produced with /A/ only.

In table 6.3, there are three sounds that are not correctly perceived at all (0.0% shown in the table) which are /bA/, /di/ and /du/. If we take a look at them a bit more closely in their formant transitions in Figure 6.17, their F2 moves in different direction comparing to those in real speech even if their F1 move in the same direction and almost in same frequencies. These support Delattre et al.'s theory on plosive perception that all of the F1, F2 transitions play an important part in that if there is anything in the sound that tends to cue another sound then there will be confusion. Moreover, there are some differences in F3 transitions but they don't have much effect on perception comparing to F2, to the different F3 transitions in /gA/ in Figure 6.17 that got 91.7% correct perception in table 6.3.

/bA/, /di/ and /du/ were not perceived. The confusion matrix is shown in Table 6.4. In detail, four participants mistook /b/ when it was produced in /bA/ for /d/, while one participant mistook /d/ in /di/ for /D/, and three participants mistook /d/ in /du/ for /g/.

Table 6.4. Confusion matrix of number of participants marking the studied phone /b/ in /bA/, /d/ in /di/ and /d/ in /du/ as a different one.

	/p/	/b/	/t/	/d/	/k/	/g/	/dh/	Not perceived any consonant
/b/ in /bA/	-	-	-	1	-	-	1	10
/d/ in /di/	-	-	-	-	-	-	1	11
/d/ in /du/	1	1	-	-	1	2	-	7

These indicate that there are some errors in /bA/, /di/ and /du/. When we examined their formant transitions more closely, we found that there were F2 locus errors in them. The locus of /bA/ was slightly too high (higher than F2 frequency) which does not match to its theoretical locus that should be lower than F2 frequency. Moreover, it is more similar to the transition in /d/ (the theoretical locus is shown in Figure 6.16), and therefore it was perceived as /d/. In the same way, the F2 locus in the synthesized /du/ was not high enough to be perceived as /d/, but was at a low frequency as for /g/ and therefore was perceived as /g/. In /di/, the locus was correct but the burst was too thin which caused an overly slow opening of the oral closure. These thin energy and wrong locus frequencies may indicate the disadvantage of adopting some trajectories from continuous speech, in that they can mislead articulatory trajectories for CV synthesis which directly affect the acoustic characteristics and therefore lead to the wrong perceived sound.

For fricatives, only the vowel /A/ was used as a carrier and only /f/, /v/, /T/, /S/ and /h/ were perceived. Almost all of the unvoiced fricatives were perceived by different numbers of participants, with the exception of /s/. /h/ and /f/ were picked up the most because of their acoustic properties that cover a wide range of frequencies. All the other fricative properties were described by resonance in the frontal cavity from constriction [58], [84], [70], [85]. /T/ and /S/ were perceived by a few participants which means that some participants can perceive low frequency resonance from the frontal cavities in synthesizing /T/ and /S/, but not /s/ which involved the shortest front cavities and resonated friction noise in high frequencies.

/v/ was the only voiced fricative perceived. This means that the voicing source with coarticulation implementation could not synthesize a proper voiced fricative acoustic property accurately when one of them (/v/) that involves a wide range of frequencies from labiodental constriction was perceived accurately but all others that resonate a narrower range of frication frequencies were not picked up by listeners (/D/, /z/, /Z/ were not labelled). However, comparing with the results from the previous experiment, synthesizing them alone, without a carrier, can be perceived by some participants.

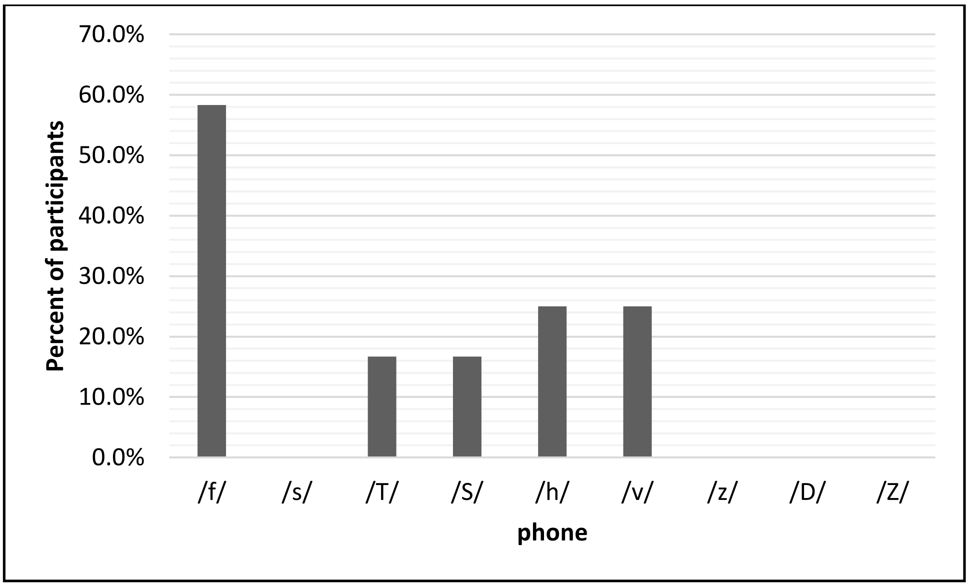


Figure 6.25. Perceivable synthetic fricative chart.

The last set of results is the number of participants who can perceive /j-A/ and /w-A/. Ten of the twelve picked up /j/, while five of them perceived /w/. This result shows that faster articulation for synthesizing a semi-vowel with an adjacent open vowel /A/ affects human perception slightly differently. Oral tract articulation from the shape for front-close vowel /i/ to back-open /A/ was more successful than that from back-close vowel /u/ to /A/ for 41.6% of participants.

In addition, all participants were asked to score the naturalness they thought they perceived from 0 to 10, when 0 means totally robotic and 10 means real speech. The CV cut chunks from 16 kHz recorded speech obtained 5.81 out of 10.0 while the synthesized CVs which were also down sampled to 16 kHz obtained 4.49 out of 10.0, on average. The low scores of those from real speech (the cut chunks) is a result of too short a length of CV with unclear boundary cut from continuous speech. The below middle score from synthetic speech is a result of robot-like characteristics. These results indicated that another set of studying sounds and their articulatory trajectories (pairs of CV) is needed for more precise comparison in the perception test. However, the lower score of the synthetic ones compared to the real ones indicates that the sounds of the proposed system are more robot-like than natural.

Figure 6.26 shows the collative of the results of the listening tests. C, CV and real show the average percentage of accuracy in each test. C is for the results from consonant perception test, CV is for consonant-vowel and real is for real speech. The figure shows that the semivowel gets the highest accuracy in total than the plosive, nasal, fricative and affricate, respectively. The black error bars indicate standard deviation in each case. This shows the success of semivowel recognition in both C and CV tests in getting not much different percentage of accuracy as same as nasal (but in lower percentage). On the other hand, plosive and fricative get bigger different percentage. This could be an effect of using different source of tract shape and trajectories and also am extra noise source in vocal tract, which is the main acoustic characteristic of both plosive and fricative.

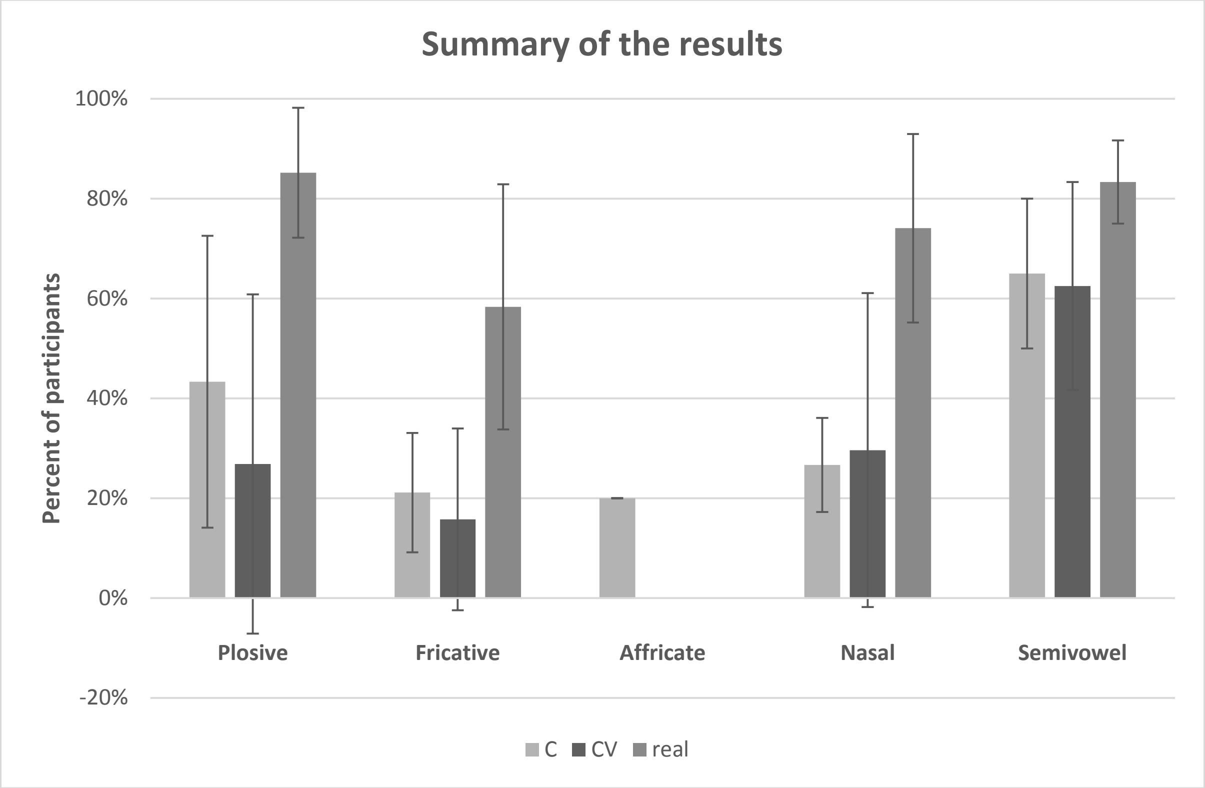


Figure 6.26. Summary of the results of the listening tests.

Chapter 7
Conclusions and Future Work

A two-dimensional digital waveguide mesh has been used as a numerical method to simulate wave propagation for English consonant production. It was used with specific configurations for acoustic consonant production. According to [55], consonants are produced by articulations in the vocal tract, and this research has demonstrated 2-D DWM acoustic performance when the cross-sectional areas of the vocal tract are changed. The manner of articulation enabled the grouping together of English phones with similar articulation. The acoustic behaviours for each manner became the main target in this study. Generally, 2-D DWM has been used with some additional configurations such as:

- attaching a side branch for nasal studies;

- injecting a secondary noise source at the place of constriction for fricative and plosive studies;

- implementing the articulatory trajectories to study semi-vowels and all other consonants in the case of CV synthesis.

The simulation itself has adopted the cosine impedance function from [5] to smooth the formant transition during articulations. The articulation in this research was implemented by a vocal tract area change sequence based on recorded articulator position changes.

7.1 Conclusion

In semi-vowel synthesis, the model was set as for synthesizing diphthongs but the transition was faster and the results show that the modelled formant transitions are similar to those found in natural speech and also that more than two thirds of participants can identify them correctly. However, the perception test also shows that responses from some English-speaking participants identify the synthesized semi-vowels, plosives, nasals and some fricatives more reliably when they are synthesized with an adjacent /A/ and /i/ but less reliably when with an adjacent /u/.

For fricatives, the synthesis was set up with an additional noise source (white noise) around the place of constriction, at junctions behind the place of constriction. The objective results showed similarities in terms of frequency range compared to those in natural speech in all attempts. However, the use of white noise did not create a natural frication source. Objectively, the results show that 2-D DWM has resonances in its fricative outputs which can be associated to the size of the front cavity or tube, but the subjective results indicate that these properties are not clear enough to be perceived reliably by English speakers. The fricatives /f/, /T/, /S/, /h/ and /v/ were perceived most reliably.

For plosives, the main acoustic features, stop gap, transient, frication interval, VOT and formant transition, are summarised in Chapter 4. The stop gap and transient burst relationship was examined. The result shows that the longer closure causes the stronger burst, but that nevertheless aspiration was not generated. Hence, an additional noise source (again, white noise) is injected at the place of constriction/closure to hold the frication energy for the aspiration interval. VOT was not examined in detail but voice source was turned on for voiced plosive after an appropriate voice onset time but turned off for unvoiced synthesis. The formant transition was examined by synthesizing voiced plosives with three vowels /i/, /u/ and /A/ to vary target vowel formant frequencies. The results show that all transitions have similar trajectories to those appear in natural speech apart from the results from /u/. It is suggested that synthesizing /u/ requires proper modelling of lip rounding which is outside the scope of this research.

For affricates, the acoustic properties are similar to plosive and fricative together (see Chapter 4). The articulation plays a more important part here, as does the nature of the frication source. The lack of MRI images for the affricates was a limitation (the place of articulation has to be post-alveolar closure and then post-alveolar constriction [55] but we only have alveolar plosives /t/ and /d/). However, the simulation was done using the alveolar place of articulation and this gives comparable acoustic properties which have been discussed. Therefore, for affricates, more accurate vocal tract shape and frication sound source are needed in future work.

For nasals, the branch tube is attached by a modification of junction topology at the velopharyngeal port location (shown in Figure 5.11). The results show that the branch tube absorbs acoustic energy at its anti-formant frequency in spectrograms not far from the expectation frequency.

The results in Figure 6.6 show convincing anti-formants when applying real tract shapes /m/, /n/ and /N/. The subjective results showed that most participants could perceive /m/, /n/ and /N/ when they were synthesized with adjacent /A/ and all participants could perceive /m/ when it was with /u/.

In addition, given the additional noise source injection location along the x-axis, on the y-axis the noise source is injected equally into all y junctions which means that the noise source is scattered at the place of injection. A more precise dipole source as described in Shadle, 1991 and Peter, 2014 could also be implemented and injected at boundary junctions only, as future work.

From the view of acoustic phonetics, the muscle gesture at the articulators has not been discussed or included in this research (readers can find an example of muscle control for great success in articulatory speech synthesis in *Functional Phonology* [39]). Only cross-sectional areas were extracted and used to guide the tract shape for each English phone. Articulatory trajectories were adopted from *mngu0* [67] and implemented in the simulations mainly to enable formant transformations. This can also help to guide the speed of the release stage in plosion generation.

7.2 Future work

There are several points that can be considered for continuing this work in the future, such as:

- Increasing the number of dimensions to three

Speed successfully evaluated his 3-D DWM vowel synthesizer in 2013 [60]. His work successfully generated similar sounds to recorded ones from the same subject who gave his/her tract shape scanned at a sampling rate of 176.8 kHz. There are possibilities of applying his 3-D synthesizer for consonants. In addition, to get more human-like sounds the system could also include a frication source (e.g. Birkholz [74]).

- A more accurate frication source

An appropriate frication noise source is essential for fricative, plosive and affricate synthesis. The noise characteristics vary by place of articulation and the nature of the source (obstacle or wall) [80]. However, Peter, 2014 [74] claims that the place of articulation cannot always rely on the area function during dynamic articulatory speech synthesis. Further studies in friction source implementation can be found in (Shadle, 1991, Badin, 1995, Peter, 2014).

- Using a more accurate vocal tract shape and articulation from the same subject

In this work, I applied vocal tract shapes that were scanned under support from York Neuroimaging Centre (YNiC) using a General Electric 3.0T HDx Excite MRI Scanner to the articulatory trajectories from another subject which was recorded at Ludwig-Maximilians-Universität München using Cartsens AG500 electromagnetic articulograph. The two subjects are both native English speakers but more accurate tract shape from the same subject is strongly needed. Then we can use trajectories to indicate coarticulation in the vocal tract properly. For the target sound recording, they can be done from the same subject in an MRI scanner just before the start to reduce background noise, then the vocal tract shape can be scanned and then the articulatory trajectories can be recorded separately.

As MRI technology is improving, real-time MRI (rtMRI) is also very helpful in guiding articulatory trajectories in the vocal tract. Kim [86] claims that he can evaluate the distance between tissue boundaries in the USC-EMO-MRI corpus more precisely; hence, we can use this set of data to control articulation as well. Note that the quickest sampling rate in recording rtMRI is 400 Hz using the NDI Wave Speech Research System which is quick enough even for capturing plosive gestures (at least 200 Hz for plosive [55, 32]).

In addition, in the case of plosive simulation, the tract is meant to be completely shut which requires the cosine function to be weighted to near zero value (it cannot be zero to avoid dividing by zero in the calculation). Then the release acts quickly after the closed phase which causes a sudden increase in the weighted cosine parameter. However, the allowance of the leakage of propagated pressure is spread over the tract width as cosine shape with the most pressure at the middle of the tract.

Appendix 1

```
////////////////////////////////////////////////////////////////////
// Scatter algorithm
// Anocha June 2011: revise for velopharyngeal 5 port scattering
////////////////////////////////////////////////////////////////////

void CVocalTract2DWaveScat::Scatter()
{
        UInt16 x, y;

        for(x=0; x<m_iXSizeMax; x++)
        {
                for(y=0; y<m_iYSizeMax; y++)
                {

                        if(x==m_iVelumIndex)                    // Anocha 5 port scattering
                        {
                                m_Pressure = 2*( (  m_PNPlus[x][y] / m_ZNorth[x][y]
                                                                + m_PEPlus[x][y] / m_ZEast[x][y]
                                                                + m_PSPlus[x][y] / m_ZSouth[x][y]
                                                                + m_PWPlus[x][y] / m_ZWest[x][y]
                                                                + m_PVelarInput[y] / m_ZVelar[y])

                                                        / (      1/m_ZNorth[x][y] + 1/m_ZEast[x][y]
                                                                + 1/m_ZSouth[x][y] + 1/m_ZWest[x][y]
                                                                + 1/m_ZVelar[y]) );
        if(fabs(m_Pressure)>1.0)
                m_Pressure=m_Pressure;



}

else                                                            // 4 port scattering
{
        m_Pressure = 2*( (  m_PNPlus[x][y] / m_ZNorth[x][y]
                                                + m_PEPlus[x][y] / m_ZEast[x][y]
                                                + m_PSPlus[x][y] / m_ZSouth[x][y]
                                                + m_PWPlus[x][y] / m_ZWest[x][y]  )

                                        / (             1/m_ZNorth[x][y] + 1/m_ZEast[x][y]
                                               + 1/m_ZSouth[x][y] + 1/m_ZWest[x][y] ) );
}

m_PNMinus[x][y] = m_Pressure - m_PNPlus[x][y];
m_PEMinus[x][y] = m_Pressure - m_PEPlus[x][y];
m_PSMinus[x][y] = m_Pressure - m_PSPlus[x][y];
m_PWMinus[x][y] = m_Pressure - m_PWPlus[x][y];

// nasal output to tract
if(x==m_iVelumIndex)
```

```
                        m_NasalTract->setInputFromTract( (m_Pressure-m_PVelarInput[y]), y);
            }
        }

        if(m_bNasal)        m_NasalTract->Scatter();

}
```

```
////////////////////////////////////////////////////////////////
// Timestep algorithm
// Anocha : revise for extra noise source injection and lip radiation
////////////////////////////////////////////////////////////////

void CVocalTract2DWaveScat::Timestep()
{
    UInt16 x, y;
    Float32 input;
    Float32 noise;
    UInt16 fricJunc = m_vCurrentVowel->m_iImpMaxIdx ;



    noise = m_SystemData->amp_coef * m_SystemData->extraNoise_coef *getNoiseExternalInputSample();
    input = m_SystemData->amp_coef * getInputSample();



    for(x=0; x<m_iXSizeMax; x++)
    {
        for(y=0; y<m_iYSizeMax; y++)
        {

            if(x==0)
                m_PWPlus[x][y] = m_fGlottalRef * m_PWMinus[x][y] + input;
            else{
```

```
        ////  Adding extra noise source at the most constricted junction

        if(m_SystemData->extraNoise_coef != 0.0 && x== (fricJunc+fricJuncCnt))

                m_PWPlus[x][y] = m_PEMinus[x-1][y] + (m_SystemData->extraNoise_coef * noise);

        else

                m_PWPlus[x][y] = m_PEMinus[x-1][y];

}



if(x==m_iXSizeMax-1)

        m_PEPlus[x][y] = m_fAirRef * m_PEMinus[x][y];

else

        m_PEPlus[x][y] = m_PWMinus[x+1][y];




if(y==0)

        m_PSPlus[x][y] = (x > m_iLipIdx)?m_fAirRef:m_fWallRef * m_PSMinus[x][y];

else

        m_PSPlus[x][y] = m_PNMinus[x][y-1];


if(y==m_iYSizeMax-1)

        m_PNPlus[x][y] = (x > m_iLipIdx)?m_fAirRef:m_fWallRef * m_PNMinus[x][y];

else

        m_PNPlus[x][y] = m_PSMinus[x][y+1];
```

```
				// velar inputs
				if(m_bNasal && x==m_iVelumIndex)
				{
					m_PVelarInput[y] = m_NasalTract->getOutputToTract(y);
				}

			}
		}

		if(m_bNasal)		m_NasalTract->Timestep();
}
```

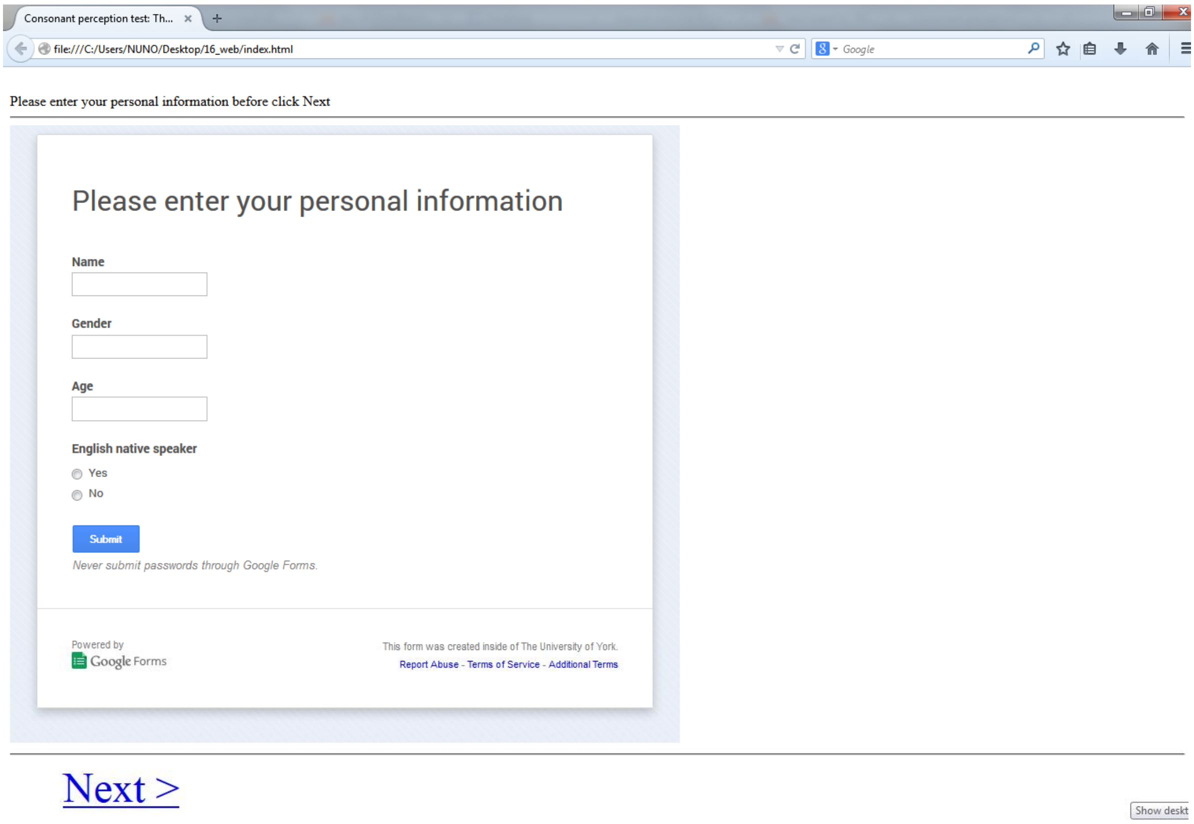
```
////////////////////////////////////////////////////////////
// Main sample increment function
// Anocha 10 May 2012 : revise for separate oral and nasal output channel and articulation
////////////////////////////////////////////////////////////

Float32 CVocalTract2DWaveScat::getSample()
{

        Scatter();
        Timestep();


        // Anocha : add m_bOral m_bNasal
        if(m_bOral)
                setPressureOut(0,(Float64)(getOutput()));
        else
                setPressureOut(0,0.0);
        if(m_bNasal)
                setPressureOut(1, m_NasalTract->getOutput());
        else
                setPressureOut(1,0.0);


        if(m_bSliding)                          Slide();
        if(m_bManualSliding)            ManualSlide();
```

```
    if(m_bVelumSliding)         VelumSlide();

    // Anocha nasal and oral output channel separatoin
    Float64 pressure=0.0f;
    for(int i=0; i<2; i++)
    {
/*          //Anocha: comment out for overclip protection
        if(getPressureOut(i)>0.5f)
            setPressureOut(i, 0.5f);

        if(getPressureOut(i)<-0.5f)
            setPressureOut(i, -0.5f);
*/
        if(m_SystemData->getCh()==1)
            pressure += getPressureOut(i);
        else
            pressure = getPressureOut(i);

        if(m_SystemData->getWriteWav())
        {
            if(m_SystemData->getCh()==1 && i==0) continue;
            m_DataOut->DumpData(&(pressure), 1);
            pressure = 0;
        }
    }
```

```
	if((	m_SystemData->samplecnt == smpStep * m_SystemData->artcnt)	//Anocha smpStep = m_SystemData->getFs/m_artFs; //Anocha for mngu0, m_ArtFs is 200
			+ m_iTimeHold + m_iTimeShift
		)
	{
			Articulate(artShp[m_SystemData->artcnt], 1.0/m_artFs);

			if ((double)m_SystemData->samplecnt/(double)m_SystemData->getFs == fricStartTime)
				m_SystemData->extraNoise_coef = m_fExtraNoiseAmp;

			if ((double)m_SystemData->samplecnt/(double)m_SystemData->getFs >= fricStopTime)
				m_SystemData->extraNoise_coef = 0.0f;

	}

	m_SystemData->samplecnt ++;



	if(m_bMute)			return 0;
	else				return 1;//pressureOut;// Anocha change for 2 channel array


}
```

```
//////////////////////////////////////////////////////////////////
// Articulation
//////////////////////////////////////////////////////////////////
void CVocalTract2DWaveScat::Articulate(VowelType tractShape, Float32 time){ // time is 0.005 s for articulation from mngu0


		m_SystemData->m_Dlg->m_VocalSystem->setVowelSlide(tractShape, time);
}
```

Appendix 2


Consonant perception test: Th...
file:///C:/Users/NUNO/Desktop/16_web/index.html
Google
Show deskt

Please enter your personal information before click Next

Please enter your personal information

Name

Gender

Age

English native speaker

○ Yes

○ No

Submit

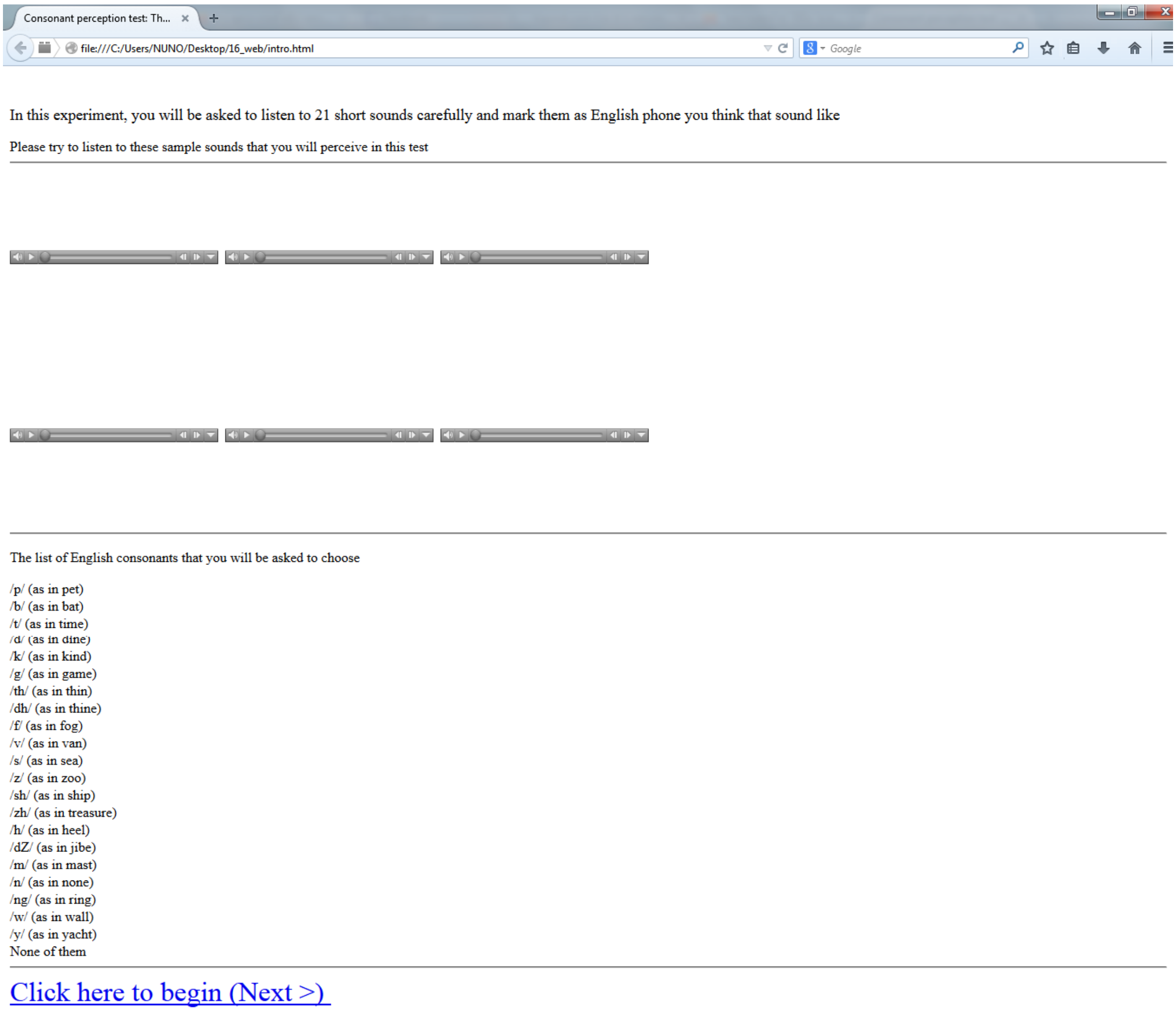
Never submit passwords through Google Forms.

Powered by Google Forms

This form was created inside of The University of York.

Report Abuse - Terms of Service - Additional Terms

Next >


Consonant perception test: Th...
file:///C:/Users/NUNO/Desktop/16_web/intro.html
Google

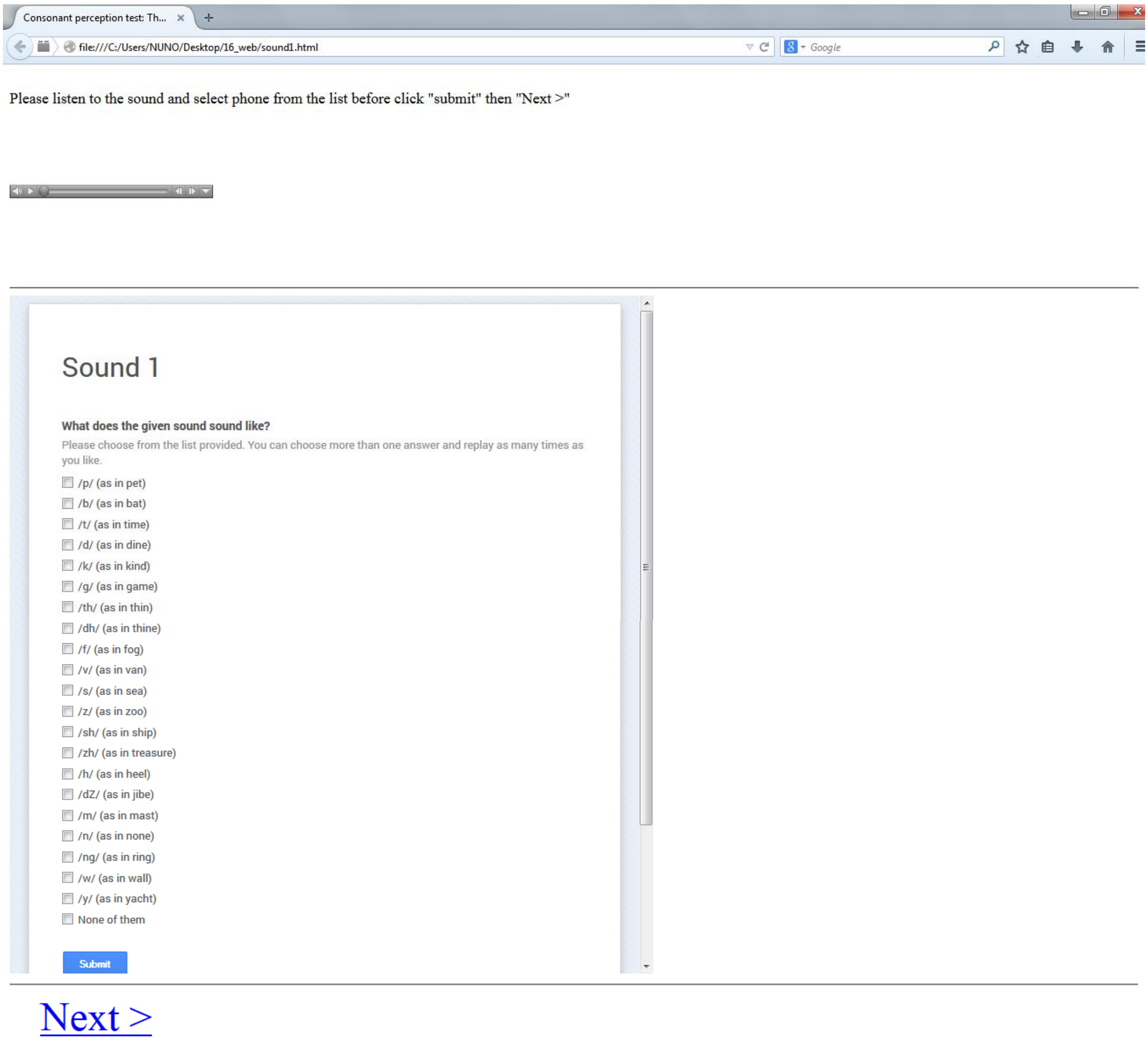
In this experiment, you will be asked to listen to 21 short sounds carefully and mark them as English phone you think that sound like

Please try to listen to these sample sounds that you will perceive in this test

The list of English consonants that you will be asked to choose

/p/ (as in pet)
/b/ (as in bat)
/t/ (as in time)
/d/ (as in dine)
/k/ (as in kind)
/g/ (as in game)
/th/ (as in thin)
/dh/ (as in thine)
/f/ (as in fog)
/v/ (as in van)
/s/ (as in sea)
/z/ (as in zoo)
/sh/ (as in ship)
/zh/ (as in treasure)
/h/ (as in heel)
/dZ/ (as in jibe)
/m/ (as in mast)
/n/ (as in none)
/ng/ (as in ring)
/w/ (as in wall)
/y/ (as in yacht)
None of them

Click here to begin (Next >)


Consonant perception test: Th...
file:///C:/Users/NUNO/Desktop/16_web/sound1.html
Google

Please listen to the sound and select phone from the list before click "submit" then "Next >"

Sound 1

What does the given sound sound like?

Please choose from the list provided. You can choose more than one answer and replay as many times as you like.

- □ /p/ (as in pet)
- □ /b/ (as in bat)
- □ /t/ (as in time)
- □ /d/ (as in dine)
- □ /k/ (as in kind)
- □ /g/ (as in game)
- □ /th/ (as in thin)
- □ /dh/ (as in thine)
- □ /f/ (as in fog)
- □ /v/ (as in van)
- □ /s/ (as in sea)
- □ /z/ (as in zoo)
- □ /sh/ (as in ship)
- □ /zh/ (as in treasure)
- □ /h/ (as in heel)
- □ /dZ/ (as in jibe)
- □ /m/ (as in mast)
- □ /n/ (as in none)
- □ /ng/ (as in ring)
- □ /w/ (as in wall)
- □ /y/ (as in yacht)
- □ None of them

Next >

Appendix 3

Synthesized English consonant sound samples using 2-D Digital Waveguide Mesh.

	Synthesized sound	Cut chunk from continuous speech
/A/	track1_1.wav	track1_2.wav
/i/	track2_1.wav	track2_2.wav
/u/	track3_1.wav	track3_2.wav
/bA/	track4_1.wav	track4_2.wav
/bi/	track5_1.wav	track5_2.wav
/bu/	track6_1.wav	track6_2.wav
/dA/	track7_1.wav	track7_2.wav
/di/	track8_1.wav	track8_2.wav
/du/	track9_1.wav	track9_2.wav
/gA/	track10_1.wav	track10_2.wav
/gi/	track11_1.wav	track11_2.wav
/gu/	track12_1.wav	track12_2.wav
/pA/	track13_1.wav	track13_2.wav
/tA/	track14_1.wav	track14_2.wav
/kA/	track15_1.wav	track15_2.wav
/fA/	track16_1.wav	track16_2.wav
/sA/	track17_1.wav	track17_2.wav
/TA/	track18_1.wav	track18_2.wav
/SA/	track19_1.wav	track19_2.wav
/vA/	track20_1.wav	track20_2.wav
/zA/	track21_1.wav	track21_2.wav
/DA/	track22_1.wav	track22_2.wav

	Synthesized sound	Cut chunk from continuous speech
/ZA/	track23_1.wav	track23_2.wav
/hA/	track24_1.wav	track24_2.wav
/mA/	track25_1.wav	track25_2.wav
/mi/	track26_1.wav	track26_2.wav
/mu/	track27_1.wav	track27_2.wav
/ni/	track28_1.wav	track28_2.wav
/Ni/	track29_1.wav	track29_2.wav
/jA/	track30_1.wav	track30_2.wav
/wA/	track31_1.wav	track31_2.wav

Bibliography

[1] H. Laboratories, “Kratzenstein's resonators,” Haskins Laboratories, 2008. [Online]. Available: http://www.haskins.yale.edu/featured/heads/SIMULACRA/kratzenstein.html. [Accessed 2014 05 1].

[2] P. Taylor, Text-to-Speech Synthesis, Cambridge: Cambridge University Press, 2009.

[3] G. J. Borden, K. S. Harris and L. J. Raphael, “7 Research Tools in Speech Science,” in *Speech Science Primer Physiology, Acoustics, and Perception of Speech Third Edition*, Baltimore, Williams & Wilkins, 1994.

[4] S. Bilbao, “Modeling of Complex Geometries and Boundary Conditions in Finite Difference/Finite Volume Time Domain Room Acoustics Simulation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1524 - 1533, 2013.

[5] J. Mullen, “Physical Modelling of the Vocal Tract with the 2D Digital Waveguide Mesh,” A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Electronics, University of York, York, 2006.

[6] D. Childers, Speech Processing and Synthesis Toolboxes, NY: John Wiley, 2000.

[7] P. B. Denes and E. N. Pinson, “The physics and biology of spoken language. (2nd edn.),” in *Chapter 1 The Speech Chain*, New York, Freeman & Co, 1993, pp. 1-9.

[8] M. Huckvale, “PALS1004 Introduction to Speech Science,” UCL, 11 02 2014. [Online]. Available: http://www.phon.ucl.ac.uk/courses/spsci/iss/week1.php. [Accessed 20 04 2014].

[9] A. J. Ellis, On early English pronunciation, New York: Greenwood Press, 1968.

[10] M. K. C. MacMahon, “The International Phonetic association,” UCL Division of psychology and language sciences, [Online]. Available: http://www.langsci.ucl.ac.uk/ipa/history.html. [Accessed 2014 05 20].

[11] -, “Bell system technical journal vol26-1947,” [Online]. Available: http://www3.alcatel-lucent.com/bstj/vol26-1947/articles/bstj26-1-213.pdf. [Accessed 2014 05 20].

[12] J. Holmes, I. Mattingly and J. Shearme, “Speech synthesis by rule,” *Language Speech,* no. 7, pp. 127-143, 1964.

[13] G. Fant, Acoustic Theory of Speech Production, Netherlands: Mouton: The Hague, 1960.

[14] W. Hardcastle, “The use of electropalatography in phonetic research,” *Phonetica,* vol. 25, pp. 197-215, 1972.

[15] F. Gibbon, “Bibliography of electropalatographic (EPG) studies in English (1957-2005),” [Online]. Available: http://www.qmu.ac.uk/casl/cleftnet/EPG_biblio_2005_september.pdf. [Accessed 2014 05 20].

[16] M. Tal Geva, “Magnetic Resonance Imaging: Historical Perspective,” *Journal of Cardiovascular Magnetic Resonance,* vol. 8, pp. 573-580, 2006.

[17] A. Fourcin, “Laryngographic examination of vocal fold vibration,” in *Wyke B. (ed.) Ventilatory and Phonatory Control System,* London, Oxford University Press, 1974.

[18] J. Perkell, M. Cohen, M. Svirsky, M. Matthies, I. Garabieta and M. Jackson, “Electro-magnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements,” *The Journal of Acoustical Society of America,* vol. 92, pp. 3078-3096, 1992.

[19] M. Uecker, S. Zhang, D. Voit, A. Karaus, K.-D. Merboldt and J. Frahm, “Real-time magnetic resonance imaging at a resolution of 20ms,” *NMR in Biomedicine,* vol. 23, pp. 986-994, 2010.

[20] D. Howard and D. Murphy, Voice Science acoustics and recording, Oxford: Pural Publishing, 2008.

[21] M. a. H. TMH Speech, “TMH KTH :: WaveSurfer,” TMH, Speech, Music adn Hearing, School of Computer Science and Communication, 05 12 2012. [Online]. Available: http://www.speech.kth.se/wavesurfer/. [Accessed 2012].

[22] L. R. Rabiner and R. W. Schafer, Digital processing of speech signals, Prentice-Hall, 1978.

[23] D. Howard and J. Angus, Acoustics and psychoacoustics, Oxford: Focal, 2009.

[24] “Talking Heads: Simulacra - Kratzenstein's resonators,” Haskins Laboratories, 2008. [Online]. Available: http://www.haskins.yale.edu/featured/heads/SIMULACRA/kratzenstein.html. [Accessed 2014 05 20].

[25] “Talking Heads: Simulacra - Von Kempelen's talking machine, 1791,” Haskins Laboratories, 2008. [Online]. Available: http://www.haskins.yale.edu/featured/heads/SIMULACRA/kempelen.html. [Accessed 2014 05 20].

[26] -, “History of Computers and Computing, Automata, Joseph Faber,” history-computer, [Online]. Available: http://history-computer.com/Dreamers/Faber.html. [Accessed 2014 05 20].

[27] H. W. Dudley, “The vocoder,” *Bell Labs Rec.*, vol. 18, pp. 122-126, 1939.

[28] S. Lemmetty, “Review of Speech Synthesis Technology,” Master's Thesis, Laboratory of Acoustics and Audio Signal Processing, Helsingki University of Technology, 1999.

[29] P. K. Tokuda, “HMM-based Speech Synthesis System (HTS),” Department of Computer Science and Engineering, 14 03 2009. [Online]. Available: http://hts.sp.nitech.ac.jp/?Publications. [Accessed 10 01 2014].

[30] E. Keller, G. Bailly, A. Monaghan, J. terken and M. Huckvale, Improvements in Speech

Synthesis COST 258: The Naturalness of Synthetic Speech, West Sussex: John Wiley, 2002.

[31] T. Ogunfunmi and M. Narasimha, “13 - Conclusion and Future Directions of Speech Coding,” in *Principles of Speech Coding*, Florida, Taylor & Francis, 2010, pp. 335-340.

[32] R. D. Kent and C. Read, “Chapter 5: The acoustic characteristics of consonants,” in *The acoustic analysis of speech - 2nd ed.*, NY, Singular Press, 2002.

[33] G. Rosen, “Dynamic Analog Speech Synthesizer,” Technical Report 353, Research Laboratory of Electronics, Massachusetts Institute of Technology, 1960.

[34] J. L. Kelly and C. C. Lochbaum, “Speech Synthesis,” in *the Fourth International Congress on Acoustics*, Copenhagen, Denmark, 1962.

[35] S. Bilbao, “1-Introduction,” in *Wave and Scattering Methods for Numberical Simulation*, Chichester, J. Wiley, 2004.

[36] B. H. Story, “A Parametric model of the vocal tract area function for vowel and consonant simulation,” *Journal of the Acoustical Society of America*, vol. 117, pp. 3231-3254, May, 2005.

[37] T. Arai, “Vocal Tract Models,” Arai Laboratory (Speech Communication Laboratory), Dept. of Information and Communication Sciences, Sophia University, [Online]. Available: http://www.splab.ee.sophia.ac.jp/Vocal_Tract_Model/index-e.htm. [Accessed 1 06 2011].

[38] P. Boersma and D. Weenink, “Praat: doing phonetics by computer,” [Online]. Available: http://www.fon.hum.uva.nl/praat/. [Accessed 08 Dec 2010].

[39] P. Boersma, “Functional Phonology,” Holland Academic Graphics, Netherlands, 1998.

[40] P. Boersma, “An articulatory synthesizer for the simulation of consonants,” in *EUROSPEECH 93*, Berlin, Germany, 1993.

[41] A. Takanishi, “Anthropomorphic Talking Robot,” Takanishi Lab, [Online]. Available:

http://www.takanishi.mech.waseda.ac.jp/top/research/voice/. [Accessed 20 04 2014].

[42] P. Birkholz, “VocalTractLab: Towards high-quality articulatory speech synthesis,” VocalTractLab, 2014. [Online]. Available: http://www.vocaltractlab.de/. [Accessed 2014 05 28].

[43] J. O. Smith III, “Physical Modeling using Digital Waveguides,” *Computer Music Journal,* vol. 16, no. 4, pp. 74-91, 1992.

[44] J. O. Smith III, “Physical Audio Signal Processing,” W3K Publishing, 2010. [Online]. Available: https://ccrma.stanford.edu/~jos/pasp/Digital_Waveguide_Mesh.html. [Accessed 14 02 2013].

[45] D. Murphy, A. Kelloniemi, J. Mullen and S. Shelley, “Acoustic Modelling using the Digital Waveguide Mesh,” *IEEE Signal Processing Magazine,* pp. 55-66, 2007.

[46] S. V. Duyne and J. Smith, “Physical modelling with the 2-D Digital waveguide mesh,” in *Computer Music Conf.,* Tokyo, Japan, 1993.

[47] G. Campos and D. Howard, “On the computational efficiency of different waveguide mesh topologies for room acoustic simulation,” *IEEE Trans, Speech Audio Processing,* vol. 13, no. 5, pp. 1063-1072, 2005.

[48] F. Fontana and D. Rocchesso, “Signal-Theoretic Characterization of Waveguide Mesh Geometries for Models of Two-Dimensional Wave Propa-gation in Elastic Media,” *IEEE Transactions on Speech and Audio Processing,* vol. 9, no. 2, 2001.

[49] J. Mullen, D. Howard and D. Murphy, “Waveguide physical model-ing of vocal tract acoustics: Flexible formant bandwidth control for increased model dimensionality,” *IEEE Trans. Speech Audio Processing,* vol. 14, no. 3, pp. 964-971, 2006.

[50] S. B. Selley, “Diffuse boundary modelling in the digital waveguide mesh,” PhD thesis, Department of Electronics, the University of York, York, 2007.

[51] A. Rugchatjaroen and D. M. Howard, “A Study On Dynamic Vocal Tract Shaping For Diphthong Simulation Using A 2D Digital Waveguide Mesh,” in *the 15th Int.*

Conference on Digital Audio Effects (DAFx-12), York, UK, 2012.

[52] P. Ladeforged and J. K., A course in phonetics-4th ed., Boston: Wadsworth/Cengage Learning, 2011.

[53] P. Ladeforged, Vowels and Consonants : an introduction t othe sounds of languages - 2nd ed., Oxford: Blackwell publishing, 2005.

[54] S. H. Christine, “The Aerodynamics of Speech,” in *The Handbook of Phonetic Sciences Second Edition*, West Sussex, Blackwell, 2013, pp. 39-80.

[55] R. Ogden, An Introduction to English Phonetics, Edinburgh: Edinburgh University Press, 2009.

[56] A. W. Black and K. A. Lenzo, “A Diphone database,” Language Technologies Institute, Carnegie Mellon University , 21 01 2007. [Online]. Available: http://www.festvox.org/festvox/c2261.html. [Accessed 01 10 2011].

[57] K. A. Lenzo and A. W. Black, “Diphone Collection and Synthesis,” in *Sixth International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, China, 2000.

[58] K. Johnson, Acoustic and Auditory Phonetics (Third Edition), Chicester: John Wiley & Sons, 2012.

[59] P. Ladefoged and I. Maddieson, “Nasals and Nasalized Consonants,” in *The sounds of the world's language*, Oxford, Blackwell, 1996.

[60] M. Speed, “Voice Synthesis using the Three-Dimensional Digital Waveguide Mesh.,” PhD thesis, University of York, York, 2012.

[61] J. Mullen, D. Howard and D. Murphy, “Real-Time Dynamic Articulations in the 2D Waveguide Mesh Vocal Tract Model,” *IEEE Transactions on Audio, Speech and Language Processing*, pp. 577-585, 2007.

[62] G. Fant, J. Liljencrants and Q. Lin, “A four-parameter model of glottal flow,” in *STL-*

QPSR, KTH, Stockholm, 1985.

[63] M. Stone, “Chapter 1: Laboratory Techniques for Investigating Speech Articulation,” in *the Handbook of Phonetics*, Oxford, Wiley-Blackwell, 2013.

[64] L. a. M. the Centre for Research on Brain, “The Electromagnetic Articulagraph (EMA),” the Centre for Research on Brain, Language, and Music, 2013. [Online]. Available: www.youtube.com/watch?v=6oIejoZI7j0. [Accessed 30 08 2013].

[65] P. A. Yushkevich, G. Gerig, O. Soldea and Y. Gao, “ITK - SNAP,” [Online]. Available: http://www.itksnap.org/pmwiki/pmwiki.php. [Accessed 01 05 2011].

[66] K. Martin, W. Schroeder and B. Lorensen, “vtk Visualization Toolkit,” [Online]. Available: http://www.vtk.org/. [Accessed 01 05 2011].

[67] K. Richmond, P. Hoole and S. King, “Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory Corpus,” in *Interspeech*, Florence, Italy, 2011.

[68] P. A. Yushkevich, H. Zhang, C. Goodlett, T. Burke and N. Tustison, “itk-SNAP,” the U.S. National Institute of Biomedical Imaging and BioEngineering and the NIH Blueprint for Neuroscience through grant R03 EB008200, 2011. [Online]. Available: www.itksnap.org. [Accessed 01 05 2012].

[69] Kitware, VTK User's Guide, Kitware Inc., 2010.

[70] C. Shadle, “The Aoustics of Fricative Consonants,” Massachusetts Institute of Technology, Research Laboratory of Electronics, Cambridge, Massachusetts, 1985.

[71] P. Jackson, “Characterisation of plosive, fricative and aspiration components in speech production,” PhD Thesis, University of Southampton, Southampton, 2000.

[72] P. Ladeforged, Vowels and Consonants An Introduction to the sounds of languages, Oxford: Blackwell, 2001.

[73] K. N. Stevens, Acoustic Phonetics, Cambridge, Massachusetts: The MIT Press, 1998.

[74] P. Birkholz, “Enhanced area functions for noise source modeling in the vocal tract,” in

10th ISSP, Cologne, 2014.

[75] F. Cox, “Consonant Acoustics: The Acoustic Characteristics of Nasals,” Macquarie University, 2008. [Online]. Available: http://clas.mq.edu.au/speech/acoustics/consonants/nasalweb.html. [Accessed 09 06 2014].

[76] A. Rugchatjaroen and D. M. Howard, “A STUDY ON DYNAMIC VOCAL TRACT SHAPING FOR DIPHTHONG SIMULATION USING A 2D DIGITAL WAVEGUIDE MESH,” in *the 15th Int. Conference on Digital Audio Effects (DAFx-12),* York, UK, 1012.

[77] K. Richmond, “Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus,” in *INTERSPEECH 2011*, Florence, 2011.

[78] P. Ladeforged, A course in phonetics - 4th ed., Boston, MA: Wadsworth/Cengage Learning, 2001.

[79] K. Sjölander and J. Beskow, “WaveSurfer,” The Department of Speech, Music and Hearing, KTH, [Online]. Available: http://www.speech.kth.se/wavesurfer/index2.html. [Accessed 08 Dec 2010].

[80] C. Shadle, “The effect of geometry on source mechanisms of fricative consonants,” *Journal of Phonetics,* vol. 19, pp. 409-424, 1991.

[81] P. C. Delattre, A. M. Liberman and F. S. Cooper, “Acoustic Loci and Transitional Cues for Consonants,” *JASA,* vol. 27, no. 4, pp. 769-773, 1955.

[82] P. Badin, “Acoustics of voiceless fricatives: production theory and data,” Speech Transmission Laboratory Quarterly Progress and Status Report, Stockholm, 1989.

[83] W. J. Hardcastle, J. Laver and F. E. Gibbon, The Handbook of Phonetic Sciences Second Edition, Oxford, UK: Wiley-Blackwell, 2013.

[84] K. N. Stevens, “Airflow and turbulence noise for fricative and stop consonants: static considerations,” *J. Acoust. Soc. Am.,* vol. 50, no. 4:2, pp. 1180-1192, 1971.

[85] K. N. Stevens, “On the quantal nature of speech,” *Journal of Phonetics,* vol. 17, pp. 3-45, 1989.

[86] J. Kim, N. Kumar, S. Lee and S. Narayanan, “Enchanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data,” in *10th ISSP Cologne,* Cologne, 2014.

[87] “The INTERNATIONAL PHONETIC ASSOCIATION,” UCL DIVISION OF PSYCHOLOGY AND LANGUAGE SCIENCES, [Online]. Available: https://www.langsci.ucl.ac.uk/ipa/fullchart.html. [Accessed 01 01 2014].

[88] I. Conexant Systems, “Conexant,” Conexant Systems, Inc., 2012. [Online]. Available: http://www2.conexant.com/Product/Audio/pchdaudio/CX20672/Pages/default.aspx. [Accessed 2014].

[89] S. Bilbao, Numerical Sound Synthesis, Edinburgh, UK: John Wiley & Sons, 2009.

[90] Y. Fujiso, A. V. Hirtum, K. Nozaki and S. Wada, “Study of unvoiced fricative speech production: Influence of initial conditions on flow development,” *J. Acoust. Soc. Am.* , vol. 133, no. 5, 2013.