Finite-Difference Simulations of Speech with Wall Vibration Losses

Chad McKell

A special project dissertation submitted in partial fulfilment
of the requirements for the degree of

Master of Science (MSc)

Acoustics and Music Technology

Acoustics and Audio Group
Edinburgh College of Art
University *of* Edinburgh

April 2017

Supervisor: Dr. Stefan Bilbao

Abstract

Physical modeling synthesis techniques are attractive computational methods for producing artificial speech. In this study, finite-difference simulations of vocal-tract sound propagation were performed to create realistic vowels, diphthongs, and phrases. Wall vibration losses in the vocal tract were included by coupling a damped forced oscillator to Webster's equation. Comparisons of the formant structures of synthesized and recorded speech suggest that finite-difference simulations are accurate tools for generating natural speech.

Declaration

I do hereby declare that this dissertation was composed by myself and that the work described within is my own, except where explicitly stated otherwise.

Chad McKell
April 2017

Contents

Abstract **i**

Declaration **iii**

Contents **v**

List of figures **vii**

List of tables **ix**

1 Introduction **1**

1.1 Physical Modeling 1

1.2 Motivation: Wall Vibration Losses 2

1.3 Summary and Results 2

2 Physical Models **3**

2.1 Acoustic Tubes 3

2.2 The Vocal Tract 4

3 Finite-Difference Simulations **7**

3.1 Vowels 7

3.2 Diphthongs 11

3.3 Wall Vibration Losses 15

3.4 Speech Phrases 16

4 Conclusions **19**

4.1 Results and Evaluation 19

4.2 Future Directions 19

A Code **21**

A.1 Code Explanation 21

A.2 Diphthong with Wall Losses 22

B Project Proposal **27**
B.1 Introduction 27
B.2 Objectives 28
B.3 Equipment 28
B.4 Plan 29

Bibliography **29**

List of Figures

3.1 Waveform of synthesized glottal input signal $g[n]$. 9

3.2 Formant structures of the synthesized vowels /a/, /i/, /ʊ/, and /u/. The area function S of each synthesized vowel was computed using MRI data from a test subject (BS). The first three formants taken from audio recordings of the vowels spoken by BS are drawn as dotted lines. In most cases, the synthesized formants generally agree with the recorded formants. 10

3.3 Plot of the interpolation function `beta`. 12

3.4 Waveform (top) and spectrogram (bottom) of the synthesized diphthong /ai/. The transition period occurs between $t = 0.16$ and 0.24 seconds. . 13

3.5 Waveform (top) and spectrogram (bottom) of the synthesized diphthong /ʊu/. The transition period occurs between $t = 0.16$ and 0.24 seconds. . 14

3.6 Formant structures of the vowels /a/, /i/, /ʊ/, and /u/—synthesized without damping (solid gray curves), synthesized with damping (solid black curves), and recorded (dashed curves). For the synthesized vowels, damping loss lowered the energy of the formant structures, bringing them into closer agreement with the formant envelopes of the recorded vowels. 16

3.7 Spectrograms of the synthesized sentence "I owe you a yo-yo". The plots correspond to synthesized speech without damping (top) and with damping (bottom). From left to right, the spectrogram plots the sounds /ai/ /au/ /iu/ /a/ /iʊ/ /iʊ/. 17

List of Tables

3.1 Measured formant frequencies of /a/, /i/, /ʊ/, and /u/. The formant frequencies f^S correspond to vowels synthesized from vocal-tract MRI data while the formant frequencies f^N correspond to natural recorded vowels. The MRI data and audio recordings were taken from the same test subject (BS). 11

Chapter 1

Introduction

Speech synthesis dates back to at least 1791 when Wolfgang von Kempelen debuted the third iteration of his mechanical speaking machine [1]. The device used a variety of parts to carefully imitate the human speech system—a bellows for the lungs, a reed for the vocal folds, tubes for the various vocal-tract geometries, and so on. By reproducing the subtleties of linguistic sounds from observations of the acoustic and physiological mechanisms of speech, Kempelen set the stage for more advanced physical modeling techniques that would emerge centuries later.

1.1 Physical Modeling

Physical modeling speech synthesis is a computational approach to artificial voice production that generates acoustic sounds by numerically solving a mathematical model of speech. Compared to other modern methods of speech synthesis, such as concatenative synthesis or statistical parametric synthesis, physical modeling synthesis appears to be more promising for improving vocal expressivity [2], synthesizer control and manipulation [3], articulatory-driven facial animation [4], and speech science pedagogy [5].

The first physically-modeled simulation of human speech was implemented in 1961 by Kelly and Lochbaum [6]. The Kelly-Lochbaum method represents the vocal tract as a one-dimensional acoustic tube consisting of a series of concatenated cylindrical sections. Instead of directly solving the wave equation for an acoustic signal moving along the tube, the method computes the volume velocity and pressure of the signal using digital delay lines and scattering methods [7]. In its infancy, the algorithm was welcomed for its superior computational speed. However, with the more sophisticated hardware capabilities that have since emerged, the technique has been surpassed by other methods. Moreover, its indirect approach to solving the wave equation introduces complications, particularly for tuning formant frequencies [8].

Direct methods of numerical simulation, such as finite-difference methods, pose attractive alternatives to the Kelly-Lochbaum method. Simulations of vowels and

diphthongs follow intuitively from an equation of motion and require little computational overhead. Moreover, since modifications to the sound output are influenced directly by physical parameters, finite-difference methods are uniquely poised to improve control and expressivity of synthetic speech.

1.2 Motivation: Wall Vibration Losses

A more refined physical speech model paired with a low-cost numerical simulation scheme is required in order to advance speech synthesis technology. In particular, a description of sound propagation along the vocal tract is one aspect of the physical model that remains incomplete.

Current descriptions of vocal-tract sound propagation do not account for several crucial phenemena, including wall vibration losses. Wall vibration losses are often modeled by coupling a forced damped oscillator to the wave equation, where the mass and stiffness of the oscillator are determined by the properties of the vocal tract. Recent work has shown that multiple oscillators arising from distinct layers of the subglottal wall material can be used to more accurately simulate formant frequencies [9]. In preparation for incorporating multiple oscillators into a loss model for the vocal tract, this dissertation provides preliminary analysis and simulations of wall losses resulting from a single oscillator.

1.3 Summary and Results

In Chapter 2, the physical theory of acoustic tubes and vocal-tract sound propagation with wall vibration losses is outlined. In Chapter 3, finite-difference simulations of vocal-tract sound propagation were performed to produce the speech sounds /*a*/, /i/, /ʊ/, /u/, /*a*i/, and /ʊu/, both with and without wall vibration losses. Comparisons of the formant structures of synthesized and recorded speech suggest that finite-difference simulations are accurate tools for producing natural speech. Chapter 4 concludes with a discussion of the results and suggestions for future work on wall losses with multiple damped oscillators.

Chapter 2

Physical Models

A review of the fundamental mathematical descriptions of acoustic tubes is provided. The results are then extended to sound propagation in a vocal tract with yielding walls.

2.1 Acoustic Tubes

Propagating sound waves in a tube satisfy the following system of equations derived from first principles in fluid dynamics [10]:

$$p_x = -\rho(u/S)_t \tag{2.1a}$$

$$u_x = -\frac{1}{\rho c^2}(pS)_t + S_t \tag{2.1b}$$

where $p = p(x,t)$ is the acoustic pressure, $u = u(x,t)$ is the volume velocity, $S = S(x,t)$ is the cross-sectional area function of the tube, ρ is the density of the propagation medium, and c is the speed of sound. Here, Eq. (2.1a) corresponds to conservation of fluid momentum and Eq. (2.1b) to conservation of total fluid mass. In terms of notation, note that $p_x = \frac{\partial p}{\partial x}$, $(u/S)_t = \frac{\partial (u/S)}{\partial t}$, and so forth. If S is independent of time, Eqs. (2.1) reduce to

$$p_x = -\frac{\rho}{S}u_t \tag{2.2a}$$

$$u_x = -\frac{S}{\rho c^2}p_t \tag{2.2b}$$

Combining Eqs. (2.2), we get Webster's Equation for pressure and volume velocity:

$$\frac{1}{S}(Sp_x)_x \;=\; \frac{1}{c^2}p_{tt} \tag{2.3a}$$

$$\frac{1}{S}(Su_x)_x \;=\; \frac{1}{c^2}u_{tt} \tag{2.3b}$$

For convenience in applying numerical schemes to Webster's equation, Eqs. (2.3) may be re-written as a single equation in terms of an arbitrary variable $\Psi = \Psi(x,t)$ [8]:

$$S\Psi_{tt} \;=\; c^2(S\Psi_x)_x \tag{2.4}$$

where $p \;=\; \rho\Psi_t$ and $u \;=\; -S\Psi_x$. Scaling the pressure, volume velocity, and variable Ψ as $p' = p/\rho c^2$, $u' = u/cS_0$, and $\Psi' = \Psi/cL$, Eq. (2.4) becomes

$$S\Psi_{tt} \;=\; \gamma^2(S\Psi_x)_x \tag{2.5}$$

after dropping primes. Here, $\gamma = c/L$ is the time required for sound to travel the full length L of the vocal tract and S_0 is a reference area chosen to be equal to the cross-sectional area at $x = 0$, i.e. $S_0 = S(0)$. The non-dimensional forms of the pressure p and volume velocity u are now given in terms of the non-dimensional variable Ψ as

$$p \;=\; \frac{1}{\gamma}\Psi_t \tag{2.6a}$$

$$u \;=\; -S\Psi_x \tag{2.6b}$$

2.2 The Vocal Tract

Webster's equation can be solved for the case of a vocal tract as long the area function and boundary conditions of the system are specified. The area function S is typically approximated by interpolating cross-sectional area measurements taken at equal intervals along the vocal tract during phonation. The boundary conditions are dictated by the nature of the input excitation signal at the glottis and the radiation load at the mouth. A stable system of conditions is given by [8]:

$$\Psi_x(0,t) \;=\; g(t) \tag{2.7a}$$

$$\Psi_x(L,t) \;=\; -\alpha_1\Psi_t(L,t) - \alpha_2\Psi(L,t) \tag{2.7b}$$

where $g(t)$ is the glottal input signal, $x = 0$ is the position of the glottis, $x = L$ is the position of the mouth, and the constants α_1 and α_2 are determined from parameters of the vocal tract. For the case of a vocal tract terminating on an infinite plane, α_1 and α_2 may be set to

$$\alpha_1 \;=\; \frac{1}{2(0.8216)^2\gamma} \tag{2.8a}$$

$$\alpha_2 \;=\; \frac{L}{0.8216\sqrt{S_0 S(L)/\pi}} \tag{2.8b}$$

Applying the boundary conditions to Eq. (2.5), a scaled form of the acoustic pressure at all points along the vocal tract may be computed.

Wall Vibration Losses

Losses arising from the vibrations of yielding walls in the vocal tract may be modeled by coupling a forced damped oscillator to Webster's equation. Moving along the coupling coordinate $w = w(w,t)$, the forced damped oscillator is described by the following relation [8]:

$$w_{tt} + 2\sigma_0 w_t + \omega_0^2 w \;=\; \epsilon S^{1/4}\Psi_t \tag{2.9}$$

Here, σ_0 is the damping coefficient, ω_0 is the fundamental frequency of oscillation, and ϵ is the coupling coefficient. The coupling coefficient ϵ is given by

$$\epsilon \;=\; c\sqrt{\frac{s\rho}{M}}\left(\frac{\pi}{S_0}\right)^{1/4} \tag{2.10}$$

where c is the speed of sound, ρ is the density of air in the vocal tract, and M is the mass per unit area of the vocal-tract walls. The oscillator is coupled to Webster's equation in the following way:

$$S\Psi_{tt} \;=\; \gamma^2(S\Psi_x)_x - \epsilon S^{1/4} w_t \tag{2.11}$$

In Section 3.3, w and Ψ are arranged as explicit updates in a finite-difference scheme, allowing the acoustic pressure to be computed at all points and time steps.

Chapter 3

Finite-Difference Simulations

In this chapter, finite-difference simulations are used to numerically compute the discrete pressure signal in a vocal tract with time-invariant and time-varying area functions. The time-invariant function produces sounds that mimic linguistic vowels while the time-varying function makes signals that resemble diphthongs. Wall vibration losses in the vocal tract are calculated by coupling a damped forced oscillator to Webster's equation. Finally, two synthesized speech phrases are compared—one with wall vibration losses and one without.

3.1 Vowels

Speech vowels were produced by inserting the boundary conditions in Eqs. (2.7) into an explicit numerical solution to the form of Webster's equation expressed in Eq. (2.5). In finite-difference notation, Webster's equation becomes [8]

$$[S]_l \delta_{tt} \Psi_l^n \;=\; \gamma^2 \delta_{x+}((\mu_{x-} S)(\delta_{x-} \Psi_l^n) \tag{3.1}$$

with the boundary conditions given by

$$\delta_{x.} \Psi_0^n \;=\; -g[n] \tag{3.2a}$$

$$\delta_{x.} \Psi_N^n \;=\; -\alpha_1 \delta_{t.} \Psi_N^n - \alpha_2 \mu_{t.} \Psi_N^n \tag{3.2b}$$

Here, n is the time step, l is the spatial step, $[S]_l$ is the area vector averaged over all spatial steps, $g[n]$ is the input signal at time step n, and $\gamma = c/L$ is the time required for sound to travel the full length of the vocal tract. The finite difference operators $\delta_{t.}$, δ_{tt}, δ_{x+}, δ_{x-}, and $\delta_{x.}$ are numerical approximations to time and spatial derivatives. The

operators $\mu_{t.}$ and μ_{x-} are averaging terms. Below is a complete list of the operators used for computing speech sounds in this report:

$$\delta_{t+} = \frac{1}{k}(e_{t+} - 1) \tag{3.3a}$$

$$\delta_{t.} = \frac{1}{2k}(e_{t+} - e_{t-}) \tag{3.3b}$$

$$\delta_{tt} = \frac{1}{k^2}(e_{t+} - 2 + e_{t-}) \tag{3.3c}$$

$$\delta_{x+} = \frac{1}{h}(e_{x+} - 1) \tag{3.3d}$$

$$\delta_{x-} = \frac{1}{h}(1 - e_{x-}) \tag{3.3e}$$

$$\delta_{x.} = \frac{1}{2h}(e_{x+} - e_{x-}) \tag{3.3f}$$

$$\delta_{xx} = \frac{1}{h^2}(e_{x+} - 2 + e_{x-}) \tag{3.3g}$$

$$\mu_{t.} = \frac{1}{2}(e_{t+} + e_{t-}) \tag{3.3h}$$

$$\mu_{x-} = \frac{1}{2}(1 + e_{x-}) \tag{3.3i}$$

$$\mu_{xx} = \frac{1}{4}(e_{x+} + 2 + e_{x-}) \tag{3.3j}$$

Here, k is the duration between time samples, h is the distance between spatial samples, e_{t+} and e_{t-} are operators that shift a function forward and backward in time, respectively, and e_{x+} and e_{x+} are operators that shift a function forward and backward in space, respectively.

To solve for the area function S corresponding to the vowels /a/, /i/, /ʊ/, and /u/, magnetic resonance imaging (MRI) measurements of the vowel shapes created by one test subject were taken from the literature [11]. The male subject (BS) was a 29 year-old native speaker of American English who had no history of speech or voice disorders. The measurements were inserted into a two-column matrix on which linear interpolation was applied to produce a continuous area functon S. Next, the explicit solution to Eq. (3.1) was determined to be

$$\begin{aligned}\Psi_l^{n+1} = {} & \frac{\lambda^2(S_{l+1} + S_l)}{2[S]_l}\Psi_{l+1}^n + \frac{\lambda^2(S_l + S_{l-1})}{2[S]_l}\Psi_{l-1}^n \\ & + \left(2 - \frac{\lambda^2(S_{l+1} + 2S_l + S_{l-1})}{2[S]_l}\right)\Psi_l^n - \Psi_l^{n-1}\end{aligned} \tag{3.4}$$

The value of Ψ_l^{n+1} was computed for all spatial points and time steps. In the MATLAB programming language, the update for Ψ_l^{n+1} between the second and penultimate spatial steps took the form

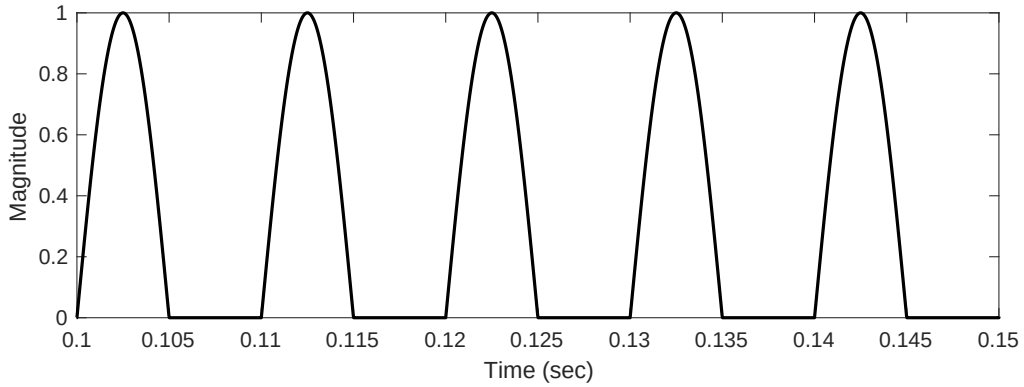


Figure 3.1: Waveform of synthesized glottal input signal $g[n]$.

```
psi(2:N) = sr.*psi1(3:N+1) + sl.*psi1(1:N-1) + s0*psi1(2:N) - psi2(2:N);
```

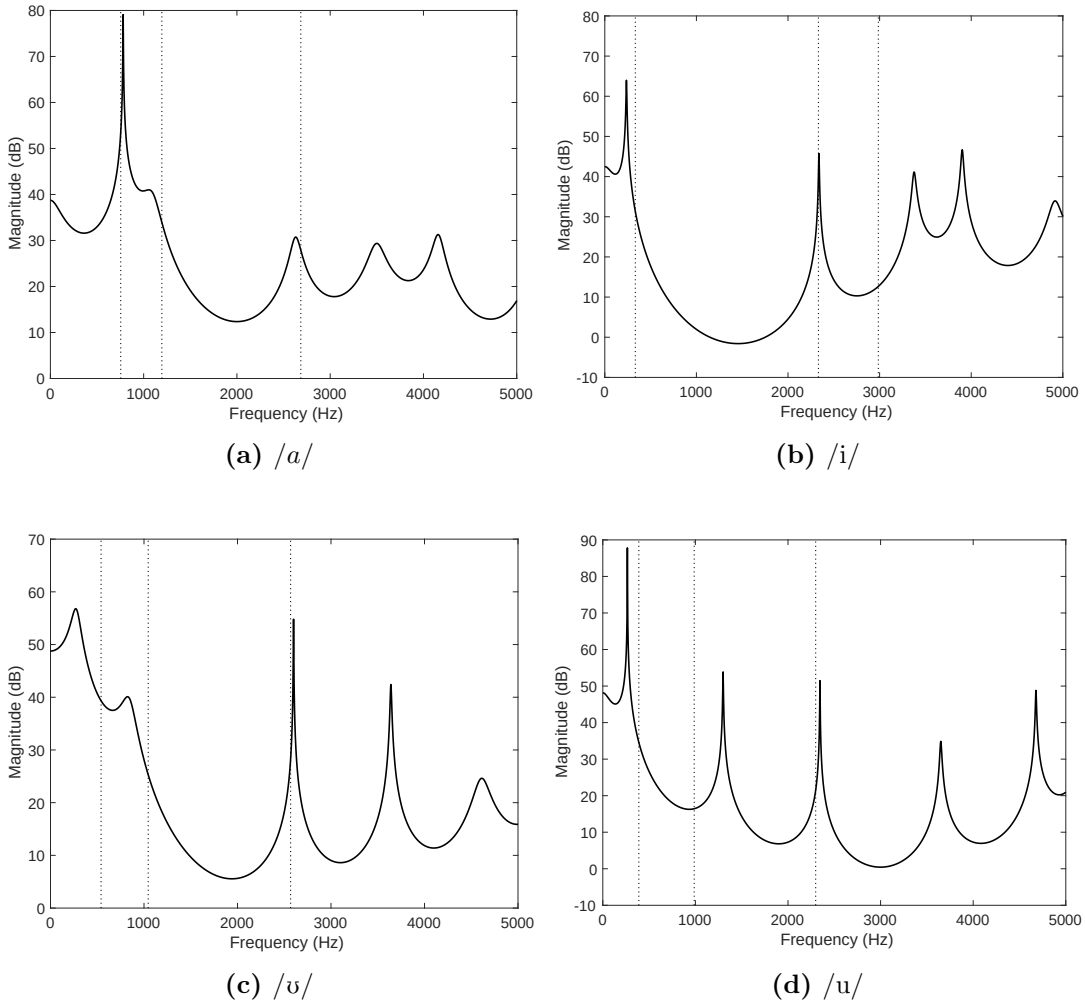
where `sr`, `sl`, and `s0` are the coefficients in Eq. (3.4), and `psi`, `psi1`, and `psi2` represent Ψ^{n+1}, Ψ^{n}, and Ψ^{n-1}, respectively. To calculate the value of Ψ^{n+1} at the glottis and at the mouth, the virtual values Ψ^{n}_{-1} and Ψ^{n}_{N+1} were first solved for in Eqs. (3.2) and then inserted into Eq. (3.4) in turn. After some algebraic manipulation, the updates for the first value Ψ^{n+1}_{0} and the last value Ψ^{n+1}_{N} were determined. In MATLAB, the expressions were written as

```
psi(1) = gr*psi1(2) + gx*uin(n) + g0*psi1(1) - psi2(1);
psi(N+1) = r*psi1(N+1) + rl*psi1(N) + r0*psi2(N+1);
```

where the coefficients `gr`, `gx`, `g0`, `r`, `rl`, and `r0` are defined in Appendix A.2. Note that 0→1 and N→N+1 in MATLAB notation. The input signal $g[n]$ = `uin(n)` was defined to be a sinusoidal wave with the troughs set to zero, as shown in the code below:

```
uin = sin(2*pi*f0*n);
uin = 0.5*(uin+abs(uin));
```

where `f0` is the fundamental frequency of the wave and `n` is the time step. A plot of $g[n]$ = `uin` is shown in Fig. 3.1. Finally, the acoustic pressure was computed by solving Eq. (2.6a) numerically as follows:



(a) /*a*/ **(b)** /i/

(c) /ʊ/ **(d)** /u/

Figure 3.2: Formant structures of the synthesized vowels /*a*/, /i/, /ʊ/, and /u/. The area function S of each synthesized vowel was computed using MRI data from a test subject (BS). The first three formants taken from audio recordings of the vowels spoken by BS are drawn as dotted lines. In most cases, the synthesized formants generally agree with the recorded formants.

$$p \;=\; \frac{1}{\gamma}\delta_{x+}\Psi_l^n \tag{3.5}$$

To simulate the output pressure signal $p = \mathrm{out(n)}$, the following line of code was added directly after the updates for `psi(1)`, `psi(2:N)`, and `psi(N+1)`:

```
out(n) = (fs/gamma)*(psi(N+1) - psi1(N+1));
```

where `fs` is the sampling frequency. The Fourier transform of the output pressure signal was computed for each vowel, and linear predictive coding was applied to extract

the corresponding formant structures. As shown in Fig. 3.2, the first three formants of the synthesized vowels approximately matched the corresponding formants of recorded vowels from the test subject (BS). Numerical values of the formant frequencies from synthesized and recorded vowels are provided in Table 3.1. The values of the natural recorded formant frequencies f^N were taken from the literature [11, Table IV] while the values of the synthesized formant frequencies f^S were measured using the Praat software tool.

Table 3.1: Measured formant frequencies of /*a*/, /i/, /ʊ/, and /u/. The formant frequencies f^S correspond to vowels synthesized from vocal-tract MRI data while the formant frequencies f^N correspond to natural recorded vowels. The MRI data and audio recordings were taken from the same test subject (BS).

Vowel	Formant	Frequency, f^N (Hz)	Frequency, f^S (Hz)
/*a*/	1	754	810
	2	1195	1371
	3	2685	2955
/u/	1	333	212
	2	2332	2242
	3	2986	3305
/ʊ/	1	541	463
	2	1045	2587
	3	2568	3594
/u/	1	389	291
	2	987	1380
	3	2299	2702

3.2 Diphthongs

The dipthongs /*a*i/ and /ʊu/ were synthesized by adding an interpolation function to the vowel simulation described in Section 3.1. In the case of /*a*i/, the interpolation function `beta` was used to transition the global area function `S` from `Sa`, the area function of /*a*/, to `Si`, the area function of /i/, over a finite transition duration. The transition algorithm was programmed as follows:

```
S = (1 - beta(n))*Sa + beta(n)*Si;
```

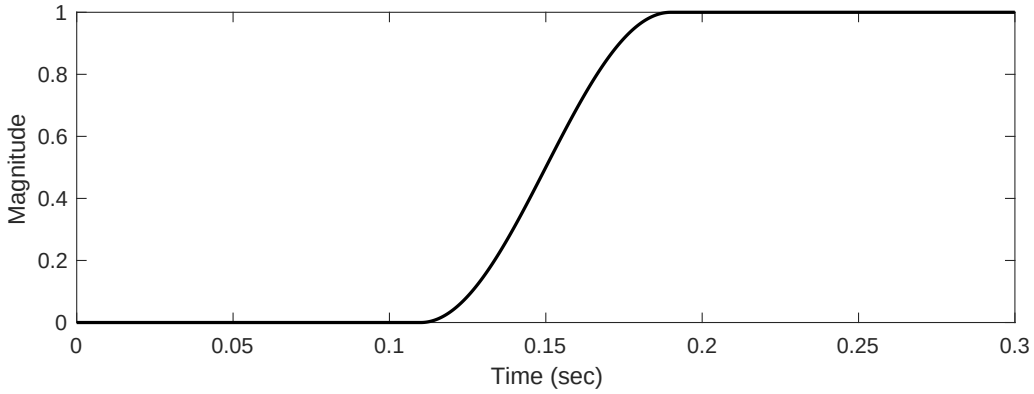


Figure 3.3: Plot of the interpolation function `beta`.

The interpolation function `beta` was defined to be a vector with three states—a 0 state, a half-cosine state, and a 1 state, as defined below

```
beta = ones(NF, 1);
beta(startT:endT-1) = 0.5*(1 - cos(pi*(0:durint-1)/durint));
beta(1:startT) = 0;
```

In time steps, `NF` is the total duration of the simulation, `startT` is the start of the transition, `endT` is the end of the transition, and `durint` is the duration of the transition. A plot of the function `beta` is provided in Fig. 3.3.

Given that the value of the area function S changed at every time step during the transition period, the coefficients `s0`, `sl`, `sr`, `g0`, `gx`, `gr`, `r0`, `rl`, and `r`, along with their defining variables, were added into the main update loop (see Appendix A.2). Waveform plots and spectrograms of the synthesized diphthongs /*a*i/ and /ʊu/ are shown in Fig. 3.4 and Fig. 3.5, respectively. The transition period is clearly visible in each figure.

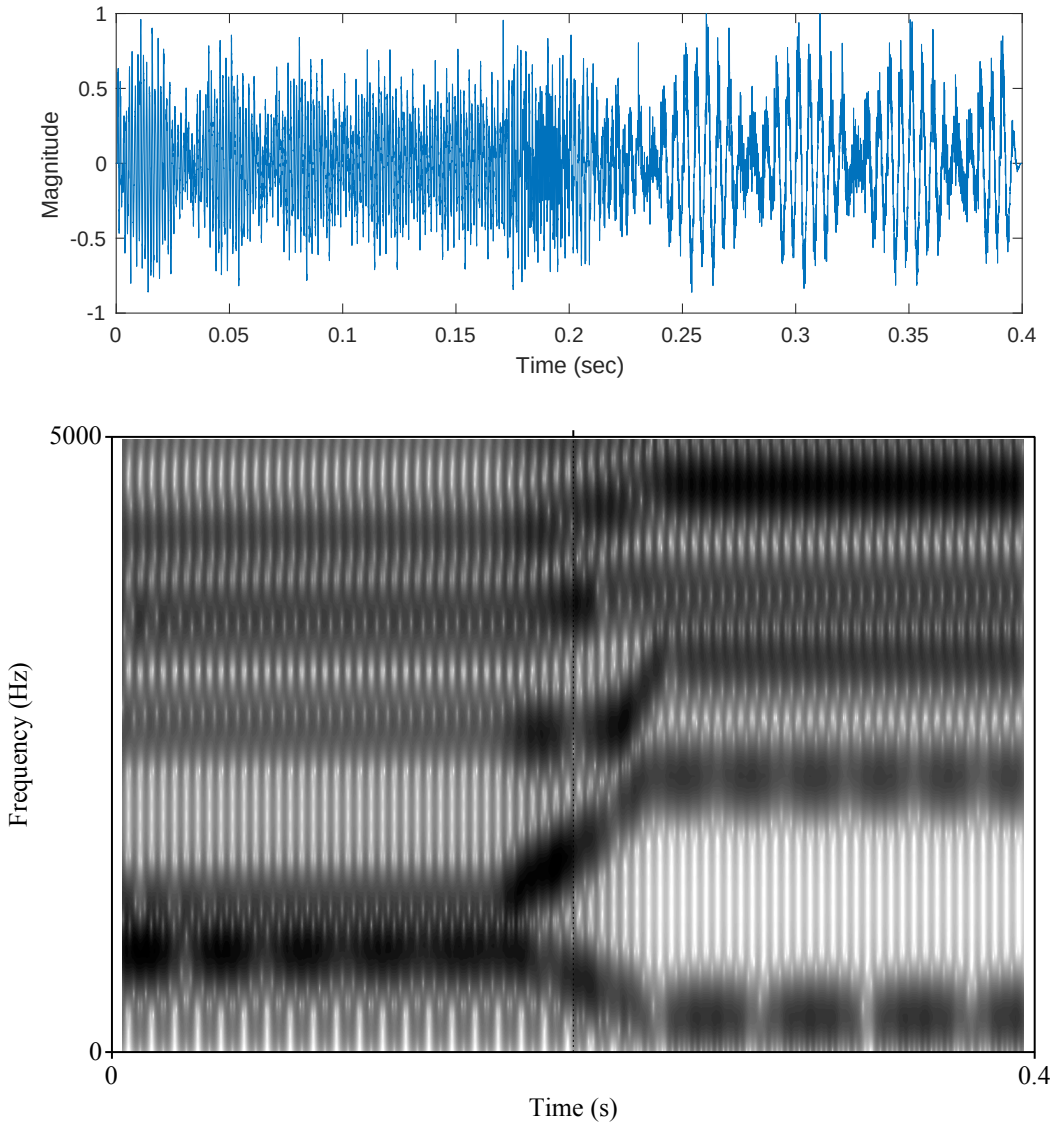


Figure 3.4: Waveform (top) and spectrogram (bottom) of the synthesized diphthong /*a*i/. The transition period occurs between $t = 0.16$ and 0.24 seconds.

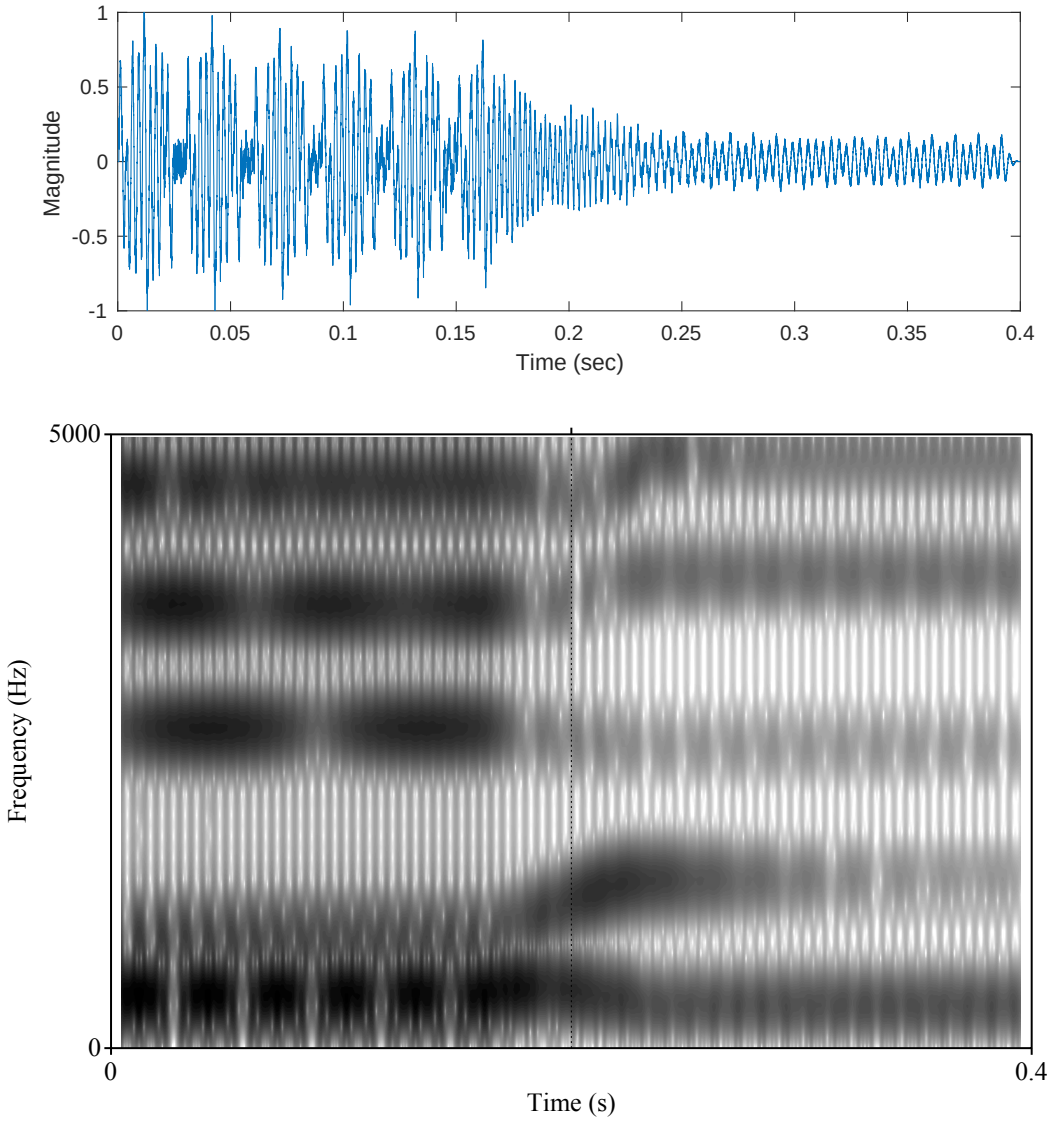


Figure 3.5: Waveform (top) and spectrogram (bottom) of the synthesized diphthong /ʊu/. The transition period occurs between $t = 0.16$ and 0.24 seconds.

3.3 Wall Vibration Losses

In Section 2.2, it was shown that damping losses resulting from the vibration of reacting walls in the vocal tract may be modeled by the system of equations Eq. (2.9) and (2.11). Solving for w_l^{n+1} in Eq. (2.9), we get

$$w_l^{n+1} = \frac{1}{\frac{1}{k^2}+\frac{\sigma_0}{k}}\left[\left(\frac{2}{k^2}-\omega_0^2\right)w^n + \left(\frac{\sigma_0}{k}-\frac{1}{k^2}\right)w^{n-1} + \frac{\epsilon S^{1/4}}{2k}\left(\Psi_l^{n+1}-\Psi_l^{n-1}\right)\right] \tag{3.6}$$

Inserting the difference $(w_l^{n+1}-w_l^{n-1})$ into Eq. (2.11), we arrive at the following solution for Ψ_l^{n+1}:

$$\begin{aligned}\Psi_l^{n+1} = \frac{1}{1+\frac{BE}{A}}\Bigg[&\frac{\lambda^2(S_{l+1}+S_l)}{2[S]_l}\Psi_{l+1}^n + \frac{\lambda^2(S_l+S_{l-1})}{2[S]_l}\Psi_{l-1}^n \\ &+\left(2-\frac{\lambda^2(S_{l+1}+2S_l+S_{l-1})}{2[S]_l}\right)\Psi_l^n \\ &+\left(\frac{BE}{A}-1\right)\Psi_l^{n-1} - \frac{BC}{A}w_l^n - \frac{BD}{A}w_l^{n-1}\Bigg]\end{aligned} \tag{3.7}$$

where the terms A, B, C D, and E are defined as follows:

$$A = \frac{1}{k^2}+\frac{\sigma_0}{k} \tag{3.8a}$$

$$B = \frac{k\epsilon S^{1/4}}{2[S]_l} \tag{3.8b}$$

$$C = \frac{2}{k^2}-\omega_0^2 \tag{3.8c}$$

$$D = -\frac{2}{k^2} \tag{3.8d}$$

$$E = \frac{\epsilon S^{1/4}}{2k} \tag{3.8e}$$

As with the undamped vowel simulation described in Section 3.1, the updates for `psi(1)`, `psi(2:N)`, `psi(N+1)`, and `out(n)` were programmed using the expression for Ψ_l^{n+1}.

Formant structures of the vowels /*a*/, /i/, /ʊ/, and /u/, both with and without damping loss, are provided in Fig. 3.6. As expected, the damping loss lowered the energy of the formant envelopes. The plots in Fig. 3.6 also show formant structures of recorded vowels provided by the UCLA Phonetics Lab website [12]. The recordings of the male test subject John Wells were analyzed. Although some variability in the positions of the formant frequencies is to be expected with different test subjects, the plots indicate that the addition of damping brought the formant envelopes from the

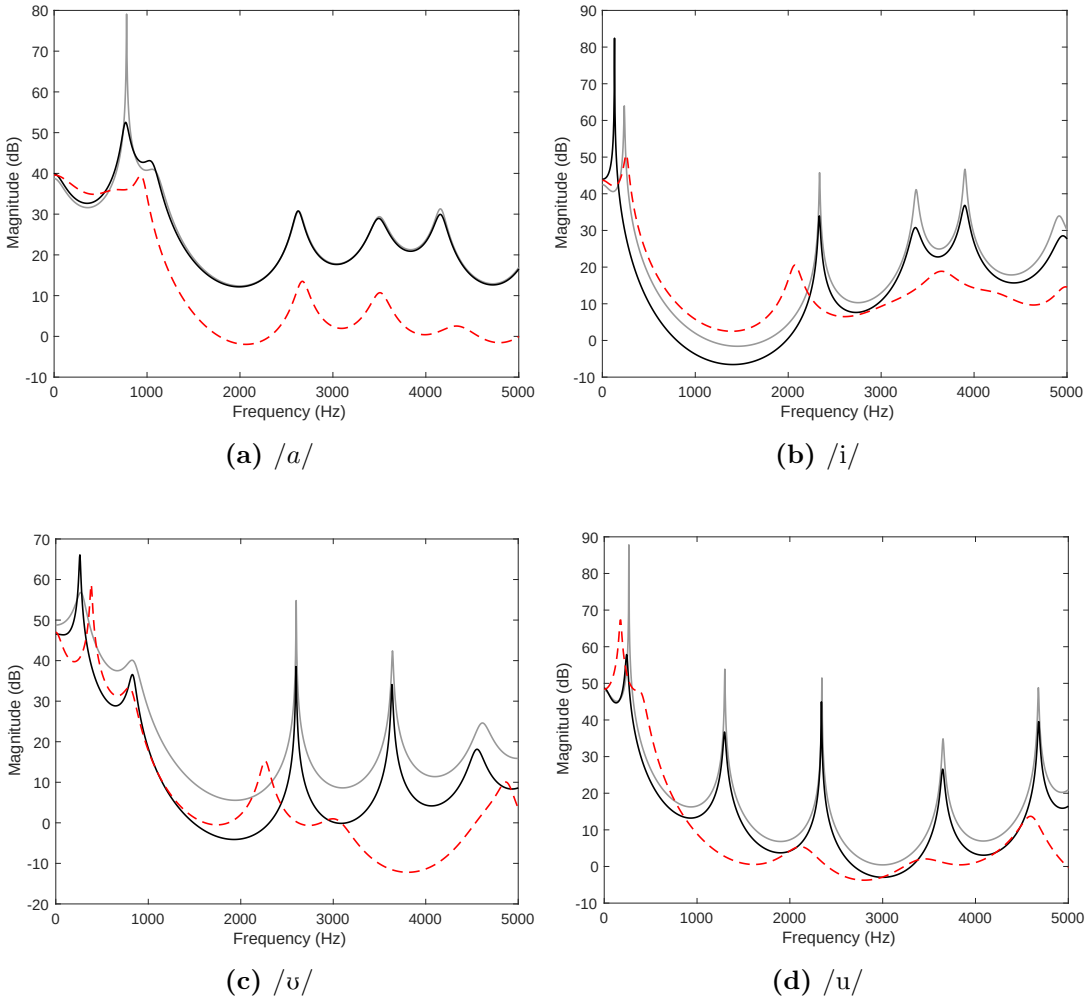


Figure 3.6: Formant structures of the vowels /*a*/, /i/, /ʊ/, and /u/—synthesized without damping (solid gray curves), synthesized with damping (solid black curves), and recorded (dashed curves). For the synthesized vowels, damping loss lowered the energy of the formant structures, bringing them into closer agreement with the formant envelopes of the recorded vowels.

synthesized vowels into closer alignment with those of the recorded vowels. This result suggests that synthesized speech benefits from the addition of wall vibration losses.

3.4 Speech Phrases

Sounds from the vowel and diphthong simulations were concatenated in order to produce two instances of the English sentence:

"I owe you a yo-yo."

which was chosen due to its lack of consonant sounds. One version of the phrase included wall vibration losses while the other did not. As shown in Fig. 3.7, the F1 formants of the speech sounds appear to contain less energy after the addition of damping losses,

particularly the F1 formant of the vowel /ʊ/ in the word "yo-yo". This loss of energy seemed to improve the intelligibility of the word.

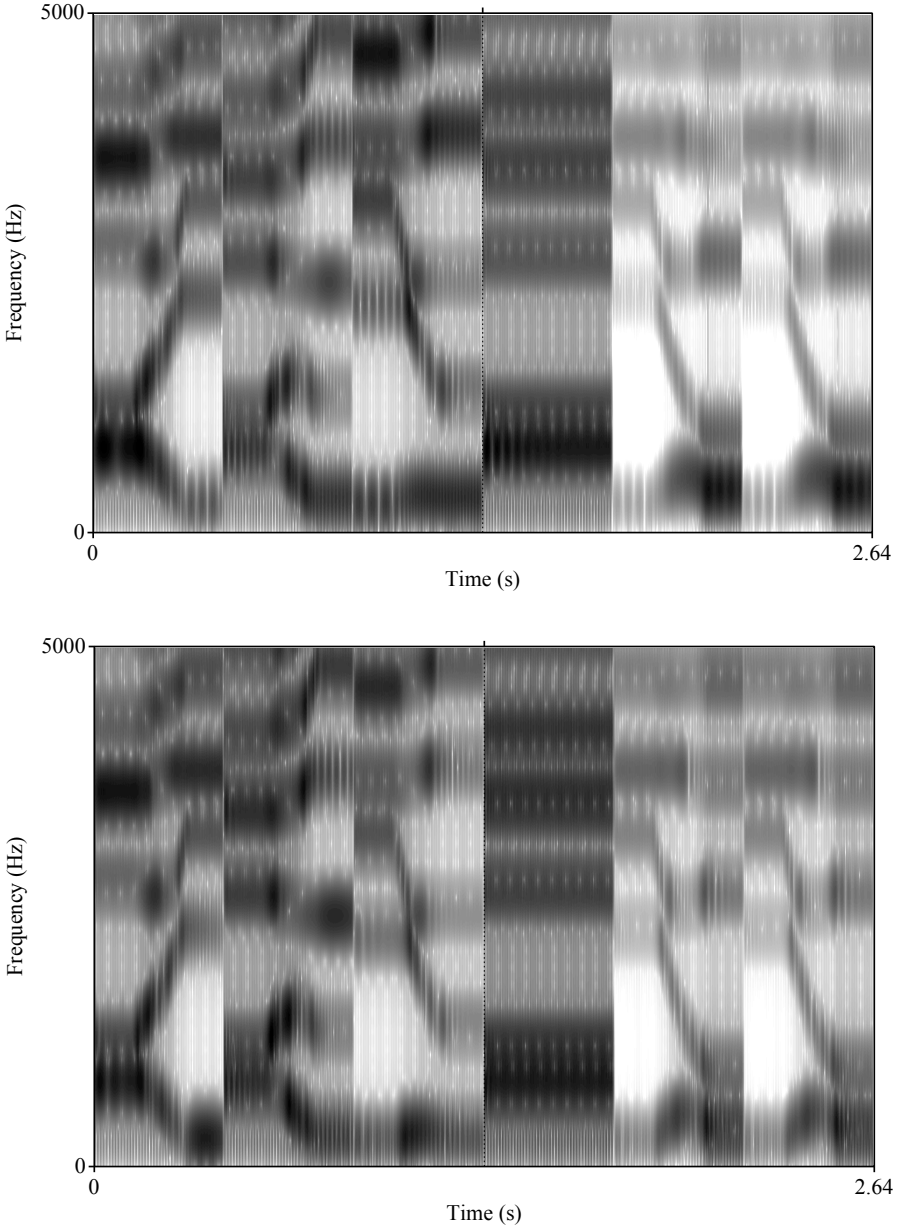


Figure 3.7: Spectrograms of the synthesized sentence "I owe you a yo-yo". The plots correspond to synthesized speech without damping (top) and with damping (bottom). From left to right, the spectrogram plots the sounds /*a*i/ /*a*u/ /iu/ /*a*/ /iʊ/ /iʊ/.

Chapter 4

Conclusions

4.1 Results and Evaluation

Comparisons of the formant structures of synthesized and recorded speech in this report suggest that finite-difference simulations are promising tools for replicating natural speech. In particular, the inclusion of wall vibration losses proved to lower the energy of the formant envelopes of synthesized vowels and diphthongs, bringing them into closer agreement with natural speech. In regards to the original project proposal (see Appendix B), some minor alterations to the plan were made. The most significant of these was the removal of the Kelly-Lochbaum scheme from the timeline and the added emphasis given to finite-difference methods.

4.2 Future Directions

A more refined physical model is required in order to produce more expressive and realistic speech synthesis. A recent study showed that including damping from two separate layers of the subglottal region more accurately simulated natural formant envelopes [9]. In addition to improving the accuracy of finite-difference simulations of vocal-tract sound propagation, coupling multiple damped oscillators to Eq. (2.5) may be beneficial for distinguishing the relative strengths of the damping effects caused by the different layers of the vocal tract. Besides wall vibration losses, explanations of other wave propagation phenomena in the vocal tract remain incomplete. In particular, competing mathematical models have yet to provide adequate explanations for viscothermal losses in the vocal-tract boundary layer [15], voice quality effects arising from subtle modifications of the vocal tract shape [16], or higher-order propagation modes associated with fricative sounds [13].

Improvements to the signal processing algorithms and evaluation tools are also needed. For example, adding an interpolation function between concatenated sounds in speech phrases would create smoother, more natural transitions (as opposed to the abrupt

changes apparent in Fig. 3.7). Moreover, developing a text-to-speech processing tool could improve psychoacoustic evaluation of the intelligibilty, clarity, and expressivity of finite-difference simulations of speech. In the end, physically-modeled synthesis techniques could prove to be competitive alternatives to the current methods of speech synthesis in use today.

Appendix A

Code

The special project submission includes 11 MATLAB files and one audio file. Detailed code from the diphthong with wall losses simulation (Item 7 in A.1) is included in Section A.2.

A.1 Code Explanation

Below is a list of the files included in the submission, as well as descriptions of each of the files.

(1) **vowel_McKell.m** — Finite-difference simulation of a vowel. This script models glottal excitation, vocal-tract sound propagation, and radiation from the mouth.

(2) **vowel_Func_McKell.m** — Same as (1) except this file is a function. It is called by (10).

(3) **diphthong_McKell.m** — Finite-difference simulation of a diphthong. This script models glottal excitation, vocal-tract sound propagation, and radiation from the mouth for two vowels. The vowels are joined using a transition function.

(4) **diphthong_Func_McKell.m** — Same as (3) except this file is a function. It is called by (10).

(5) **vowelLoss_McKell.m** — Finite-difference simulation of a vowel with wall loss. This script models glottal excitation, vocal-tract sound propagation with wall vibration, and radiation from the mouth.

(6) **vowelLoss_Func_McKell.m** — Same as (5) except this file is a function. It is called by (9) and (10).

(7) **diphthongLoss_McKell.m** — Finite-difference simulation of a diphthong with wall losses. This script models glottal excitation, vocal-tract sound propagation

with wall vibration, and radiation from the mouth for two vowels. The vowels are joined using a transition function. See Section A.2.

(8) **diphthongLoss_Func_McKell.m** — Same as (7) except this file is a function. It is called by (10).

(9) **formant_McKell.m** — This script computes the formant structure of a vowel using linear predictive coding.

(10) **phrase_McKell.m** — This script synthesizes two phrases using finite-difference simulations of speech—one with wall vibration losses and one without. The following English sentence is used: "I owe you a yo-yo."

(11) **phrase_Rec_McKell.wav** — This is a 16-bit audio recording of the phrases synthesized in (10).

(12) **Read_Me.m** — An accompanying file included in the code submission that contains the list of file descriptions shown above.

A.2 Diphthong with Wall Losses

```
Filename: diphthongLoss_McKell.m
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%
% Variable Preamble
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%

% Set variables
fs = 44100; % sampling frequency [samples/sec]
f0 = 100; % fundamental frequency [Hz]
c = 340; % speed of sound [meter/sec]
L = 0.1746; % length of vocal tract [meter] (/open_a/)
T = 0.3; % total duration of simulation [sec]
k = 1/fs; % sampling duration [sec]
NF = floor(T*fs); % total length of simulation [samples]
d = c*k; % distance sound travels between samples ('sample distance') [meter]
h = d/L; % fraction of tube length represented by one 'sample distance'
N = floor(L/d); % number of complete 'sample distances' in tube
h = 1/N; % 'h' increased so that 'sample distances' fit evenly in tube
lambda = (d/L)*(1/h); % courant no. (ratio of old 'h' to new/bigger 'h')
gamma = c/L; % time required for sound to travel length L [sec]
p = 0.01; % scaling factor for surface areas

% /open_a/
Sa = [0 0.45;1 0.20;2 0.26;3 0.21;4 0.32;5 0.30;6 0.33;...
    7 1.05;8 1.12;9 0.85;10 0.63;11 0.39;12 0.26;13 0.28;...
    14 0.23;15 0.32;16 0.29;17 0.28;18 0.40;19 0.66;20 1.20;...
    21 1.05;22 1.62;23 2.09;24 2.56;25 2.78;26 2.86;27 3.02;...
```

```
    28 3.75;29 4.60;30 5.09;31 6.02;32 6.55;33 6.29;34 6.27;35 5.94;...
    36 5.28;37 4.70;38 3.87;39 4.13;40 4.25;41 4.27;42 4.69;43 5.03];

% Normalize first column of Sa
Sa(:,1) = Sa(:,1)/max(Sa(:,1));

% Compute interpolated values of the function Sa(:,2)(Sa(:,1))
Sa = interp1(Sa(:,1), Sa(:,2), 0:h:1)';

% Scale surface areas of Sa
Sa = p*Sa;

% Set surface areas of tube slices
% /i/
Si = [0 0.33;1 0.30;2 0.36;3 0.34;4 0.68;5 0.50;6 2.43;...
    7 3.15;8 2.66;9 2.49;10 3.39;11 3.80;12 3.78;13 4.35;...
    14 4.50;15 4.43;16 4.68;17 4.52;18 4.15;19 4.09;20 3.51;...
    21 2.95;22 2.03;23 1.66;24 1.38;25 1.05;26 0.60;27 0.35;...
    28 0.32;29 0.12;30 0.10;31 0.16;32 0.25;33 0.24;34 0.38;35 0.28;...
    36 0.36;37 0.65;38 1.58;39 2.05;40 2.01;41 1.58];

% Normalize first column of Si
Si(:,1) = Si(:,1)/max(Si(:,1));

% Compute interpolated values of the function Si(:,2)(Si(:,1))
Si = interp1(Si(:,1), Si(:,2), 0:h:1)';

% Scale surface areas of Si
Si = p*Si;

% Initialize time-varying cross-sectional areas
S = Si*0;

%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%
% Define damping oscillator parameters
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%
omega0 = 500; % fundamental frequency of vocal tract walls
sigma0 = 405000; % damping coefficient
rho = 1.225; % density of air [kg/m^3]
M = 4.76; % mass per unit area of vocal tract walls [kg/m^2] (Titze,1988)

%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%
% Define transition function 'beta'
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%

% Set transition parameters
dur = 0.08; % duration of transition [sec]
durint = dur*fs; % duration in samples
```

```
startT = floor(NF/2 - durint/2); % start of transition
endT = floor(NF/2 + durint/2); % end of transition

% Initialize transition function
beta = ones(NF, 1);

% Set transition window
beta(startT:endT-1) = 0.5*(1 - cos(pi*(0:durint-1)/durint));

% Set beginning state
beta(1:startT) = 0;

%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%
% Define input impulse train
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%
t = 0:k:T; % time bins
uin = sin(2*pi*f0*t); % sine wave with fundamental frequency f0
uin = 0.5*(uin+abs(uin)); % convert negative sinusoidal values to zeros

%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%
% Compute output signal
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%

% Initialize output sound signal
out = zeros(NF,1);

% Initialize velocity potential vectors. N+1 is last slice
psi = zeros(N+1,1); % velocity potential (n+1 time step)
psi1 = zeros(N+1,1); % velocity potential (n time step)
psi2 = zeros(N+1,1); % velocity potential (n-1 time step)

% Initialize coupling vectors (N-1 is the last slice)
w = zeros(N-1,1); % current time step
w1 = zeros(N-1,1); % 1 time step back
w2 = zeros(N-1,1); % 2 time steps back

for n = 1:NF

    % Calculate time-varying area function
    S = (1 - beta(n))*Sa + beta(n)*Si;

    % Compute coupling coefficient
    eps = c*sqrt(2*rho/M)*(pi/(S(1)))^(1/4);

    % Set average surface area at glottis, in the vocal tract, and at the lips
    Sav = [S(1); 0.25*(S(3:N+1)+2*S(2:N)+S(1:N-1)); S(N+1)];

    % Calculate coefficients of glottal excitation component
```

```
g0 = 2*(1-lambda^2);
gr = 2*lambda^2;
gx = (k^2*gamma^2/h/S(1))*(3*S(1)-S(2));

% Calculate coefficients of vocal-tract propagation component
A = 1/k^2 + sigma0/k;
B = eps*S(2:N).^(0.25)*k./(2*Sav(2:N));
C = 2/k^2 - omega0^2;
D = -2/k^2;
E = eps*S(2:N).^(0.25)/(2*k);
F = sigma0/k - 1/k^2;
sl = 0.5*lambda^2*(S(2:N)+S(1:N-1))./Sav(2:N);
s01 = g0;
s02 = B.*E/A - 1;
w01 = B*C/A;
w02 = B*D/A;
sr = 0.5*lambda^2*(S(3:N+1)+S(2:N))./Sav(2:N);

% Calculate coefficients for lip radiation component
alf1 = 1/(2*0.8216^2*gamma);
alf2 = L/(0.8216*sqrt(S(1)*S(N+1)/pi));
a = 0.5*lambda^2*h*(3*S(N+1)-S(N))/S(N+1);
q1 = alf1*a/k;
q2 = alf2*a;
rl = gr/(1+q1+q2);
r0 = (q1-q2-1)/(1+q1+q2);
r = g0/(1+q1+q2);

% Calculate coefficients for coupling component
w03 = C/A;
w04 = F/A;
wp = E/A;

% Calculate velocity potential at the glottis
psi(1) = gr*psi1(2) + gx*uin(n) + g0*psi1(1) - psi2(1);

% Calculate velocity potential in the vocal tract
psi(2:N) = s01*psi1(2:N) + s02.*psi2(2:N) + sl.*psi1(1:N-1) ...
            + sr.*psi1(3:N+1) - w01.*w1 - w02.*w2;

% Calculate velocity potential at the lips
psi(N+1) = r*psi1(N+1) + rl*psi1(N) + r0*psi2(N+1);

% Calculate pressure at the lips
out(n) = (fs/gamma)*(psi(N+1) - psi1(N+1));

% Compute coupling term
w = w03*w1 + w04*w2 + wp.*(psi(2:N) - psi2(2:N));
```

```
    % Set values equal to next grid line in time
    psi2 = psi1;
    psi1 = psi;
    w2 = w1;
    w1 = w;
end
```

Appendix B

Project Proposal

B.1 Introduction

Articulatory speech synthesis is the artificial generation of human speech achieved by numerical simulation of a physical speech model. Compared to other methods of speech synthesis, such as concatenative synthesis or statistical parametric synthesis, articulatory synthesis appears to be more promising for improving vocal expressivity [2], synthesizer control and manipulation [3], articulatory-driven facial animation [4], and speech science pedagogy [5].

A more refined physical speech model paired with a low-cost numerical simulation scheme is required in order to advance articulatory synthesis technology. In particular, a description of sound propagation along the vocal tract is one aspect of the physical model that remains incomplete. Vocal-tract sound propagation is often described by Webster's equation, a lossless form of the linearlized Navier-Stokes equation. However, Webster's equation does not account for several crucial phenemena, including wave propagation losses. Moreover, competing mathematical models have yet to provide adequate explanations for viscothermal losses in the vocal-tract boundary material [15], voice quality effects arising from subtle modifications of the vocal tract shape [16], or higher-order propagation modes associated with fricative sounds [13]. Further theoretical work is required in order to model these effects.

The first physically-modeled numerical simulation of vocal-tract sound propagation, known as the Kelly-Lochbaum (KL) method, discretized the solution to Webster's equation [6]. In the KL method, the vocal tract was represented as a one-dimensional acoustic tube consisting of a series of concatenated cylindrical sections. The volume velocity and pressure of an acoustic signal moving along the tube were computed using digital delay lines and scattering methods [7]. Several enhancements to the KL method have been developed since its emergence. These enhancements have included the addition of conical tube sections [17, 18], elliptical vocal-tract cross-sections [19], a non-quantized vocal-tract length for continous length changes [20], and a time-varying cross-sectional area function [20]. Alternative numerical simulations based on other mathematical

models have successfully synthesized other wave propagation phenomena in the vocal tract, including resonances and anti-resonances caused by higher-order modes [13] and viscothermal losses [14].

B.2 Objectives

The starting point of any study on articulatory speech synthesis must be a review of the competing physical speech models and the numerical simulations that seek to discretize them. Furthermore, in order to gain insight about the simulation techniques, it is crucial to program each method by hand. This special project focuses on the numerical simulations of vocal-tract sound propagation and aims to accomplish the following specific objectives:

(1) review the mathematical descriptions of vocal-tract sound propagation discussed in the literature, including Webster's equation;

(2) implement the corresponding numerical simulations of vocal-tract sound propagation using the MATLAB programming language, including the KL method and its enhancements;

(3) discuss the limitations and potential improvements relating to the physical models and numerical simulations that were studied.

Emphasis for this special project is given to the review of Webster's equation and the implementation of the KL method. To a lesser extent, the project seeks to review some of the competing models and numerical simulations of vocal-tract sound propagation, including viscothermal loss models, finite-element methods (FEM) and finite-difference methods (FDM), and the multimodal propagation theory and simulation method. Finally, the project aims to lay the groundwork for an adequate theoretical explanation of viscothermal losses in the vocal-tract boundary material.

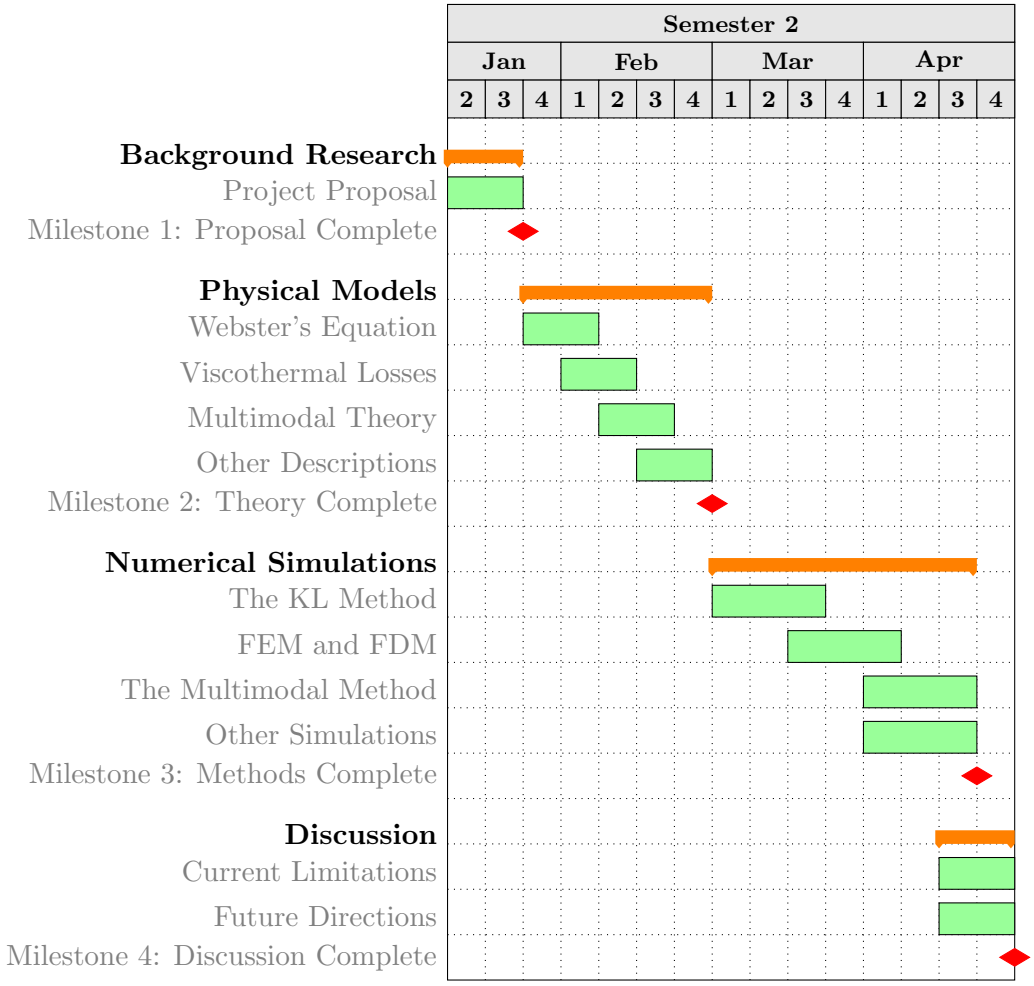
B.3 Equipment

The numerical simulations for this special project will be carried out on a MacBook Pro notebook computer (Retina, 13-inch, Mid 2014) running OS X El Capitan. The relevant hardware specifications are listed below:

Processor	2.8 GHz Intel Core i5
Memory	8 GB 1600 MHz DDR3
Number of Cores	2

B.4 Plan

To achieve the objectives outlined in Section B.2, the following timeline of study topics and key milestones is proposed. All research will be supervised by Stefan Bilbao of the Acoustics and Audio Group at the University of Edinburgh.



Bibliography

[1] W. Kempelen, H. Füger, and J. G. Mansfeld, *Wolfgangs von Kempelen kk wirklichen Hofraths Mechanismus der menschlichen Sprache: nebst der Beschreibung seiner sprechenden Maschine: mit XXVII Kupertafeln*. Bei JB Degen, 1791.

[2] C. H. Shadle and R. I. Damper, "Prospects for articulatory synthesis: A position paper," 2002.

[3] P. Birkholz, D. Jackèl, and B. J. Kroger, "Construction and control of a three-dimensional vocal tract model," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006.

[4] P. Badin, P. Borel, G. Bailly, L. Revéret, M. Baciu, and C. Segebarth, "Towards an audiovisual virtual talking head: 3d articulatory modeling of tongue, lips and face based on mri and video images," in *5th Speech Production Seminar*, pp. 261–264, 2000.

[5] A. J. Teixeira, R. Martinez, L. N. Silva, L. M. Jesus, J. C. Príncipe, and F. A. Vaz, "Simulation of human speech production applied to the study and synthesis of european portuguese," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 1435–1448, 2005.

[6] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," 1962.

[7] S. D. Bilbao, *Wave and scattering methods for the numerical integration of partial differential equations*. PhD thesis, Stanford University, 2003.

[8] S. Bilbao, *Numerical sound synthesis: finite difference schemes and simulation in musical acoustics*. John Wiley & Sons, 2009.

[9] S. M. Lulich and H. Arsikere, "Tracheo-bronchial soft tissue and cartilage resonances in the subglottal acoustic input impedance," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. 3436–3446, 2015.

[10] M. R. Portnoff, *A quasi-one-dimensional digital simulation for the time-varying vocal tract.* PhD thesis, Massachusetts Institute of Technology, 1973.

[11] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 537–554, 1996.

[12] U. C. L. A. Phonetics, "http://www.phonetics.ucla.edu/course/chapter1/wells/wells.html," Accessed on 19-04-2017.

[13] R. Blandin, M. Arnela, R. Laboissière, X. Pelorson, O. Guasch, A. V. Hirtum, and X. Laval, "Effects of higher order propagation modes in vocal tract like geometries," *The Journal of the Acoustical Society of America*, vol. 137, no. 2, pp. 832–843, 2015.

[14] S. C. Thompson, T. B. Gabrielson, and D. M. Warren, "Analog model for thermoviscous propagation in a cylindrical tube," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 585–590, 2014.

[15] I. R. Titze, A. Palaparthi, and S. L. Smith, "Benchmarks for time-domain simulation of sound propagation in soft-walled airways: Steady configurations," *The Journal of the Acoustical Society of America*, vol. 136, no. 6, pp. 3249–3261, 2014.

[16] B. Story, "The vocal tract in singing," in *The Oxford Handbook of Singing*, 2011.

[17] V. Välimäki and M. Karjalainen, "Improving the kelly-lochbaum vocal tract model using conical tube sections and fractional delay filtering techniques.," in *ICSLP*, 1994.

[18] N. Amir, U. Shimony, and G. Rosenhouse, "A discrete model for tubular acoustic systems with varying cross section–the direct and inverse problems. part 1: Theory," *Acta Acustica united with Acustica*, vol. 81, no. 5, pp. 450–462, 1995.

[19] M. Arnela and O. Guasch, "Finite element computation of elliptical vocal tract impedances using the two-microphone transfer function method," *The Journal of the Acoustical Society of America*, vol. 133, no. 6, pp. 4197–4209, 2013.

[20] K. van den Doel and U. M. Ascher, "Real-time numerical solution of webster's equation on a nonuniform grid," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1163–1172, 2008.