Modeling the Liquid, Nasal, AND Vowel Transitions OF North American English Using Linear Predictive Filters and Line Spectral Frequency Interpolations for Use in a Speech Synthesis System

Jacob G. Nieveen

Follow this and additional works at: https://digitalcommons.usu.edu/gradreports

Part of the Electrical and Computer Engineering Commons

Recommended Citation
MODELING THE LIQUID, NASAL, AND VOWEL TRANSITIONS OF NORTH AMERICAN ENGLISH USING LINEAR PREDICTIVE FILTERS AND LINE SPECTRAL FREQUENCY INTERPOLATIONS FOR USE IN A SPEECH SYNTHESIS SYSTEM

by

Jacob G. Nievven

A report submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE in

Electrical Engineering

Approved:

Dr. Jacob H. Gunther
Major Professor

Dr. Todd K. Moon
Committee Member

Dr. Jeffery B. Larsen
Committee Member

UTAH STATE UNIVERSITY
Logan, Utah
2013
Abstract

Modeling the Liquid, Nasal, and Vowel Transitions of North American English Using Linear Predictive Filters and Line Spectral Frequency Interpolations for Use in a Speech Synthesis System

by

Jacob G. Nieveen, Master of Science
Utah State University, 2013

Major Professor: Dr. Jacob H. Gunther
Department: Electrical and Computer Engineering

A speech synthesis system with an original user interface is being developed. In contrast to most modern synthesizers, this system is not text to speech (TTS). This system allows the user to control vowels, vowel transitions, and consonant sounds through a simple 2-d vowel pad and consonant buttons. In this system, a synthesized glottal waveform is passed through vowel filters to create vowel sounds. Several filters were calculated from recordings of vowels using linear predictive coding (LPC). The rest of the vowels in the North American English vowel space were found using interpolation techniques with line spectral frequencies (LSF). The effectiveness and naturalness of the speech created from transitions between these filters was tested.

In addition to the vowel filters, filters for nasal and liquid consonants were found using LPC analysis. Transition filters between these consonants and vowels were determined using LSFs. These transitions were tested as well.
Public Abstract

Modeling the Liquid, Nasal, and Vowel Transitions of North American English Using Linear Predictive Filters and Line Spectral Frequency Interpolations for Use in a Speech Synthesis System

by

Jacob G. Nieveen, Master of Science
Utah State University, 2013

Major Professor: Dr. Jacob H. Gunther
Department: Electrical and Computer Engineering

A speech synthesis system with an original user interface is being developed. In contrast to most modern synthesizers, this system is not text to speech (TTS). This system allows the user to control vowels, vowel transitions, and consonant sounds through a simple 2-d vowel pad and consonant buttons. In this system, vowel sounds are created by applying special processing techniques to a synthesized signal. These techniques have parameters that can be derived from recorded speech. Parameters corresponding to a few vowels were found from speech using linear predictive coding (LPC). The parameters corresponding to the rest of the vowels were found using techniques that can “mix” parameters of different vowels. Line spectral frequencies (LSFs) were used to mix the parameters. The effectiveness and naturalness of the speech created from these parameters was tested.

In addition to the vowel parameters, parameters for certain consonants were found using LPC. Additional parameters for transitions were found using LSFs. These transitions were tested as well.
To Sarah, because you are awesome.
Acknowledgments

Thanks to Dr. Gunther for his incredible helpfulness and inspiring enthusiasm. It has been a pleasure. Also, thanks to Brandon Graham, who has been a really good research and project partner. I hope I get to work with you in the future, Brandon. We will do great things.

Additionally, I would like to thank David G. Sant whose gift to the university allowed for the construction of the Sant Building. Much of the development of this project has occurred inside its walls.

Jacob G. Nieveen
# Contents

<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Public Abstract</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td>List of Figures</td>
<td>x</td>
</tr>
<tr>
<td>Acronyms</td>
<td>xii</td>
</tr>
<tr>
<td>1 Introduction and Background</td>
<td>1</td>
</tr>
<tr>
<td>1.1 The Direct Input Digital Speech Synthesis System</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Vowels and Periodic Consonants</td>
<td>1</td>
</tr>
<tr>
<td>1.3 Tube Model and LPC</td>
<td>2</td>
</tr>
<tr>
<td>1.4 Interpolation Using Line Spectral Frequencies</td>
<td>2</td>
</tr>
<tr>
<td>1.5 Hardware</td>
<td>4</td>
</tr>
<tr>
<td>1.5.1 User Interface Hardware</td>
<td>4</td>
</tr>
<tr>
<td>1.5.2 Microcontroller and USB</td>
<td>5</td>
</tr>
<tr>
<td>1.5.3 PC System</td>
<td>5</td>
</tr>
<tr>
<td>2 Pad Layout and Interpolation</td>
<td>8</td>
</tr>
<tr>
<td>2.1 Overview</td>
<td>8</td>
</tr>
<tr>
<td>2.2 Classification of Vowels, Nasals, and Liquids</td>
<td>8</td>
</tr>
<tr>
<td>2.3 Vowel Recordings</td>
<td>9</td>
</tr>
<tr>
<td>2.3.1 LPC Filters</td>
<td>10</td>
</tr>
<tr>
<td>2.3.2 Formant Estimation for Stand-Alone Vowels</td>
<td>11</td>
</tr>
<tr>
<td>2.4 Proposed Final Layout of Pad</td>
<td>12</td>
</tr>
<tr>
<td>2.5 Interpolation Algorithm</td>
<td>12</td>
</tr>
<tr>
<td>2.6 Proposed Layout Findings</td>
<td>15</td>
</tr>
<tr>
<td>2.6.1 /u/ to /æ/ Transition</td>
<td>15</td>
</tr>
<tr>
<td>2.6.2 /u/ to /a/ Transition</td>
<td>17</td>
</tr>
<tr>
<td>2.6.3 /i/ to /æ/ Transition</td>
<td>18</td>
</tr>
<tr>
<td>2.7 Modifications to Proposed Layout Based on Test Results</td>
<td>19</td>
</tr>
<tr>
<td>2.8 Alternative Pad Layout Analysis</td>
<td>19</td>
</tr>
<tr>
<td>2.8.1 Vowel Formant Maps</td>
<td>19</td>
</tr>
<tr>
<td>2.8.2 Vowel Transition Testing</td>
<td>21</td>
</tr>
<tr>
<td>2.9 Subjective Analysis of Interpolated Approximation of Vowels Removed from Layout of Pad</td>
<td>32</td>
</tr>
<tr>
<td>2.9.1 Vowels Absent from Revised Layout</td>
<td>33</td>
</tr>
</tbody>
</table>
2.9.2 Vowels Absent from Test Layout .................................................. 33
2.9.3 Conclusions Regarding the Effectiveness of the Revised and Test Layouts 33

3 Derivation of Liquid and Nasal Consonant Filters and Their Transitions with Vowels ................................................................. 35
  3.1 Derivation of Filters and Stand-Alone Filter Quality ....................... 35
  3.2 Transitions Between Periodic Consonants and Vowels ..................... 35
    3.2.1 Transitions to and from /r/ and /l/ ........................................ 36
    3.2.2 Transitions to and from /m/ and /n .................................... 37

4 Future Work ....................................................................................... 40
  4.1 Pad Interpolation ................................................................. 40
  4.2 Derivation of Nasal Consonant Filters ......................................... 40
  4.3 Transitions Between Nasals, Liquids, and Vowels ......................... 41

5 Conclusion ....................................................................................... 42

References ......................................................................................... 44
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 IPA vowels.</td>
<td>9</td>
</tr>
</tbody>
</table>
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Overview of DIDSS system.</td>
<td>5</td>
</tr>
<tr>
<td>1.2 User interface layout.</td>
<td>6</td>
</tr>
<tr>
<td>1.3 Schematic of user interface hardware.</td>
<td>6</td>
</tr>
<tr>
<td>1.4 Prototype of DIDSS user interface hardware.</td>
<td>7</td>
</tr>
<tr>
<td>2.1 American English vowel chart.</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Transfer function of LPC filter derived from /ae/ recording.</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Formant 1 vs. formant 2 for recorded vowels.</td>
<td>13</td>
</tr>
<tr>
<td>2.4 Proposed layout of vowel pad.</td>
<td>14</td>
</tr>
<tr>
<td>2.5 Pad interpolation process.</td>
<td>16</td>
</tr>
<tr>
<td>2.6 Spectrogram of /u/ to /ae/ transition of the proposed pad layout.</td>
<td>16</td>
</tr>
<tr>
<td>2.7 Spectrogram of spoken /u/ to /ae/ transition.</td>
<td>17</td>
</tr>
<tr>
<td>2.8 Spectrogram of /u/ to /a/ transition of the proposed pad layout.</td>
<td>17</td>
</tr>
<tr>
<td>2.9 Spectrogram of spoken /u/ to /a/ transition.</td>
<td>18</td>
</tr>
<tr>
<td>2.10 Spectrogram of /i/ to /æ/ transition of the proposed pad layout.</td>
<td>18</td>
</tr>
<tr>
<td>2.11 Spectrogram of spoken /i/ to /æ/ transition.</td>
<td>19</td>
</tr>
<tr>
<td>2.12 Alternate to the original proposed layout.</td>
<td>20</td>
</tr>
<tr>
<td>2.13 Layout based on only four corner vowels.</td>
<td>21</td>
</tr>
<tr>
<td>2.14 Contour map of first formant of revised layout.</td>
<td>22</td>
</tr>
<tr>
<td>2.15 Contour map of second formant of revised layout.</td>
<td>23</td>
</tr>
<tr>
<td>2.16 Contour map of first formant of test layout.</td>
<td>24</td>
</tr>
<tr>
<td>2.17 Contour map of second formant of test layout.</td>
<td>25</td>
</tr>
</tbody>
</table>
2.18 Spectrogram of /u/ to /æ/ transition using the revised pad layout. . . . . . 25
2.19 Spectrogram of /u/ to /æ/ transition using the test pad layout. . . . . . . 26
2.20 Spectrogram of /u/ to /a/ transition using the revised pad layout. . . . . 26
2.21 Spectrogram of /u/ to /a/ transition using the test pad layout. . . . . . . 26
2.22 Spectrogram of /i/ to /æ/ transition using the revised pad layout. . . . . 27
2.23 Spectrogram of /i/ to /æ/ transition using the test pad layout. . . . . . . 27
2.24 Spectrogram of /a/ to /i/ transition using the revised pad layout. . . . . 28
2.25 Spectrogram of /a/ to /i/ transition using the test pad layout. . . . . . . 28
2.26 Spectrogram of spoken /a/ to /i/ transition. . . . . . . . . . . . . . . . . . . 28
2.27 Spectrogram of /u/ to /i/ transition using the revised pad layout. . . . . 29
2.28 Spectrogram of /u/ to /i/ transition using the test pad layout. . . . . . . 30
2.29 Spectrogram of spoken /u/ to /i/ transition. . . . . . . . . . . . . . . . . . . 30
2.30 Spectrogram of /a/ to /æ/ transition using the revised pad layout. . . . . 30
2.31 Spectrogram of /a/ to /æ/ transition using the test pad layout. . . . . . . 31
2.32 Spectrogram of spoken /a/ to /æ/ transition. . . . . . . . . . . . . . . . . . . 31
3.1 Spectrogram of synthesized /r/ to /a/ transition. . . . . . . . . . . . . . . . 36
3.2 Spectrogram of spoken /r/ to /a/ transition. . . . . . . . . . . . . . . . . . . 37
3.3 Noise present at filter transition in synthesized /r/ to /a/ transition. . . . . 38
3.4 Plot of clean /r/ to /a/ transition waveform. . . . . . . . . . . . . . . . . . . 38
3.5 Noise present at filter transition in synthesized /r/ to /u/ transition. . . . . 39
# Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTS</td>
<td>Text to Speech</td>
</tr>
<tr>
<td>DIDSS</td>
<td>Direct-Input Digital Speech Synthesis</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear Predictive Coding</td>
</tr>
<tr>
<td>LP</td>
<td>Linear Prediction</td>
</tr>
<tr>
<td>LSP</td>
<td>Line Spectral Pairs</td>
</tr>
<tr>
<td>LSF</td>
<td>Line Spectral Frequency</td>
</tr>
<tr>
<td>CELP</td>
<td>Code Excited Linear Prediction</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction and Background

1.1 The Direct Input Digital Speech Synthesis System

Most modern speech synthesizers are text to speech (TTS) devices. Many of these devices are able to produce high-quality, natural sounding speech. However, a main detriment to the naturalness of their synthesized speech is the inability of TTS synthesizers to determine natural duration, volume, and pitch, i.e. prosody, from text. This report gives details regarding the Direct Input Digital Speech Synthesis (DIDSS) system, a synthesizer created to give the user control over prosody of synthesized speech. It does this by allowing a user to string together speech phones while controlling pitch and volume of the synthesized speech.

This report is focused on a small portion of the DIDSS system, as explained in the following section. The rest of this chapter discusses basics of certain speech synthesis techniques and gives background for the DIDSS system.

1.2 Vowels and Periodic Consonants

This report focuses mainly on the implementation of vowels within the DIDSS device. The user of the DIDSS device is given the ability to completely control vowels and vowel transitions of synthesized speech through a 2-d pad. This pad allows for the creation of any vowel sound in North American English, and any transition from one vowel to another that occurs naturally in speech. A good example of one such vowel transition is the word “eye.” To say “eye,” a person first pronounces the sound “ah,” and transitions to the sound “i.”

In addition to vowels, this report gives details regarding the implementation of certain periodic consonants and their transitions with vowels.
1.3 Tube Model and LPC

Vocalized speech is created when the vocal tract shapes the sound produced by the vocal folds. This buzzing sound is created by the periodic opening and closing of the vocal folds. It is sometimes called the glottal source. Different types of speech sounds are produced by changing the configuration of the vocal tract. In speech synthesis, the vocal tract is often modeled as a filter used to shape a synthesized glottal waveform. In this way, sounds similar to speech sounds can be created. A common means to extract filters from recorded speech is called Linear Predictive Coding (LPC).

LPC uses a common model for speech synthesis called the source/filter or the tube model. This model treats the vocal tract as a series of tubes of different lengths and diameters that filter the sound generated by the glottal source. More information about this model can be found in an early work from Gunnar Fant [1]. LPC is a method that determines the coefficients of an all-pole filter that characterize the vocal tract. LPC coefficients can be found from recorded speech. A synthesized glottal source can then be generated and filtered using these coefficients. The resulting synthesized waveform is very similar to the original recorded speech. LPC is often used in low bitrates, such as 8K or 16K samples per second, because it produces high quality synthesized speech even at those bit rates. More information regarding LPC analysis can be found in an early article by B.S. Atal and S.L. Hanauer [2].

1.4 Interpolation Using Line Spectral Frequencies

Speech is often encoded, transmitted, and decoded because the encoded speech contains much less data than the original speech. LPC coefficients, however, are very sensitive to noise and can easily become corrupted along a transmission line. Because of this, LPCs are often converted to Line Spectral Pairs (LSPs) before transmission. This is done because LSPs are much more resilient to channel noise [3].
LSPs can be converted to and from LPCs. The basis for this comes from the all-pole nature of the LPC filter. A p-order LP polynomial representation is given as

\[ A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_p z^{-p}. \]  

(1.1)

This polynomial can be decomposed into two separate polynomials, P and Q, given as

\[ P(z) = A(z) + z^{-(p+1)} A(z^{-1}), \]  

(1.2)

\[ Q(z) = A(z) - z^{-(p+1)} A(z^{-1}). \]  

(1.3)

The original LP polynomial can be determined from the LSP polynomials in the following manner:

\[ A(z) = .5[P(z) + Q(z)]. \]  

(1.4)

LSPs model the vocal tract as two different tubes, corresponding to the two different polynomials. The P polynomial gives the response of the vocal tract with the glottis closed, and the Q polynomial gives the response when the glottis is open.

The nature of the P and Q polynomials causes all of their roots to lie interleaved on the unit circle, symmetrical across the real axis. Because the poles all lie on the unit circle, the resonant frequencies are not damped. This creates a spectrum with vertical lines at resonances instead of smooth peaks. The presence of these lines is the basis for the name: line spectral pairs [4].

Because the poles all lie on the unit circle, they are often converted to angles. Knowing the angles between vectors pointing to the poles and a vector in the (1,0) direction is as good as knowing the locations of the roots themselves. These angles are called Line Spectral Frequencies (LSF). LSFs and LSPs were introduced in a paper by Itakura [5].

In addition to their robust behavior along transmission channels, LSFs are easily interpolated. In contrast, interpolating LPC coefficients can result in distorted or unstable
filters [6]. Some types of LPC systems that encode and transmit LPC filters require interpolation of these filters. Additionally, some benefits of interpolation of LPC filters can be found in the literature. Research has found that speech reconstructed from linear interpolations of LPC frames is of better subjective quality than that reconstructed from the originally calculated LPC coefficients in certain situations [7].

LSFs are not the only method for interpolating between LPC filter values, in fact, many other methods exist. Other common types of LPC representations used for LPC interpolation are reflection coefficients (RC), log area ratios (LAR), and autocorrelations (AC). A study by Islam and Kabal concludes that of all of these, LSFs show the least spectral distortion during interpolation [8]. It is for this reason that they are used in this project.

LSFs are used in the DIDSS project as a means to create transition filters from one given vowel filter to another on the vowel pad.

1.5 Hardware

As part of the DIDSS project, a hardware user interface was designed and prototyped. This section describes the user interface, its hardware, and the details of its connection with a PC.

Figure 1.1 gives a complete overview of the user interface, hardware, and PC system described in this section.

1.5.1 User Interface Hardware

The DIDSS system uses a USB device to provide a user interface. The layout of the user interface is shown as Figure 1.2.

As can be seen in the figure, there are two pads and three rows of buttons. These are actual interfaces provided to the user by physical devices. The pads are provided by resistive pads, and the buttons are provided by three resistive strips. The strips are divided into sections corresponding to the separate consonant sounds.
1.5.2 Microcontroller and USB

All of the user input devices are connected to the ADC inputs of an ATMEGA32U4. This microcontroller polls these inputs and sends data to the computer via USB when the computer requests, using the Raw Human Interface Device (Raw HID) protocol. A schematic of the hardware is shown as Figure 1.3. More information about the Teensy, the microcontroller board used in this project, is provided by PJRC [9]. This information includes details regarding the pin numbers shown in the schematic.

A prototype of the hardware was built and is shown in Figure 1.4. With the future addition of labels to the hardware, the prototype’s layout is very similar to the proposed layout in Figure 1.2.

1.5.3 PC System

A PC computer is used to communicate with the user interface device and output audio as necessary. The PC uses Raw HID to communicate with the ATMEGA microcontroller. It also uses PortAudio running inside a Windows environment to use the sound hardware.
Fig. 1.2: User interface layout.

Fig. 1.3: Schematic of user interface hardware.
Fig. 1.4: Prototype of DIDSS user interface hardware.
Chapter 2
Pad Layout and Interpolation

2.1 Overview

This chapter gives background needed to understand the layout of vowels devised for the vowel pad. It also gives details regarding a few different tested layouts and the resulting synthesized speech from each of these layouts.

2.2 Classification of Vowels, Nasals, and Liquids

Vowels are made when airflow in the mouth is mostly unconstricted. Different shapes of the vocal tract create different vowel sounds. Nasals are consonant sounds created when the nasal cavity is used as the main resonant structure for sound creation rather than the mouth. Liquids are made by partially blocking the airflow of the mouth with the tongue. The two nasals are /m/ and /n/ and the two liquids are /l/ and /r/. In contrast to other consonants, these consonants are periodic and can be held indefinitely.

Each individual language has a different vowel system. The DIDSS project’s aim is to synthesize general North American English speech. Because of this, the project only uses a subset of all possible vowel sounds. For this report, International Phonetic Alphabet (IPA) symbols are to represent these vowel sounds. Table 2.1 below shows each of these symbols, with its respective sound. For more information regarding the IPA system, refer to the IPA handbook [10].

A common classification for vowels is by their height, which is somewhat related to the position of the tongue in the mouth when they are being produced. Higher vowels are produced with the tongue nearer the front of the mouth, and lower vowels are produced with the tongue nearer the back of the mouth. Figure 2.1 gives the relative layout of American
English vowels. These types of charts are very common in phonetics research. More discussion about these charts can be found in Ladefoged’s authoritative book on Phonetics [11]. The symbols used to represent the vowels in the chart follow IPA convention.

Vowels are also frequently characterized by the location of their formants, or in other words, the frequencies at which the resonances occur in the vocal tract. It is quite common to compare formant 1 and formant 2 for each of the separate vowels. This characterization is discussed more deeply later in this report.

### 2.3 Vowel Recordings

Eleven vowels spoken by a male voice were recorded at a sampling rate of 44.100 kHz.
for roughly two seconds each. Every vowel in Table 2.1 was recorded. These recordings were analyzed and tested using LPC analysis. Transitions between vowels were also recorded and analyzed. The filters derived from these recordings were used both directly in the vowel pad and to derive additional filters for the vowel pad.

2.3.1 LPC Filters

LPC filters were made from the recorded vowels. These filters were calculated by finding the least-squares prediction error. A Hamming window of 1500 samples, or approximately 34 ms was applied to a location where the waveform was in steady-state. This window reduced the possible effects of discontinuities at the edge of the windowed waveform, which could result in inaccurate data filters. This window is slightly longer than a standard window length for LPC, which is generally 20-25 ms [12]. LPC windows are usually made short in order to account for rapid changes in speech waveforms that occur when certain aperiodic consonant sounds such as /t/ or /ch/ are created. Vowels vary less rapidly, which allows for a larger window length. This slightly larger window length creates a better defined filter.

A sampling rate of 44.1 kHz is higher than the 8 or 16 kHz that is often used for LPC. Speech sounds have low power in frequency content above 8 kHz, partially because the glottal waveform has an approximately -12 dB/octave dampening effect on the power spectrum of speech [4]. A main motivation behind using the higher sampling rate in the DIDSS project is found in possible future uses of the vowel pad’s processing capabilities. Music and sounds from musical instruments do not have the same attenuation of higher frequencies as speech does. When music or instruments are used as an input to the LPC filters in place of a glottal source, musically interesting sounds can result.

A side effect of the higher sampling rate is that a much higher order LPC filter is needed to achieve satisfactory results. It is common to find an 8-12 order LPC filter in use with a sampling rate of 8 kHz. In contrast, a good filter order for a 44.1 kHz sampling rate was subjectively determined to be approximately 60. At this order, there are no noticeable sound quality issues. This order is somewhat higher than Taylor’s rough estimate that there should be one pole for each 1000 Hz of sampling frequency plus an additional 4-6 [4]. At
order 60, there are 44 poles, one for each 1000 Hz, plus an additional 16. Since LPC is usually used in low sampling rate applications, it is very likely that Taylor’s rule-of-thumb was not intended to be used around 44.1 kHz. This would explain why 16 extra poles were needed instead of 6.

2.3.2 Formant Estimation for Stand-Alone Vowels

Formants for particular vowels vary greatly between specific speakers, and from recording to recording. Therefore, in order to determine the proper layout for the vowel pad, formants were estimated for each of the ten recorded vowels. A common method for determining formant locations from LPC data is to use the frequency response of the filters. The location of peaks in the frequency response corresponds to the frequencies of formants of the vocal tract. Figure 2.2 shows the transfer function of the filter derived from the vowel /ae/, as determined from LPC coefficients found using a 34 ms window of the recorded vowel.

As a means to prevent incorrect formant data, many separate windows from the same recorded vowel waveform were used to compute LPC coefficients and find the first two formant locations. The mean of each of these locations was found. For each of the recorded vowels, it was rare to find a specific window within the waveform where the formants varied significantly from the mean, although there was some variance from window to window. A chart showing the location of the first two formants plotted against each other is shown in Figure 2.3. This chart shows great similarity to a chart given by Dew and Jensen [13] in their book on phonetics. This gives validity to the filters derived from these recordings. Some of the formants frequencies are still somewhat different, however. These differences

![Fig. 2.2: Transfer function of LPC filter derived from /ae/ recording.](image-url)
did not affect the proposed layout of the pad. The derivation of this layout is discussed in the next section.

2.4 Proposed Final Layout of Pad

The vowel pad was laid out so that the user could make vowel to vowel transitions as easily as possible. Most transitions can be made by drawing straight lines between the two vowels in question. The layout shown in Figure 2.4 shows the location of the vowels on the pad. Each of these locations corresponds to a filter made using LPC analysis on a voice recording. The filters corresponding to areas between these vowels is found by interpolation.

The layout for the vowel pad was derived from two separate sources: the vowel height chart, and the information about the first two formants of the recorded vowels. These charts and plots are shown as Figure 2.1 and Figure 2.3, respectively. These charts all have very similar layouts. The proposed layout is very similar to the vowel height chart turned on its side. It can also be seen that proper formant transitions can be created by drawing appropriate lines on the vowel pad. The justification for using the formant and vowel height charts to lay out the pad is that for each vowel transition, the formants need to transition from those of the first vowel to those of the second. It is intuitive that the vowels should maintain a similar relative layout to that of the formant chart, to maintain these transitions. The tongue also has to move positions from front to back and high to low or vice versa. The vowel height chart implies transitions that must be made by the tongue between vowels. It is intuitive that the pad layout should maintain these transitions as well.

2.5 Interpolation Algorithm

In order to fulfill the needs of the DIDSS Project, a 200x200 grid of LPC coefficient vectors was calculated from various recordings of vowels. Each location on this grid corresponds to an LPC filter. When the user of the DIDSS system places his finger on a particular location of the vowel pad, a filter that corresponds to that location is used in DIDSS system. In order to achieve smooth transitions from point to point on the grid, an interpolation algorithm was applied.
Fig. 2.3: Formant 1 vs. formant 2 for recorded vowels.

To create the LPC vector grid, a 200x200 cell of 60 order vectors was instantiated in MATLAB. Several points of this grid were instantiated with LPC coefficients found from vowel recordings. Figure 2.4 shows the locations on the grid of each of these vectors.

With these locations initialized, the corresponding LPC vectors were converted to LSFs. Each 60 order LPC vector was converted to a 60 order LSF vector. Every pole of an LPC filter corresponds to two LSP zeros, one for the P polynomial and one for the Q polynomial. However, because the locations of these zeros are symmetric across the real axis of the unit circle, only half of these values were needed.

The first step of the interpolation algorithm requires interpolation along vertical lines of the grid. This interpolation is performed by taking linear combinations of two LSF vectors with the same horizontal coordinates. As an example, consider the LSF vectors corresponding to the vowels /u/ and /i/. This is the first column of vectors in the 200x200 grid. The interpolated values for the LSF vectors are found using the following equation, which uses MATLAB indexing:
\[ \vec{g}(1, n) = \frac{(200 - n)}{199} \vec{i} + \frac{(n - 1)}{199} \vec{u}, \]  

(2.1)

where \( \vec{g}(m, n) \) is a function that gives an LSF vector for any point on the vowel pad’s grid, \( m \) and \( n \) are integer coordinates on the grid, and \( \vec{u} \) and \( \vec{i} \) are the LSF vectors for the /u/ and /i/ vowels. It can be seen from the equation that the interpolation uses simple linear combinations of the vectors in question. The other interpolations are found using similar equations.

The left and right edges of the grid were completed by interpolating between the LSF vectors of the corners. This means that interpolation occurs between the LSFs corresponding to the /u/ and /i/ vowels, and between the LSFs corresponding to the /a/ and /æ/ vowels. Interpolation also occurred between any other already defined vectors which lay at the same horizontal coordinates. In cases where a full vertical line could not be created because of a missing endpoint on the top or bottom, a suitable endpoint was created by interpolating between the two nearest points on the same horizontal line.

Once all of the necessary vertical lines were created, the horizontal gaps are filled in. This was done by interpolating between vertical lines. Each line was interpolated with its
two nearest neighbors to fill in the gap. In the case of the vertical lines on the far left and far right, the interpolation only occurred between those lines and their one nearest neighbor. Figure 2.5 illustrates the full interpolation process.

This method of interpolating the vowel pad values treats vertical transitions differently than horizontal ones. This is because a few vertical transitions are calculated first, then the horizontal transitions are calculated. It can be shown mathematically that slightly different vectors result when calculating the horizontal transitions first. To ameliorate this effect, a second vowel pad was calculated. This pad was calculated beginning with horizontal lines. The horizontal lines were created by interpolating between vowels at the same vertical position, similar to how the vertical lines were created, as described in the paragraphs above. Once the horizontal lines were created, the vertical gaps were calculated.

The final LSF values for each point in the table were found by interpolating between the LSF values at the same point in both tables. In this manner, a table that varies well in both the vertical and horizontal directions was created.

2.6 Proposed Layout Findings

The proposed pad layout, as shown in Figure 2.4, was implemented on the DIDSS device and tested both subjectively and by using spectrograms. The results were mostly positive, but there were a few vowel transitions that did not work properly. These findings are given in the sections below.

2.6.1 /u/ to /æ/ Transition

The DIDDS device was configured to be used with vowel pad filter values that were calculated based on the proposed pad layout. A subjective listening test resulted in the conclusion that the transition from /u/ to /æ/ did not sound natural to human listeners. Figure 2.6 is a spectrogram of a recording of the synthesized transition. Figure 2.7 shows the spectrogram of an actual speech recording of this transition for comparison. This figure was made using a recording of the synthesized vowel transition and the WASP software created by Mark Huckvale of the University College London [14].
Comparison of these figures leads to the conclusion that the synthesized waveform’s higher formants (formant 2 and above) are not behaving properly to create a natural sounding transition. These formants shift up and back down when they should either stay close to constant (formants 3 and above), or shift upwards only (formant 2).

To create this transition on the vowel pad, the user draws a line from the corner representing /u/ to the diagonally opposite line representing /ae/. This line transitions through four different vowels /u/, /ʊ/, /ʌ/, and /æ/. The main problem with this transition arises between the vowels ʊ and ʌ.
2.6.2 /u/ to /a/ Transition

A /u/ to /a/ transition is made when a user draws a straight line across the top of the vowel pad from the /u/ corner to the /a/ corner. The result of a subjective listening test is that this transition sounds natural using the proposed layout, provided that the user draws the horizontal line at the very top of the pad. If the transition is made slightly lower on the pad by drawing a horizontal line approximately 1/8 of the distance from the top, the transition no longer sounds natural. Figure 2.8 shows a spectrogram of this transition. Figure 2.9 shows a spectrogram of a speech recording for comparison.

In a manner similar to the /u/ to /ae/ transition, the higher formants of the /u/ to /a/ transition shift upwards and then downwards, where they should only shift upwards. This is caused by transitions to and from the /ū/ vowel.
2.6.3 /i/ to /æ/ Transition

A subjective test lead to investigation of the /i/ to /æ/ transition on the proposed vowel pad. The test concluded that this transition sounded unnatural. Figure 2.10 is a spectrogram of the recorded synthesized vowel transition. Figure 2.11 is a spectrogram of an actual speech recording of the same transition.

The synthesized waveform’s spectrogram shows a dip in the third and forth formant frequencies that account for its unnatural sound. This dip is not present in the spoken waveform.

To transition from /i/ to /æ/, the user of the DIDSS device draws a straight line along the bottom of the pad. This line corresponds to transitions between four vowels: /i/, /e/, /ɛ/, and /æ/. The slight dip in formant frequencies is a result of the inclusion of the /e/ vowel in the interpolation.

Fig. 2.10: Spectrogram of /i/ to /æ/ transition of the proposed pad layout.
2.7 Modifications to Proposed Layout Based on Test Results

In order to remove the unnaturalness from the tested vowel transitions, a new layout was implemented. This layout is nearly the same as the tested layout, as the only modification is the removal of two vowels: / الماضي / and /e/. By removing these two vowels, the disruption of the natural formant transition was prevented. The new layout is shown as Figure 2.12.

Clearly, an issue with this new layout is the possibility that the two removed vowels are no longer well defined on the vowel pad. However, while implementing this new layout it was expected that approximations to these vowels would exist on the vowel pad. Further sections give details regarding the effectiveness of the interpolated filter values approximating these vowels.

As an additional test, an additional vowel pad filter table was created by interpolating using only the four corner vowels: /u/, /a/, /æ/, and /i/. These filter values were used to test transitions, as well as to test how well-defined the interpolated vowels sound. Figure 2.13 shows this layout.

2.8 Alternative Pad Layout Analysis

2.8.1 Vowel Formant Maps

Contour maps of the first two formants of the vowel tables were created. Figures 2.14 and 2.15 are contour maps of the primary alternative layout, which uses eight vowels to determine the entire vowel space. Figures 2.16 and 2.17 show the contour maps for the test
alternative layout that only uses four vowels.

It is important to note that the maps of the second formant were rotated, in order to make the map more visible. Whereas the first formant maps have the point (1,1) as the bottom left corner, the second formant maps have the point (1,200) as the bottom left corner. This is important to note because the corner corresponding to /i/ is at the bottom left for the first formant maps, but the vowel corresponding to /u/ is at the bottom left for second formant maps.

An initial glance at both sets of formant contours shows that the formants vary more smoothly for the test layout than for the revised layout. This is not surprising since including additional vowels can be seen as including additional intermediary constraints that prevent the smoothest possible interpolation.

It is visible in Figure 2.15 that the second formant changes relatively quickly from the /I/ vowel to the /ʌ/ and /o/ vowels. There is a large change in the shape of the frequency response and second formant frequencies in these filters, so this steep “cliff” was somewhat expected. It should be noted, however, that this cliff is not due entirely to steep changes in formant frequency, but also to some nonlinear behavior of LSF interpolation. This is the biggest anomaly in the first two formant maps for the revised layout. There are other small
Fig. 2.13: Layout based on only four corner vowels.

variations from the very mellow curves given as the formant contours of the test layout, but overall the contour maps are relatively smooth with a uniform gradient.

2.8.2 Vowel Transition Testing

Several vowel transitions were tested using both the revised and test layouts. Each of these tests involved a subjective listening test and analysis of relevant spectrograms to check for anomalies. Six transitions were tested. These six transitions test roughly the entirety of the pad.

/ʊ/ to /æ/ Transition

By removing the /ʊ/ vowel from the original proposed layout, the quality of the /ʊ/ to /æ/ transition was greatly improved. This is evidenced both by a subjective listening test and by spectrograms of the transitions. The listening tests concluded that the synthesizer sounds like it is saying “wa” as in “wham.” Figure 2.18 shows a spectrogram of this transition, as synthesized using the DIDSS device and the revised layout. Figure 2.19 shows a similar spectrogram as synthesized using the test layout.
The fluctuations of the higher formants that was present in the original proposed layout (and seen in Figure 2.6) is not seen in the spectrograms of the transitions for the revised and test layouts. Both these spectrograms also show a shift upwards of the first two formants, something that is seen in the spectrogram of the recorded spoken speech, given as Figure 2.7.

Despite both the synthesized speech samples following the same general trend as the recorded spoken speech, there are still some differences in the spectrograms. The first formant shifts much more in the synthesized speech, as do formants three and above. Because the vowels used in filter interpolation were stand-alone and not a part of any transition, it is not surprising that the formant frequencies do not line up as well as those of the actual spoken speech. Despite this small anomaly in the transition, subjectively the transition sounds natural in both cases.

There is a difference in the way that the first two formants transition relative to one another in the revised and test layouts. The first two formants of the synthesized speech both increase in roughly the same manner in the case of the revised layout. In the case

Fig. 2.14: Contour map of first formant of revised layout.
of the test layout, formant 2 seems to increase before formant 1. This difference can be explained by the inclusion of the vowel /ʌ/, which forces the formants to vary more equally in the case of the revised layout. This difference in the way the formants vary does effect the subjective nature of the transition sound, but both transitions still sound natural. The difference in sound was deemed unimportant.

/u/ to /a/ Transition

Removing the /ʊ/ vowel from the proposed layout also had a positive effect on the transition from /u/ to /a/. A subjective listening test concluded that the transition sounded much more natural, and that the synthesizer sounds like it is saying the word “wah” when using this transition. Figures 2.20 and 2.21 show the spectrograms of the /u/ to /a/ transition for both the revised and test layout.

Both of these spectrograms show an absence of the rising the falling behavior of the higher formants seen in the synthesized waveforms from the original pad layout, as shown
in Figure 2.8. This accounts for the increase of naturalness discovered in the subjective listening test.

The spectrograms corresponding to both the revised and test layouts are nearly identical. The first formant shifts upwards while the second formant stays constant in both of them. The higher formants have extremely similar behavior between both spectrograms. This implies that the recorded /o/ vowel is very close to being along the natural interpolation path between /u/ and /a/.

These spectrograms are also very similar to the spectrogram of the natural recording of the /u/ to /a/ transition given as 2.9. The first formant frequency is increased while the second formant frequencies stay constant. The third and forth formants lose energy at around 2000-3000 Hz in all cases. Those specific resonances seem to disappear in the natural speech recording, but only seem to diminish in both synthesized waveforms. In all cases, a new resonance seems to appear slightly below 4000 Hz. Overall, both synthesized and natural spectrograms are very similar.
/i/ to /æ/ Transition

The removal of the /e/ vowel from the proposed layout increased the naturalness of the synthesized /i/ to /æ/ transition. Both subjective and spectrographic analysis confirm this. When made using the DIDSS device, this transition sounds like the word “yeah.” Figures 2.22 and 2.23 show spectrograms of this transition synthesized using both the revised and test layouts.

A comparison of these two spectrograms with Figure 2.10 shows that the dip present...
in the higher formant frequencies of the transition made by using the original pad layout is not present in the revised and test layouts. The absence of this dip is the result of removing /e/ from the vowel pad layout.

As is the case with many other transitions, these spectrograms are very similar for both the revised and test layouts. They are also very similar to the spectrogram corresponding to the natural speech recording of the transition from /i/ to /æ/, given as Figure 2.10.

Fig. 2.21: Spectrogram of /u/ to /a/ transition using the test pad layout.
In both the synthesized and natural waveforms, the first formant frequency rises as the second formant frequency decreases. The higher frequencies are approximately the same, with variances that are easily accounted for by LPC encoding and vocal tract differences between recordings.

/a/ to /i/ Transition

The transition between the vowels /a/ and /i/ was tested on the revised and test layouts. This transition was made using the DIDSSS device by drawing a straight line diagonally from the upper right corner to the lower left corner. This transition was confirmed to sound natural in a subjective listening test. This transition sounds like the word “eye.” The spectrograms were also analyzed to look for anomalies compared to a spectrogram of a natural speech sample. Figures 2.24 and 2.25 are spectrograms of the synthesized speech. Figure 2.26 is a spectrogram of natural speech of the transition included for comparison.

It is clearly visible in these spectrographs that the synthesized waveforms and the
Fig. 2.24: Spectrogram of /a/ to /i/ transition using the revised pad layout.

Fig. 2.25: Spectrogram of /a/ to /i/ transition using the test pad layout.

natural waveform have similar frequency content. The first two formant frequencies are closely spaced at the beginning of the transition and separate as time goes on. Some of the higher resonant frequencies decrease with a pattern approximately the same as the first formant.

Minor differences are visible between the transitions created from the revised layout and those created by the test layout. There is a very slight variation from the smooth

Fig. 2.26: Spectrogram of spoken /a/ to /i/ transition.
transition of formant frequencies of the test layout seen in the revised layout. This variation is barely noticeable in Figure 2.24. It implies that the /ε/ vowel does not fit perfectly into the transition between /a/ and /i/. This variation is not noticeable in a subject listening test of the transition. It is considered unimportant, although it might be fixed by re-recording the vowel with the speaker’s vocal tract in a very slightly different shape, or by changing the location of the vowel in the layout. Both of these solutions were deemed unnecessary.

/ụ/ to /i/ Transition

A subjective listening test yields the conclusion that the transition between the vowels /ụ/ and /i/ sounds natural. This transition is made by drawing a straight line from the upper left corner to the lower left corner. This transition sounds like the word “we.” Spectrograms of the transitions from the revised layout, the test layout, and a recorded sample of natural speech are given as Figures 2.27, 2.28, and 2.29.

It is intuitively obvious from the revised and test pad layouts that the transitions between /ụ/ and /i/ should be nearly identical for both cases. In both cases, there are no intermediary vowels along the transition line, and the influence of the vowel /I/ for the revised case is minimal.

The spectrograms of the synthesized speech are very similar to the spectrogram of the natural speech. The first formant stays nearly constant and loses power, while the second formant frequency rises. The frequencies of formant 3 and above change very little.

![Fig. 2.27: Spectrogram of /ụ/ to /i/ transition using the revised pad layout.](image-url)
/a/ to /ae/ Transition

The final transition tested was the transition from /a/ to /ae/. This transition is made by drawing a line straight down from the upper right corner to the lower right corner of the vowel pad. In contrast to the other transitions tested, this transition does not sound like any common English word. Figures 2.30, 2.31, and 2.32 show the spectrograms of the two synthesized and one natural speech waveforms.

---

Fig. 2.28: Spectrogram of /u/ to /i/ transition using the test pad layout.

Fig. 2.29: Spectrogram of spoken /u/ to /i/ transition.

Fig. 2.30: Spectrogram of /a/ to /æ/ transition using the revised pad layout.
The spectrograms for the two synthesized waveform layouts are nearly identical. This is because there are no intermediary vowels between /a/ and /ae/ and the nearest vowels /ʌ/ and /ɛ/ have very little effect on the transition.

The spectrograms of the synthesized waveforms and the natural waveforms are similar, with the first two formants running together at fairly close frequencies before the second formant frequency splits off and increases. The first formant frequency increases with the transition in the natural waveform, but does not as much in the synthesized waveforms. There is an increase in the frequency of a resonance around 4000 Hz in both the synthesized and natural waveforms as the sound transitions from /a/ to /ae/. The spectral power between 2000 and 3000 Hz is more constant in the natural waveform. This difference in the waveforms is likely partially due to the imperfect nature of LPC representation of the spectral envelope.

Despite the differences in the spectrographs of the synthesized and natural waveforms, the waveforms sound very similar.
Conclusions Found from Test Transitions

The transitions tested were all very similar to the natural spoken waveform transitions in terms of the first one or two formants. However, it was discovered that anomalous variations in higher formant frequencies such as those shown occurring in Figures 2.8 and 2.10 can affect the naturalness of synthesized speech, and affect the perception of the vowel transition in question. It seems that rapid fluctuations in these higher formants caused the most issues. In contrast, some transitions had third and fourth formant frequencies that varied in nature slightly from those of the natural waveforms, but still did not cause the synthesized speech to sound unnatural. Those transitions in which the behavior of the third and forth formant frequencies was somewhat different than the natural waveform, but that was still steadily increasing, decreasing, or unchanging (rather than increasing then quickly decreasing, or vice versa) sounded natural.

In all cases, the interpolation at various points between two distinct vowel filters followed an extremely similar pattern to the transitions found in actual recorded speech. Interpolation between LSFs is capable of creating very effective vowel transitions.

The removal of the vowels /_PKG/ and /e/ caused several transitions made using the vowel pad to sound more natural. It is expected that some form of these removed vowels can be fit in the vowel pad, since sounds that approximate them can be made using the revised layout. By making new recordings of these vowels with a slightly different vocal tract shape, equivalent vowels might be found that would fit into the layout better.

An analysis of the sounds approximating the removed vowels is found in Section 2.9.

2.9 Subjective Analysis of Interpolated Approximation of Vowels Removed from Layout of Pad

In order to justify the complete removal of the vowels /_PKG/ and /e/ from the layout of the vowel pad, the revised pad was subjectively tested to find the closest approximations to these sounds. The test layout was also tested to find the closest approximations to all vowels not used in the interpolation.
2.9.1 Vowels Absent from Revised Layout

The vowel sound corresponding to /Ὠ/ is present in the revised pad, although it sounds dull compared to the sound created with the filter derived from a recording. For the purposes of this project, the vowel was removed from the layout of the pad, but it could be added again after LPC coefficients from a careful re-recording was made. This recording would need to have formant frequencies that fit into the various transitions that can be made by the vowel pad.

The vowel sound corresponding to /e/ is present in the revised pad, and it sounds slightly dull in comparison to the sound made using the coefficient filter derived completely from the recording of /e/. The /e/ sound on the revised pad is relatively better than the sound for /Ὠ/. In fact, the difference between the interpolation-derived filter and the natural derived filter is slight and subtle, but still noticeable.

2.9.2 Vowels Absent from Test Layout

There are six vowels absent from the test layout. Each of these vowel sounds can be created approximately using the DIDDS device with the test layout.

The /e/ and /习近平新/ vowels are of roughly the same quality in both the revised and test layouts. This is not surprising, given that both layouts do not use this vowel in the interpolation process.

The / Cupertino/ sounds do not occur in the same places in both layouts, although they sound very similar. On the test layout, the / Cupertino/ sound occurs in a location that is much closer to the /e/ sound. A benefit of the revised layout over the test layout is therefore a more staggered spacing of these vowels, which could be important for usability.

The / Cupertino/ and / Cupertino/ sounds are approximated on the test layout, but poorly. They are much more well-defined on the revised layout.

The / Cupertino/ sound is of decent quality in the test layout.

2.9.3 Conclusions Regarding the Effectiveness of the Revised and Test Layouts

The revised layout is a much better choice than the test layout for implementation with
the DIDSS device. Although all of the transitions tested were at least roughly the same, various vowel sounds are much more well-defined in the revised layout. The spacing of the vowels in the revised layout is also more even.

Despite certain vowel sounds sounding less well-defined using the test layout, all necessary sounds can be approximated while using the layout.

The addition of extra vowels in the revised layout creates some irregularities in the formant contour maps. There are some steep cliffs and some formant transitions that are not as smooth as those in the test layout. Spectrograms of certain vowel transitions show slight anomalies not present in the test layout’s waveforms. Subjectively, these are at worst barely noticeable, and most are not noticeable at all.

It is interesting to note that while the vowel pad is moderately effective in modeling the vowel space of American English, it is far from perfect. There are many vowel sounds that can be spoken that are only approximated with the vowel pad. The vowel pad was created with the intent that vowel transitions sound natural, but it by no means contains every possible vowel transition. This is because the vowel space cannot be mapped completely onto a 2-d grid. The formants for a particular vowel sound change between different instances of speech, and not always in a way that is directly mapable onto a 2-d vowel space chart. This is seen when two vowels with nearly identical first two formants have third and fourth formants that differ more, as was the case found in some of the vowel transitions of the proposed layout. However, all in all, the mapping used in this project is a reasonable approximation of the vowel space.
Chapter 3

Derivation of Liquid and Nasal Consonant Filters and Their Transitions with Vowels

Part of the DIDSS device's functionality is to allow the user to create certain periodic consonant sounds for arbitrary amounts of time. These consonant sounds are the liquids and nasals: /l/, /r/, /m/, and /n/. A filter was derived for each of these sounds. Transitions between each consonant filter and every filter used in the vowel pad were created. Several of these transitions were tested.

3.1 Derivation of Filters and Stand-Alone Filter Quality

Filters were derived using LPC from recordings of each of the nasals and liquids. These filters are all 60 order, and were made to be used with 44.1 KHz audio, just like the vowel filters. Subjectively, the /r/ sound is the most well-defined and correct sounding when listened to stand-alone. Its quality is comparable to the best vowel sounds made using the vowel pad. The /l/ sound is fairly natural sounding as well. Both of the nasal filters, the filters for /m/ and /n/, have poorer sound quality. This is not surprising, since it is documented that LPC has a weakness in encoding the nasal sounds [12].

3.2 Transitions Between Periodic Consonants and Vowels

Transitions between the periodic consonants and the vowel sounds corresponding to the vowel pad were made. All transitions were made to those vowels included in the revised layout. Each transition was made to be approximately 30 ms long, by using three separate filters. Each filter, therefore, is used for 10 ms of the transition. Each filter in a transition was found by interpolating between LSF representations of the LPC filters in question.
3.2.1 Transitions to and from /r/ and /l/

Subjectively, the transitions between the liquids /r/ and /l/ and most vowels are generally of good quality. Figures 3.1 and 3.2 show spectrograms of a transition between the /r/ liquid and the /a/ vowel for both synthesized and natural cases.

These spectrograms show many similar features, such as similar formant transition behavior. The transition times are also similar, both approximately 30 ms. There is a slight difference in these two transitions, since the synthesized waveform shows a low-power noise present at around 5000 Hz during a brief time during the transition. The natural waveform shows energy at the higher frequencies during much of the recording. This is likely a result of the speaker’s breathy voice while recording the transition.

The energy present at around 5000 Hz in the synthesized waveform is a result of a fluctuation in filter output when changing from one filter to the next. This is caused by the filter values being too far different from one another to produce completely smooth output. Figure 3.3 shows some of the fluctuations that occur in the synthesized waveform directly after switching between filters. Figure 3.4 shows the waveform after a few cycles. It is apparent that most of the fluctuations are not present in Figure 3.4.

Subjectively, the fluctuations shown are not noticeable while listening to the synthesized waveform for the transition from /r/ to /a/. This is not the case for transitions from vowels to periodic consonants. In many cases, the high frequency content is just barely noticeable in a transition from /r/ or /l/ to a given vowel. In the case of the transition from /r/ or /l/ to /u/, the high frequency content is particularly high-powered. A very audible click is heard while listening to this waveform. Figure 3.5 shows the fluctuations present at this

![Fig. 3.1: Spectrogram of synthesized /r/ to /a/ transition.](image)
A discussion on what can be done to reduce this noise is included in Chapter 4.

3.2.2 Transitions to and from /m/ and /n

The transitions between /m/, /n/, and the vowels are of poor quality. Subjectively, it is difficult to determine by listening what sounds are being synthesized. It is difficult to discern whether an /m/ or /n/ is being spoken, and in some cases, it sounds more like neither is being synthesized, but a different consonant or vowel sound altogether. This is not surprising, given the weakness LPC encoding has with nasals.

In addition to this problem, transitions between the nasals and some vowels show the problem of high-frequency content that is created when switching from one filter to the next. As with the liquids, this problem is most pronounced when transitioning from the nasals and the /u/ vowel.

Possible remedies to these problems will be discussed in Chapter 4.
Fig. 3.3: Noise present at filter transition in synthesized /r/ to /a/ transition.

Fig. 3.4: Plot of clean /r/ to /a/ transition waveform.
Fig. 3.5: Noise present at filter transition in synthesized /r/ to /u/ transition.
Chapter 4

Future Work

4.1 Pad Interpolation

While interpolation between all the vowels of the vowel pad was largely successful, two vowels were removed from the original vowel pad layout. The /e/ vowel was removed, but a similar and well-defined sound can still be created with the revised layout. The /ʊ/ vowel was also removed, but the closest vowel creatable with the revised layout is not as natural or well-defined as that created by the originally included vowel filter.

It may be possible to solve this problem through various methods. A possible solution is to re-record the spoken vowels that did not fit in the layout, with slight variations in the vocal tract of the speaker. With some experimentation, a filter that fits into the layout well might be found. It may also be possible to move the location of a particular vowel on the vowel pad. A proper location might be found by estimating the formants of the filters created by interpolation on the revised pad, and comparing those to the formants of the recorded vowel. Whichever location is closest to the recorded vowel in terms of formants and spectral shape would be used as the vowel’s location. Finally, a different method of interpolation might be used, such as the pole-shifting technique given by Goncharoff and Kaine-Krolak [6]. This technique has been shown to reduce some unnatural behaviors found in other forms of interpolation.

4.2 Derivation of Nasal Consonant Filters

The filters used to produce nasal consonants are not very natural or clear sounding. Nasals are a well-established weakness of LPC. Sometimes, to account for this weakness, LPC-based encoding will not only encode a filter, but will also encode the residual of a
speech signal. When passed through the LPC filter, this residual will produce very high quality sound. A common form used to encode the residual is called the code exited linear prediction (CELP) coder. More information on CELP coders and residual excitation can be found in Schroeder and Atal’s original paper [15].

It may be possible to use a residual as input to the nasal filters, and transition to the standard glottal waveform while transitioning to the vowel filters.

Additionally, it may be possible to increase the effectiveness of the nasal LPC filters by increasing their order. LPC does a poor job at modeling nasal filters because of its all-pole nature. A good nasal filter would have both poles and zeros [16]. The effect of zeros on a filter can be approximated using many poles, hence the addition of more poles may account for problems in the LPC filter. However, if the order of the nasal filters was increased, the order of the vowel filters would also need to be increased to interpolate properly.

4.3 Transitions Between Nasals, Liquids, and Vowels

Certain transitions between nasals, liquids, and vowels contain audible clicks due to large differences in filter coefficients from one filter to the next. This problem could be rectified in a few ways. One solution is to use filters that have values that are closer together. This might be done by adding more filters into the transition, while lowering the time each filter is in use. Instead of using three filters for 10 ms each, 9 filters might be used for 3.33 ms each. Another solution is to add a low-pass filter to reduce clicks and noise when filters are changed.
Chapter 5

Conclusion

LPC filters were derived from recordings of ten different vowels. A grid of 200x200 different filters was derived from these vowels using LSF interpolation. The filters in this grid are able to approximate any vowel in the North American English vowel space.

Interpolation of LPC filters using LSF coefficients was tested. Three different layouts for the vowel grid were used to guide interpolation. The initially proposed layout used all ten filters from the original vowel recordings. Testing on this layout revealed unnatural transitions between certain vowels. Because of this, a revised layout was created that excluded two of the vowels. This layout gave satisfactory results, although there was a decrease in how well-defined the sounds corresponding to the removed vowels sounded. A third layout was tested that used only four vowels to determine the entire vowel space. While this layout can be used to approximate most vowel sounds, the approximations do not always sound well-defined, and are sometimes of poor quality.

The interpolation of the revised layout’s vowels proved that linear interpolation of LSFs does not always lead to linear interpolation of formant frequencies. The second formant changes in a nonlinear fashion along certain vowel transitions on the vowel pad.

LPC filters were determined for liquid and nasal consonants. The filters for the liquids produce high-quality speech sounds, but the filters for the nasals do not produce natural sounding speech.

Additionally, transitioning filters were found using LSFs for transitions from liquids or nasals to the vowels. Due to large difference in the filters for the consonants and the vowels, some of synthesized transitions have audible clicks in them when the filters switch during the transition.
Overall, LSFs were effective at producing realistic sounding vowel and consonant transitions from relatively few data points.
References


