USING A WHITE NOISE SOURCE TO CHARACTERIZE A GLOTTAL SOURCE WAVEFORM FOR IMPLEMENTATION IN A SPEECH SYNTHESIS SYSTEM

by

Brandon R. Graham

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

of

MASTER OF SCIENCE

in

Electrical Engineering

Approved:

_____   _____
Dr. Jacob H. Gunther   Dr. Todd K. Moon
Major Professor   Committee Member

_____
Dr. Jeffery B. Larsen
Committee Member

UTAH STATE UNIVERSITY
Logan, Utah

2013

Abstract

Using A White Noise Source to Characterize A Glottal

Source Waveform for Implementation

in a Speech Synthesis System

by

Brandon R. Graham, Master of Science

Utah State University, 2013

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

A novel speech synthesizer is being developed which needs a source waveform that represents the sound created by the vocal folds before it is shaped by the rest of the vocal cavity. Methods already exist for extracting this waveform, but this report explores a new method. The method involves finding a model for the vocal tract. A system identification technique is applied that uses a white noise audio source emitted into the oral cavity via a tube as the input. The effects of the tube are characterized and accounted for to allow for greater accuracy in the estimation of the true vocal tract properties. The vocal tract model is then used to extract the source waveform from a vocalized speech recording.

Common properties of the source waveform will also be characterized and synthesized. These properties include the changes in harmonic content of the source based on vocal effort, and the natural aperiodic fluctuations in pitch and amplitude of the source waveform. All of these properties, when properly synthesized, will help to create a more natural-sounding glottal source waveform.

(53 pages)

Public Abstract

Using A White Noise Source to Characterize A Glottal

Source Waveform for Implementation

in a Speech Synthesis System

by

Brandon R. Graham, Master of Science

Utah State University, 2013

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

A novel speech synthesizer is being developed which needs a source waveform that represents the sound created by the vocal folds before it is shaped by the rest of the vocal cavity. Methods already exist for extracting this waveform, but this report explores a new method. The method involves exciting the vocal tract with white noise, which is introduced into the mouth via a tube. While this has been attempted before, the effects of the tube itself on the white noise were not previously accounted for. This approach accounts for the affects of the tube in order to obtain a more accurate model of the vocal tract and source waveform.

Also, the natural pseudo-random fluctuations in pitch and amplitude of the source waveform are studied, and a simple but effective solution is proposed for their implementation in the new speech synthesizer.

For Brady.

Acknowledgments

Thanks to Jake Nieveen and Jacob Gunther for helping this project to be a success. And a special thanks to my family for supporting me in my scholastic efforts.

Brandon R. Graham

Contents

List of Figures

Acronyms

DIDSS   Direct-Input Digital Speech Synthesis

DFT   Discrete Fourier Transform

F0   Fundamental Frequency

FFT   Fast Fourier Transform

LP   Linear Prediction

LTI   Linear Time-Invariant

RMS   Root Mean Square

TTS   Text-To-Speech

VT   Vocal Tract

Chapter 1

Introduction

1.1 History of Speech Synthesizers

Modern day speech synthesizers produce very intelligible speech and are highly useful for a broad range of applications. The best synthesizers currently available for general use are text-to-speech (TTS) systems. These systems utilize sophisticated programs which analyze written text and decide which sounds of speech, or phonemes, should be produced. The best TTS systems currently utilize a form of concatenative synthesis referred to as unit selection [1], which was proposed by Andrew Hunt and Alan Black in 1996 [2]. Although these modern systems are capable of producing very intelligible speech, they would still not be mistaken for an actual human if listened to for any significant duration of time.

What makes the current systems sound inhuman is the lack of correct usage of the pitch and duration (also known as the prosody) of speech. TTS systems will always struggle with trying to infer information about the prosody of speech from text, because most prosodic information is not contained in the written text to begin with [3]. A solution for the prosody problem is to let a human give the input to the synthesizer instead of assigning the task to a computer.

Such direct-input systems are not a new concept, and they actually predate text-to-speech synthesis by more than a century. The first documented systems reproduced the sounds of speech by purely mechanical means, the most famous of which was Wolfgang Von Kempelen's talking machine [4]. This machine could generate recognizable vowels and some consonant sounds. The next significant improvement was an analog electrical speech synthesizer made by Homer Dudley for AT&T in the 1930's [5]. Dudley's Vodor, as he named it, could produce speech of a much higher quality than its mechanical predecessors, and was manually played by an operator on a keyboard.

With the advent of digital electronics, focus quickly turned to automatic speech synthesis based on interpretation of text by a computer. Great strides have since been made in the intelligible digital reproduction of speech, and it is the goal of this project to use some of these advances to make a more intelligible digital counterpart to Dudley's Vodor. This new system will hereafter be referred to as the Direct-Input Digital Speech Synthesis (DIDDS) system. This report addresses only a portion of the new speech synthesis system. It outlines a new method of extracting the glottal source waveform, as well as ways to process the waveform in order to make it sound more natural.

1.2   The Glottal Source Waveform

The glottal source is the basic sound generated by the vocal folds before it has been filtered by the rest of the vocal tract. The pitch of a speech signal is controlled by the glottal source waveform. The glottal source waveform varies among different people and is important in identifying the speaker [6]. The glottal source controls the pitch, intensity, and quality of the voice. Here, the word "quality" refers to whether a person's voice sounds soft, shaky, raspy, etc. An accurate portrayal of these emotions is essential for a voice synthesizer whose main goal is greater flexibility in expression of emotion. The various aspects of the glottal source will now be covered in more detail.

1.2.1   Pitch

In order to accurately control the pitch of synthesized speech, the glottal source waveform must be modified without changing the characteristics of the rest of the vocal tract. If the final waveform were to be modified instead, not only would the pitch of the source change, but the location of the formants (peaks in the frequency response of the vocal tract) would change as well. Changing the formants has the effect of making the voice sound more like a man, woman, or child (depending on which way the formants shift). So in order to independently change the pitch and the formants of a synthesized voice, the glottal source waveform must be separated from the effects of the vocal tract. This is commonly known as the source-filter model of speech. Chapters 2 and 3 describe different techniques for

extracting the glottal source waveform from recorded speech signals.

In the DIDSS system, the glottal source waveform will be stored in memory on a computer. During periods of vocalization, samples of the source waveform will be streamed to the audio port of the computer. The choice of which samples to read depends on the desired pitch of the signal. For instance, to synthesize a higher pitch, more samples of the original waveform are skipped as they are streamed to the audio card at a constant rate. This has the effect of raising the perceived pitch of the glottal source.

The human ear does not perceive pitch on a linear scale. Pitch is perceived as roughly the log base two of the emitted frequency. In order to create an intuitive user interface for the DIDSS system, the pitch should change in a manner congruent with natural perception. In order to achieve this, the frequency increments and decrements at a rate of 2^x instead of a linear x. This is done by incrementing the index by an amount of 2^x, where x would be a value linearly proportional to the movement of the user's finger on the interface.

1.2.2   Intensity and Quality

In order to create a realistic synthesized voice, the intensity and quality of the glottal source need to be appropriately modeled. These effects are more complex to mimic than pitch changes, so they will be discussed in more detail later in the report. Chapter 4 of the report will discuss how to accurately mimic changes in amplitude of the source waveform, while Chapter 5 will address issues relating to the perceived quality of the glottal source.

Chapter 2

Existing Methods for Glottal Waveform Extraction

2.1 Overview

The vocal folds are the source of the sound that is heard when a person vocalizes (otherwise known as the glottal source). By varying the tension and air flow through the vocal folds, a person can vary the pitch at which they are vocalizing. Different vowel and consonant sounds are perceived when the person changes the configuration of their vocal tract, which is the term given to the pharynx, the oral cavity, and the nasal cavity. As the vocal tract changes shape, it alters the amplitudes of the glottal source and harmonic frequencies in different ways. This altering of frequency amplitudes is known as filtering. Thus, it is common to think of vocalization as taking a source (the glottal waveform) and passing it through a filter (the vocal tract) in order to get the result that is perceived by the listener.

This source-filter model of speech treats the glottal waveform as the source and lumps the effects of the vocal tract into a filter which modifies the source. This model is the basis for the two methods of glottal waveform extraction which will be discussed in this chapter. However, the model assumes that the filter does not change with time and is independent of the source. This report explores these assumptions to the practical limit by attempting to characterize the vocal tract completely independently of any glottal activity. In reality, the behavior of the vocal tract changes constantly and is not independent of the glottal source. Chapters 3 and 6 will address this issue in more detail. For more information about the source-filter model of speech, see the classic book by Gunnar Fant [7].

2.2 Linear Predictive (LP) Analysis

Linear predictive analysis is one common method used to separate the glottal source

from the vocal tract filter. It involves using the input speech signal to create a linear time-invariant (LTI) filter which represents the vocal tract. This LTI filter is an all-pole filter whose coefficients are typically derived from the autocorrelation of the input signal. For more information on LP analysis, see the article by B. S. Atal and S. L. Hanauer [8].

The spectral response of the LP filter can be used to counteract the effects of the vocal tract. Since the frequency domain of the voiced signal can be viewed as the multiplication of the frequency domains of the glottal source and vocal tract filter, dividing by the LP filter response in the frequency domain will have the effect of cancelling out the effect of the vocal tract. Thus, what is left over is the glottal source signal.

2.2.1 Benefits of LP Analysis

LP analysis in its simplest form is easy to perform on a signal. It proves a fast and reasonably accurate estimation of the properties of the vocal tract. A more complex but more accurate version of LP analysis may be performed on certain portions of the vocalized waveform if greater accuracy is desired.

2.2.2 Drawbacks of LP Analysis

As LP analysis involves modeling the vocal tract as an all-pole filter, the peaks of its spectrum tend to be too pronounced, especially for vowel synthesis [1]. This is due to the fact that in vowel representations, the LP filter pole magnitudes lie close to the unit circle, so small changes in the coefficients can result in large changes in formant bandwidth estimation.

The amplitudes of the harmonics of the glottal source slope downwards at a rate of approximately -12dB/decade as the frequency increases. The overall slope of a vocalization is around -6 dB/decade. Since this slope is gradual compared to the peaks of the glottal wave harmonics, it is mistakenly captured by the LP filter as a characteristic of the vocal tract. Thus, if the LP filter is used to inverse filter a vocalized signal, the resulting glottal source harmonics will not have a downward frequency trend at all. A -12 dB/decade slope

can be applied after processing to get a more accurate model of the glottal waveform, but it will be an approximation only.

2.3 Cepstral Analysis

Cepstral analysis is another common method used to separate the glottal source from the vocal tract filter. In essence, the result obtained by cepstral analysis is the frequency domain of the logarithmic frequency domain of the original signal. By properly utilizing this information, information about the glottal source and the vocal tract filter can be extracted from a vocalized recording.

Normal speech can be thought of as the result of convolution of the glottal source and impulse response of the vocal tract filter. In the frequency domain, this convolution equates to multiplication of their spectra, and in the logarithm of the frequency domain, it equates to addition of their spectra. Thus, the logarithmic magnitude response of a vocalization can be thought of as the sum of the logarithmic magnitude responses of the glottal source and vocal tract filter.

The cepstrum (the result of cepstral analysis) is usually defined as the inverse Discrete Fourier Transform (DFT) of the log magnitude of the DFT of a signal. It gives a frequency analysis of the spectrum of a signal. The lower elements of the cepstrum contain information about the more gradually changing frequency characteristics of the vocal tract. The higher elements of the cepstrum contain information about the spikes in the frequency domain of the signal caused by the fundamental frequency and harmonics of the glottal source. By eliminating the lower cepstral elements (which correspond to the spectral response of the vocal tract) and re-transforming back to the original time domain, a representation of the glottal source can be extracted. For more information on cepstral analysis, see the article by A. V. Oppenheim and R. W. Schafer [9].

2.3.1 Benefits of Cepstral Analysis

Due to the lower fundamental frequency of the male voice, the harmonics are spaced closer together in the frequency domain. This makes it easy to separate these rapidly occur-

ring peaks and valleys of the glottal source from the more slowly changing characteristics of the formants of the vocal tract with cepstral analysis.

The amount of spectral precision in estimation of the vocal tract and glottal source can be modified by choosing how many of the cepstral coefficients are used in each case. This allows the user to "dial-in" the best estimations in order to get a more accurate separation of source and filter.

2.3.2 Drawbacks of Cepstral Analysis

Due to the higher fundamental frequency of female and child voices, the harmonics are spaced farther apart in the frequency domain. This makes it more difficult to accurately separate the glottal source from the vocal tract with cepstral filtering. Some of the glottal source characteristics will thus be mistaken as vocal tract characteristics and vice versa. This problem will always occur to some extent even with male voices, however it is much more noticeable with female and child voices.

Just like with LP analysis, the attenuation of glottal harmonics as frequency increases will be mistaken as part of the vocal tract filter. This effect can be compensated for, but again, it will only be an approximation.

Chapter 3

Proposed Method for Glottal Waveform Extraction

3.1 Overview

A white noise audio signal is delivered via a tube and emitted in the oral cavity of a specific individual. The noise is recorded in order to characterize the frequency response of the vocal tract for a certain configuration of the oral cavity. The specific individual then vocalizes with the same oral cavity configuration (and with the tube still in the mouth), and the sounds are recorded. The noise recordings are used to inverse filter the vocalized recordings in order to extract a glottal source waveform.

3.2 The White Noise Source

When a source sound is filtered by the vocal cavity, the frequency content of the resulting signal is simply the element-wise product of the frequency responses of the source sound and the vocal cavity [10]. White noise was chosen as a source because of the fact that its frequency distribution is (ideally) spectrally flat over a defined bandwidth. Thus, if using white noise as the source sound, the magnitude frequency response of the recorded output is the same as the magnitude frequency response of the vocal cavity, being different only by a scalar factor. White noise is thus used so that the frequency response of the vocal cavity can be found by directly analyzing the recorded audio output.

3.3 White Noise Generation and Recording

Ten seconds of white Gaussian noise is generated at a sampling rate of 44.1 kHz via MATLAB. This ensures a spectrally "flat" response for frequencies ranging from zero Hz to half the sampling rate, or 22.05 kHz. The human ear can detect frequencies up to about 20 kHz. The bandwidth of 22.05 kHz was chosen for the noise so that frequencies in the entire

audible range of human hearing would be accurately characterized by this experiment. This ensures that all audible formants will be defined, thus allowing for characterization of the glottal waveform and its harmonics in the entire audible range.

3.3.1 Recording Setup

The white noise is played through the audio ports of a computer into a powered speaker inside a tube. This speaker and tube configuration are known as a "talk box" and are commercially produced for audio special effects. The tube is inserted into the subject's mouth, with the mouth held in whatever shape is desired to create a certain sound of speech (commonly known as a phoneme). The noise will travel through the tube and into the subject's mouth, resonating in the mouth, throat, and nasal cavity (together known as the vocal tract) before exiting and being recorded by a high-quality condenser microphone and stored digitally on a computer. The following information is included to enable easy reproduction of this experiment.

- Recording System Information

    - Talk Box: Rocktron Banshee Talk Box

    - Microphone: Marshall Electronics MXL 993 condenser microphone

    - Distance from microphone when recording noise: roughly 8 cm

    - A pop filter was used in order to prevent high amplitude, low frequency pressure waves from reaching the microphone.

    - Noise floor: Roughly -48 dB

    - All recordings are taken for exactly ten seconds.

    - All recordings are taken at a sampling rate of 44.1 kHz.

- Vocalization Information

    - Target fundamental frequency: 160 Hz

    - Target RMS amplitude levels: -15 dB, -9 dB, and -3 dB (These correspond to low, medium, and high vocal effort levels, respectively.)

3.3.2 Recording and Processing the Noise Signals

By the time the audio signal has been recorded, it has not only been modified by the vocal tract, but also by the audio amplifiers, speaker, and tube that the sound must pass through. Additionally, the spectral response of the microphone is not completely flat [11], causing certain audio frequencies to be amplified or diminished with respect to the others. In order to compensate for the effects of these unwanted filters, the noise recordings must be pre-processed before they are able to be properly analyzed. Because a white noise source is used, this task is possible by simply making a recording of the noise through the same speaker, tube, and microphone configuration but without passing the noise through the vocal tract. The overall frequency response due to these unwanted contributors will be used to make a pre-processing filter to compensate for their effects.

Figure 3.1 outlines the process of how the white noise source $x_1(t)$ is passed through the talk box, vocal tract, and microphone before being recorded by the computer as signal $y_1(t)$.

Figure 3.2 outlines the process of how the white noise source $x_1(t)$ is passed through the talk box and microphone before being recorded by the computer as signal $y_2(t)$. The Fourier transform of $y_2(t)$ yields the frequency response of the talk box and microphone without the vocal tract:

$$\mathcal{F}\{y_2(t)\} = F_{TalkBox}F_{Microphone}.$$

The preprocessing filter is made from the Fourier transform of y$_2$(t):

$$F_{PreProcessing} = \frac{1}{F_{TalkBox}F_{Microphone}} = \frac{1}{\mathcal{F}\{y_2(t)\}}.$$
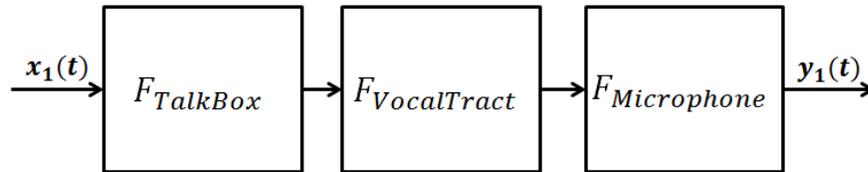
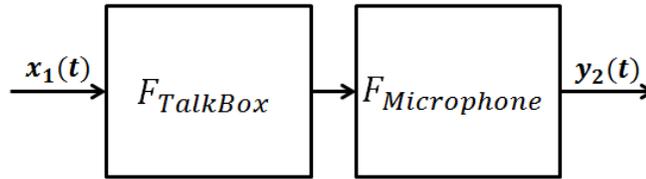Fig. 3.1: Vocal tract white noise recording process.

Fig. 3.2: Environmental effects white noise recording process.

The original signal $y_1(t)$ is then filtered with the pre-processing filter in order to obtain the corrected frequency response of the vocal tract.

$$y_1(t)_{Corrected} = \mathcal{F}^{-1}\{\frac{\mathcal{F}\{y_1(t)\}}{\mathcal{F}\{y_2(t)\}}\}$$

Figure 3.3 shows the magnitude frequency domain of the white noise before and after it has been passed through the tube and recorded with the microphone. The changes are very significant, and it is clear that these need to be accounted for if an accurate spectrum is going to be derived for the other recorded noise signals.

Figure 3.4 shows the smoothed inverse spectrum of the talk box noise signal which is used as the pre-processing filter to account for the effects of the talk box and microphone. All smoothed spectra were smoothed with a moving average window with a size of 2000 samples.

Figure 3.5 shows the smoothed spectrum of the /u/ noise signal which was created by passing noise through the oral cavity while it was shaped in the same configuration that would typically produce the /u/ phoneme. We can see that it is essential to correct the spectra for the effects due to the talk box and microphone.

Figure 3.6 shows the spectrum of the inverse filter for the /u/ phoneme. This filter is used to extract glottal waveforms for all recordings with the /u/ phoneme. Identical processes are performed for the noise recordings for /a/ and /i/.

3.3.3 Precautions

Careful consideration must be taken as to the state of the individual's vocal folds during the white noise recording. If the individual holds the vocal folds in a relaxed state (such as
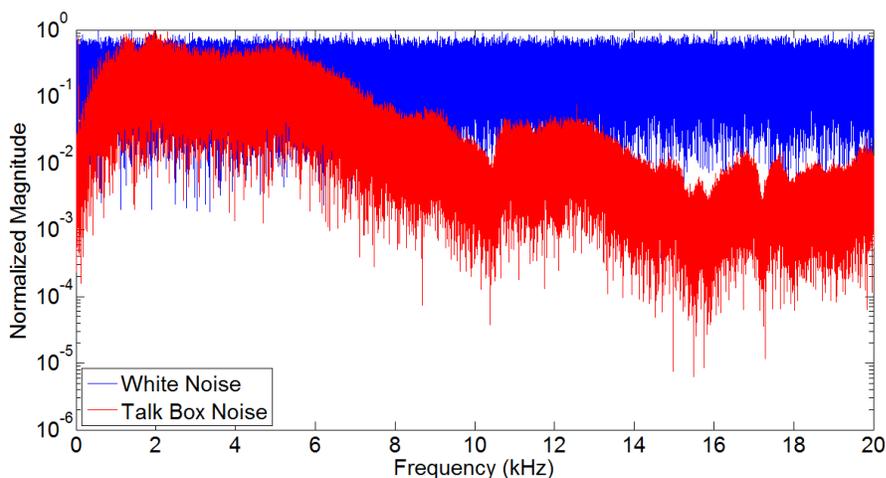
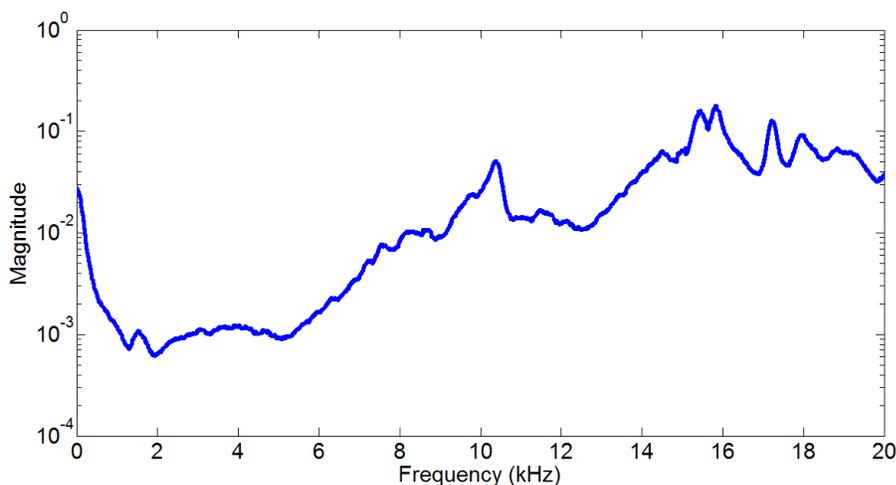Fig. 3.3: White noise spectra vs. talk box noise spectra.

Fig. 3.4: Environment pre-processing filter.

when breathing normally), the glottis will be open and the input noise will resonate in the portion of the trachea below the vocal folds known as the subglottal tract. In effect, this lengthens the vocal tract, causing the vocal tract formants to shift by a significant amount. This problem can be overcome by ensuring that the individual closes their vocal folds during the white noise characterization. This can be done by having them hold their breath while keeping the mouth in the desired configuration. When an individual holds their breath, the vocal folds are pressed tightly together, allowing no air to pass through them.
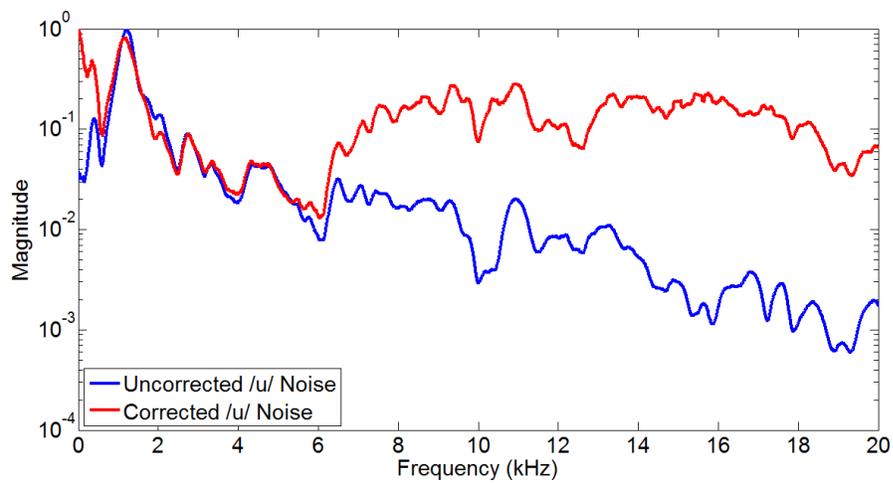
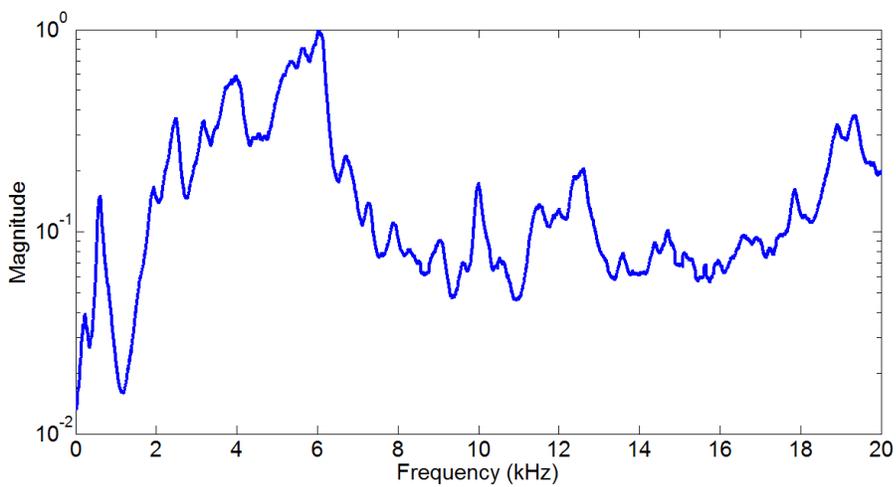Fig. 3.5: Uncorrected vs. corrected smoothed /u/ noise spectrum.

Fig. 3.6: Inverse filter for /u/ phoneme.

3.4   Recording and Processing the Vocalized Signals

Now that the /u/, /a/, and /i/ formants have been characterized with the noise source, the vocalized recordings must be processed with their corresponding filters to extract estimates of the glottal source. This is done by taking the Fourier transform of the recording, multiplying its magnitude response with the appropriate vowel inverse filter spectrum, and then taking the inverse Fourier transform. The result (assuming a perfect estimate of the formants by the inverse vowel filter) is a time domain representation of the glottal source.

To be more correct, the result would actually be the glottal source filtered by the frequency response of the microphone, because the effects of non-ideal frequency response of the microphone were not accounted for in the vocalized recordings. In order to obtain more accurate results, an inverse filter could be developed which would account for these effects. Due to time constraints, that filter will not be developed for this report. It is thought that neglecting these effects will not have significant ramifications, due to the fact that the spectral response of the microphone changes much more gradually with respect to frequency than the formants which are being characterized.

3.5   Benefits of the White Noise Method

- High resolution spectral response characterization of the vocal tract.

    - The nature of the white noise recording scenario allows for long audio segments to be recorded and analyzed, resulting in a very high spectral resolution. For instance, with 10-second segments recorded at 44.1 kHz, the resulting digitized signal will contain 441,000 samples, allowing for a frequency resolution of 0.1 Hz. However, the spectrum of the recorded noise signals will have to be smoothed in order to be useful, so the usable frequency resolution is somewhat less than the ideal 0.1 Hz.

    - LP and cepstral analysis could also be used for longer vocalized audio segments, however, due to the nature of their formant estimation, they will not be able to

reveal high resolution details of the vocal tract that can be found with the white noise method. If they were tuned to reveal more detail, they would begin to include information about the glottal source and mistakenly attribute it to the vocal tract.

- The glottal source is almost completely separated from the vocal tract filter.

  – While the source/filter model has its limitations when it comes to glottal source separation (see next section), the white noise experiment is a useful tool to study the limits of this model. Because the vocal folds are not in motion during the noise recording, the vocal tract filter is characterized completely devoid of the effects of vocal fold movement. The presence of the vocal folds, their shape, and position, will all of course still effect the characterization of the vocal tract with the noise method.

- Flatter spectral response (no rolloff due to the glottal source).

  – LP and cepstral analysis attribute all general trends in the frequency domain to the vocal tract. The problem is, the harmonics of the glottal source also follow a general trend, in that as frequency increases, the magnitude of the harmonics falls off at a rate of roughly -12 decibels per decade. This gradual overall change in the spectrum is attributed to the vocal tract by LP and cepstral analysis, thus distorting the model of the vocal tract. The noise method has a spectrally flat excitation as a source, so that all trends in the final recording will actually be due to the vocal tract and not the source.

3.6 Drawbacks of This Method

- The noise method requires special tools and more time than LP or cepstral analysis.

  – Whereas LP or cepstral analysis can be performed on a standard vocalized signal, the noise method requires the use of a talk box or similar apparatus. The

process of characterizing the vocal tract with the noise, along with making the other necessary vocalized recordings, causes the noise method to take longer than traditional methods. However, the total amount of required time is still not large, as an individual could make the recordings in a matter of minutes.

- This method cannot be used to characterize the vocal tract for nasal consonants.

  - The characteristics of the vocal tract during phonation of nasal consonants such as /m/ and /n/ cannot be found with the white noise method. This is because the mouth is closed during the production of those sounds, and thus the noise cannot be introduced into the vocal tract.

- The noise source is not in same location in the vocal tract as the actual glottal source.

  - This is unavoidable if the recordings are to be done with a live, conscious person. The complete ramifications of this issue are not known. It is postulated that some of the noise could be reflected at the oral cavity and then recorded before it could pass through the rest of the vocal tract. This would cause the effects due to the lower vocal tract to appear less pronounced than actual. Also, in order for the noise to pass through the entire vocal tract, it must do so at least twice (once going in and then again going out). This may have the effect of making all formant peaks appear more pronounced than actual.

- The tube blocks a significant portion of the opening at the lips.

  - This is also unavoidable, because the tube needs to be fed into the mouth in order for the sound to be emitted in the vocal tract. This causes a change in the structure of the formants, because it causes a significant change in the geometry of a critical junction in the vocal tract [12]. Luckily, this adverse affect can be accounted for simply by making vocalized recordings with the tube in the mouth, which was done in this case.

- The glottal source is almost completely separated from the vocal tract filter.

– Sadly, this seems to be more of a disadvantage than an advantage when attempting to extract the glottal source. If the glottal source and vocal tract filter were completely independent, the white noise method would be a very appealing choice. However, in actuality the glottal source and vocal tract filter are coupled together, and one cannot be fully characterized if the other is removed from the experiment.

– Also, since the frequency response of the vocal tract is known to change even during one glottal cycle [1], obtaining high resolution information about the vocal tract devoid of the glottal source may not be of much use for natural speech synthesis.

Chapter 4

Changes in Glottal Harmonic Content Due to Change in Vocal Effort

4.1 Background

In a simple speech synthesizer, the amplitude of the voice can be synthesized by simply scaling the speech waveform to be larger or smaller before it is sent to the speakers. However, in reality the properties of the glottal source change as the effort expended by the speaker varies. For instance, when somebody shouts, it does not sound the same as if they spoke softly and then turned up the volume of their speech. Part of this effect is due to the fact that shouted speech is generally spoken at a higher pitch than softly spoken speech. But even if a word was spoken and then shouted by the same individual at the same pitch and the two instances were compared at equal amplitudes, they would sound different. This is because the harmonic content of the glottal source varies depending on the effort expended by the speaker [1].

Every periodic signal can be decomposed into its constituent sinusoidal waveforms. These sinusoids oscillate at either the fundamental frequency of the signal or at integer multiples of the fundamental frequency. The sinusoids which oscillate at integer multiples of the fundamental frequency are called the harmonics. The strength of these harmonics affect how the signal sounds (also known as the timbre of the sound). For instance, if a human and a violin both sustain a note at the same frequency, they sound different because of the different strength of their harmonics. This is the same reason that quietly spoken speech sounds differently than loud speech. In general, louder speech tends to have relatively more power in the higher harmonics than softer speech [13].

4.2   Characterization

Harmonic analysis was performed on the glottal source signals after they had been extracted from the vocalized recordings via cepstral analysis (see Chapter 6 for reasons why cepstral analysis was used). Figure 4.1 shows the plot of harmonic strength as a ratio to the fundamental frequency F0. The x axis for the plots represents the first seven harmonics. Ideally, the three plots corresponding to each vocal effort level would be identical. We can see from these plots that they differ greatly in shape and strength. This indicates that the glottal source extraction method is not perfect, and the sources still contain some information about the formants which were removed as much as possible. The general trend is still noticeable, however, that the strength of the harmonics does tend to increase as the vocal effort increases.

Figures 4.2, 4.3, and 4.4 help to show how the harmonic strength change with respect to vocal effort for the /u/, /a/, and /i/ phonemes, respectively.

It can be seen that the /i/ phoneme is the noticeable exception in this case, with harmonic strength actually decreasing at the highest vocal effort. It is not known whether this is truly a characteristic of the source, or whether these results are simply skewed by an inaccurate glottal source extraction. Because the results do not agree well with one another, no further attempt was made at developing a model to mimic the changes in harmonic content of the glottal source due to vocal effort.
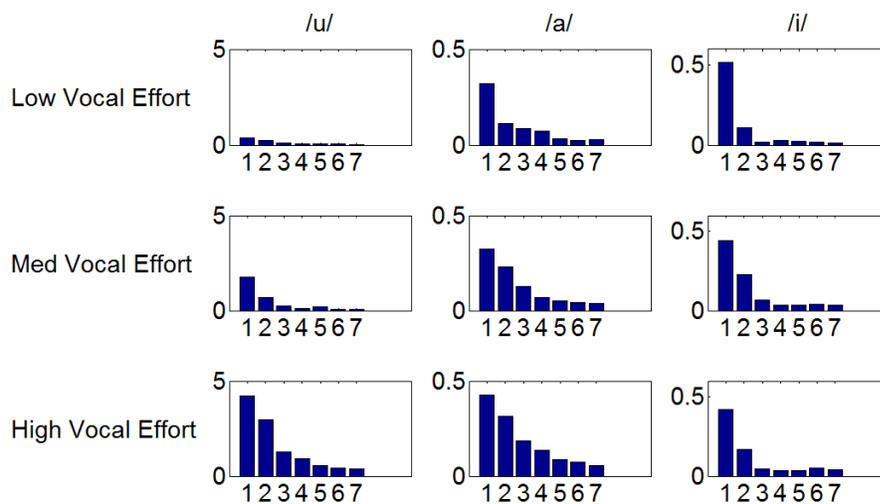
Fig. 4.1: Harmonic strength for all test vowels at all vocal effort levels.
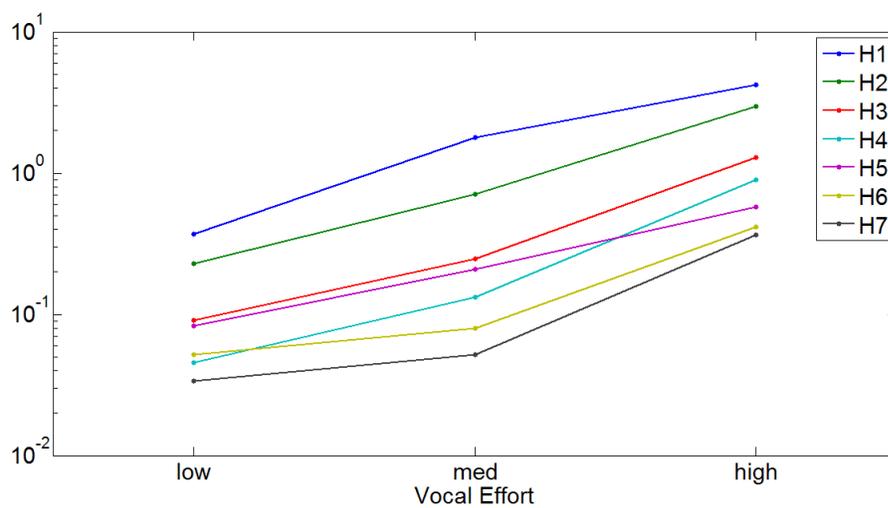

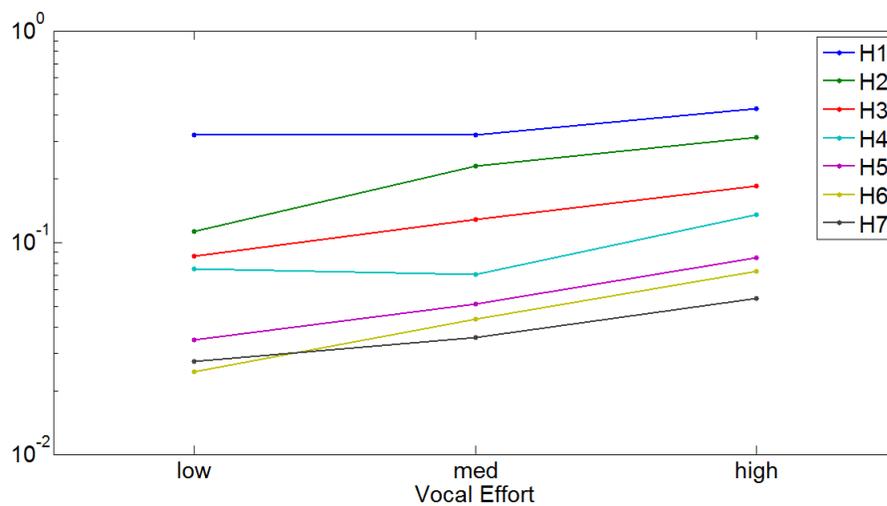Fig. 4.2: Harmonic strength vs. vocal effort for the /u/ phoneme.

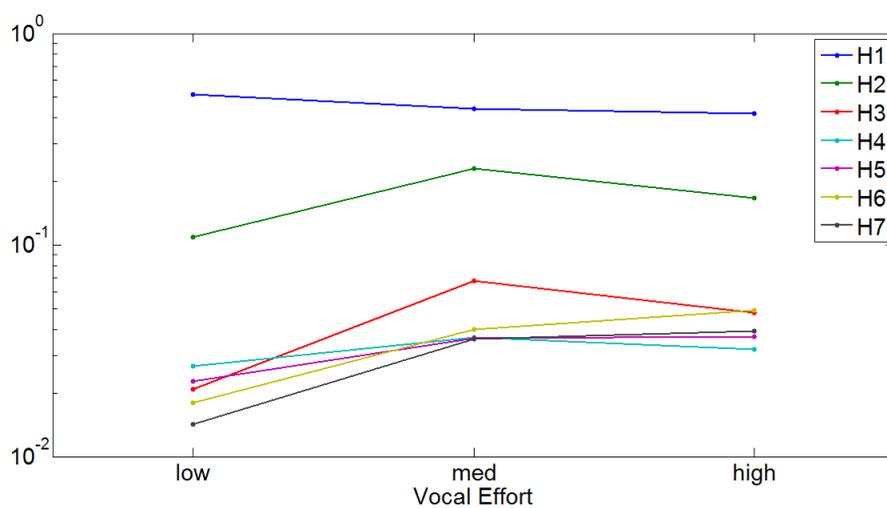Fig. 4.3: Harmonic strength vs. vocal effort for the /a/ phoneme.

Fig. 4.4: Harmonic strength vs. vocal effort for the /i/ phoneme.

Chapter 5

Synthesizing Glottal Frequency Jitter and Amplitude Shimmer

5.1   Background

Even during periods of sustained vocalized speech (such as when a vowel is being spoken), the glottal source fluctuates slightly in both pitch and average amplitude from moment to moment. These variations appear to be caused by vocal fold asymmetry, involuntary muscle activity, and fluctuations of airflow and pressure [14]. In normal sustained speech, the frequency and amplitude fluctuations are about 1% and 6%, respectively [14].

Without these cycle-to-cycle variations in the glottal waveform, it would tend to sound unnaturally steady and machine-like [15]. Thus, in order to create a natural-sounding speech synthesizer, the effects of variation in pitch (also known as jitter) and variation in amplitude (also known as shimmer) must be accounted for.

While models have been created for both jitter and shimmer [16–18], they are still not well-understood phenomena, and attempts to accurately synthesize these effects still fail to sound completely natural, although they are improving [17, 18]. Thus, it was desired to characterize the jitter and shimmer of the vocalized recordings made for this study in attempt to create a reasonably accurate model for a specific individual.

5.2   Jitter

Jitter is the term for the cycle-to-cycle variations in the fundamental frequency (commonly known as F0) of the glottal waveform. In subjective listening tests, it was found that jitter is a more dominant factor than shimmer for perceived naturalness [15]. The following sections will outline the process of creating a model that will accurately model jitter signals for an individual person.

5.2.1 Characterization

In order to track the changes in fundamental frequency over time, the vocalized speech was split into overlapping frames, and a chirp-z transform was used to estimate the fundamental frequency for each frame. The chirp-z transform used 1024 frequency data points which ranged from 150 to 170 Hz (because the frequency target for the vocalized recordings was 160 Hz). Each frame was 150 ms long, and each consecutive frame was shifted by an increment of 10 ms. At a frequency of 160 Hz the cycle period is about 6.25 ms, so the frame encompasses roughly 24 cycles while the frame is shifted each time by an increment of less than two complete cycles.

Because each fundamental frequency estimate encapsulates multiple glottal cycles, the more rapid changes in fundamental frequency will be averaged out. This was done intentionally, because the audible jitter effects that are desired to be characterized happen relatively slowly, on the order of a few times per second. Faster changes in pitch would result in a source that simply sounds noisy or hoarse. Jitter synthesis attempts by other groups [17] have reported in subjective listening tests that listeners thought their synthesized jitter sounded unnaturally hoarse. It is thought that modeling only the slower changes in fundamental frequency will yield a more natural-sounding synthetic voice.

Figure 5.1 shows the plot of fundamental frequency versus time for the /i/ phoneme at high vocal effort level. Note that even with the large frame size, multiple fundamental frequency fluctuation cycles are captured per second.

Recall that nine vocalized recordings were made, with three vowel phonemes and three vocal effort levels per vowel. Jitter analysis was performed on all nine recordings, and their spectra were computed via fast Fourier transform (FFT) and then averaged together. The resulting spectrum is shown in Figure 5.2.

5.2.2 Synthesis

Note in Figure 5.2 that there are no large spikes in the spectrum, indicating that the jitter signal would be well-characterized by some form of broadband excitation. White noise would be a poor choice, because its flat spectral response does not accurately reflect the
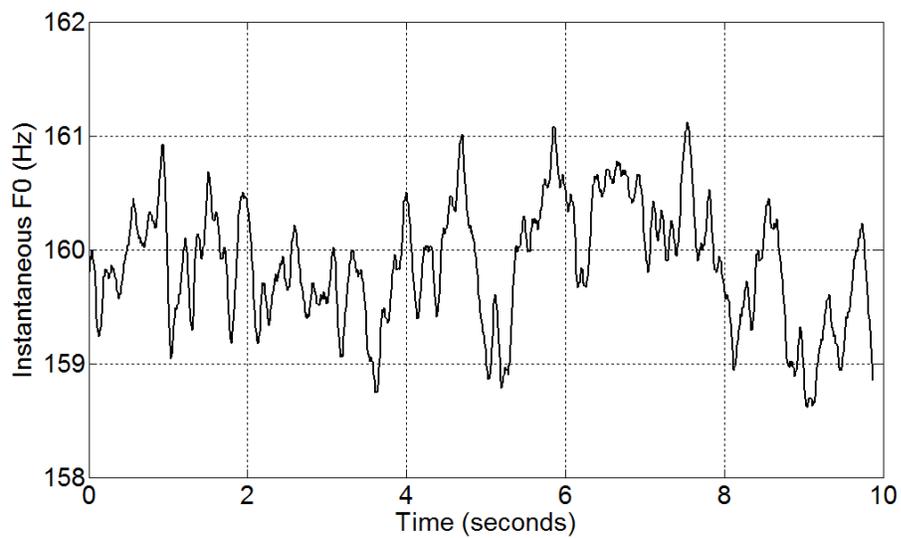
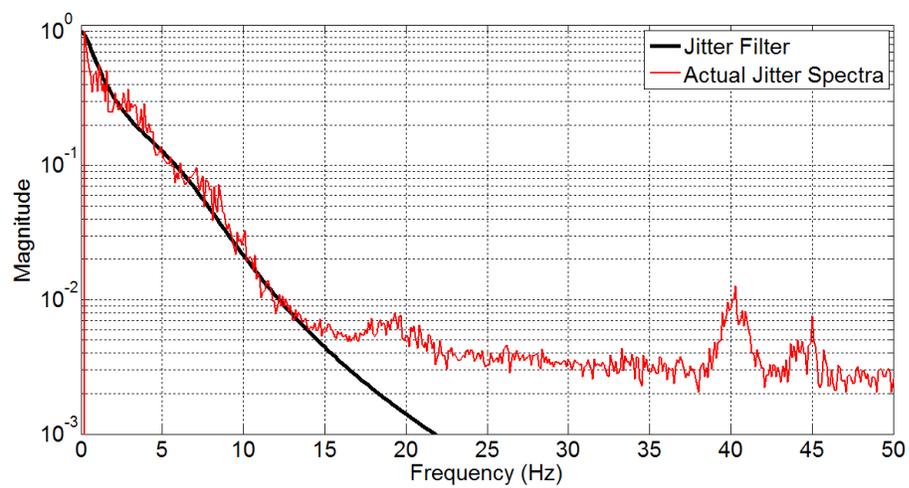Fig. 5.1: Jitter plot (F0 vs. time) for /i/ at high vocal effort.

Fig. 5.2: Jitter filter spectrum vs. actual jitter spectrum.

overall shape of the jitter spectrum. However, if the noise is filtered, a close approximation is possible.

It was found that a good synthetic jitter signal could be reproduced by processing Gaussian white noise through a cascade of two low pass filters. The first filter is a second order Butterworth low pass filter with a passband frequency of 0.2 Hz, a stopband frequency of 3 Hz, and a stopband attenuation of 13 decibels. The filter coefficients are as follows:

$$\text{Numerator} = [4.908813832115902\text{e-}05, 4.908813832115902\text{e-}05, 0],$$
$$\text{Denominator} = [1, -0.999901823, 0].$$

The second filter is a third order Butterworth low pass filter with a passband frequency of 3 Hz, a stopband frequency of 32 Hz, and a stopband attenuation of 40 decibels. The filter coefficients are as follows:

$$\text{Numerator} = [1.183535152\text{e-}10, 3.550605455\text{e-}10, 3.550605455\text{e-}10, 1.183535152\text{e-}10],$$
$$\text{Denominator} = [1, -2.998035452, 2.99607283, -0.998037381].$$

These two filters are cascaded together into a new fifth order filter which is used to synthesize the jitter signal. The frequency response of this digital filter is shown in Figure 5.2. Note that at the lower frequencies, the spectra of the jitter filter and actual jitter signals agree quite well. A synthetic jitter signal is created by passing the Gaussian white noise through the filter. A typical resulting synthetic jitter signal is shown in Figure 5.3. Comparing the synthetic signal to the real signal of Figure 5.1, we can see that the filtered noise does a good job at accurately approximating the general behavior of the jitter signal.

5.2.3 Implementation

Since the glottal waveform is stored and processed in the time domain and frequency changes are synthesized simply by changing how the glottal waveform is read from memory, the effects of jitter are relatively easy to implement in the DIDSS system.

The fundamental frequency of the output waveform depends on the rate it is read from memory. If the waveform was originally recorded with an F0 of 160 Hz, then reading every
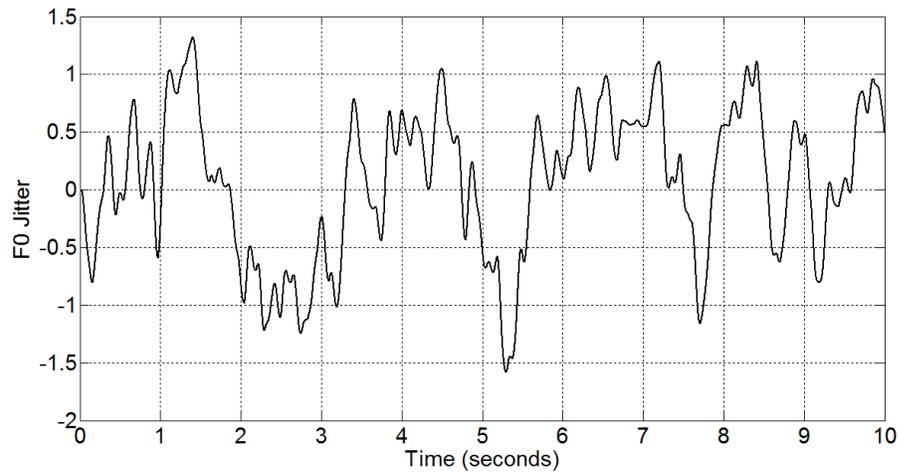
Fig. 5.3: Synthetic jitter signal (centered at 0 Hz).

sample out of memory at the rate it was recorded would yield an output signal with an F0 of 160 Hz. If every other sample were skipped (if the index were incremented by 2 every time instead of 1), but the data was still read out at the original sampling rate, the output signal would have an F0 of 320 Hz.

The indexing rate is directly proportional to the output F0. If the original signal had an F0 of 160 Hz, then each increment of 1 in the index equates to an increment of 160 Hz in the output waveform. If it is desired to increase the frequency of the output waveform from 160 to 165 Hz (an increase of 5 Hz), then the index must be incremented at a rate of $(1 + \frac{5}{160})$. So if the frequency of the output waveform is desired to jitter by ± 2 Hz, then the index must vary by $\pm \frac{2}{160}$.

5.3   Shimmer

Shimmer is the term for the aperiodic variations in amplitude of the glottal waveform. Because the amplitude of any waveform changes from sample to sample, it does not make sense to talk about the instantaneous amplitude of the signal. The concept of the root mean square (RMS) value is thus used, which gives a good estimate of the average power of the signal over a given time period. The RMS value of a vector $\mathbf{x}$ is estimated by

$$\mathbf{x}_{RMS} = \sqrt{\frac{(x_1^2 + x_2^2 + ... + x_n^2)}{n}} \ , \ \mathbf{x} = (x_1, x_2, ..., x_n).$$

The following sections will outline the process of creating a model that will accurately model shimmer signals for an individual person.

5.3.1   Characterization

In order to track the changes in the RMS value over time, the vocalized speech was split into overlapping frames, and the RMS value was computed for each frame. Each frame was 50 ms long, and each consecutive frame was shifted by an increment of 1 ms. At a frequency of 160 Hz a single cycle takes about 6.25 ms, so the frame size is roughly 8 cycles while the frame is shifted each time by an increment of about 16% of 1 cycle.

Each RMS estimate encapsulates multiple glottal cycles for the same reason as the jitter estimation method. However, it was found that using a smaller frame size and much smaller frame increment size revealed finer detail in the shimmer signal that seemed important. At these rates some high frequency artifacts start to appear, which are believed to be related to the moving window and not related to the actual shimmer signal. Figure 5.4 shows the shimmer signal for the vocalized /i/ phoneme at high vocal effort. Figure 5.5 shows a closer look at a noisier portion of the shimmer signal. These high frequency artifacts do not appear to be random, and are assumed not to be a part of the actual shimmer signal.
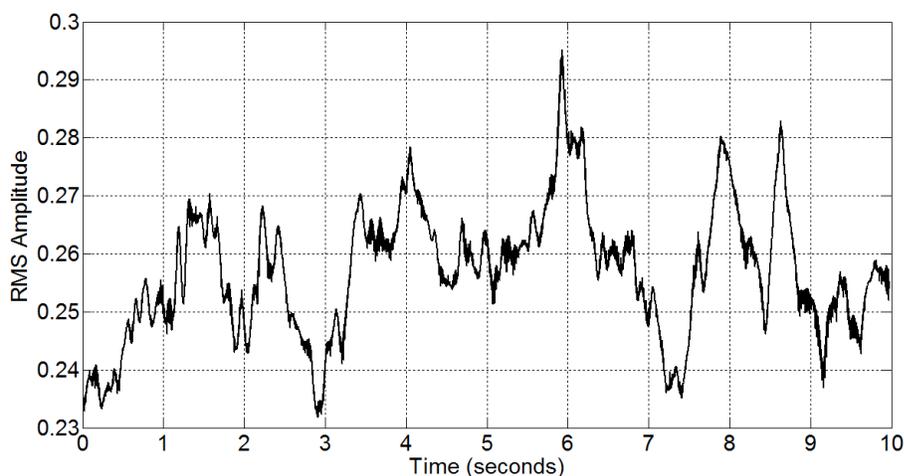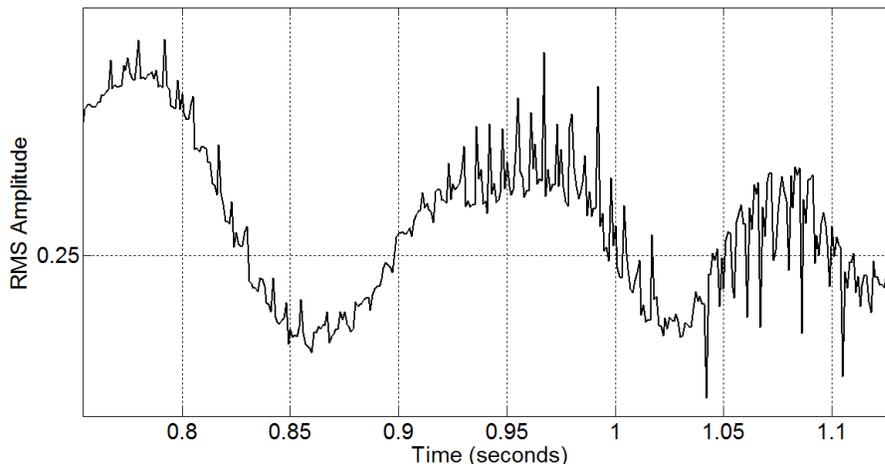

Fig. 5.4: Shimmer signal for /i/ at high vocal effort.

Fig. 5.5: Supposed artifacts in shimmer signal.

5.3.2 Synthesis

It is thought that a more natural-sounding shimmer signal will be realized by modeling the shimmer signal without these high frequency features. The argument for this is the same as given in the jitter section, namely, that the high frequency fluctuations would start to represent a vocal quality of hoarseness instead of the slower fluctuations in amplitude typically thought of as shimmer.

A shimmer filter was designed using the same method as with the jitter filter. It was found that a good shimmer approximation could be reproduced by processing Gaussian white noise through a cascade of two low pass filters.

The first filter is a second order Butterworth low pass filter, passing frequencies up to 0.2 Hz, and attenuating frequencies past 50 Hz by at least 30 dB. The filter coefficients are as follows:

$$\text{Numerator} = [2.799895233e\text{-}05, 2.799895233e\text{-}05, 0],$$
$$\text{Denominator} = [1, -0.999944002, 0].$$

The second filter is a third order Butterworth low pass filter, passing frequencies up to 9.5 Hz, and attenuating frequencies past 50 Hz by at least 30 dB. The filter coefficients are as follows:

$$\text{Numerator} = [6.081087171e\text{-}10, 1.824326151e\text{-}09, 1.824326151e\text{-}09, 6.081087171e\text{-}10],$$

$$Denominator = [1,-2.996609225,2.993224196,0.9966149661].$$

The two filters are cascaded together into a new fifth order filter which is used to synthesize the shimmer signal. The frequency response of this digital filter overlayed on the average spectrum of the actual shimmer signals is shown in Figure 5.6. The synthetic shimmer signal is created by passing the Gaussian white noise through the filter. A typical resulting synthetic shimmer signal is shown in Figure 5.7. Comparing the synthetic signal to the real signal of Figure 5.4, we can see that the filtered noise does a good job at accurately approximating the shimmer signal, without the high frequency artifacts attributed to the shimmer estimation method.

5.3.3   Implementation

The effects of shimmer will be replicated by appropriately scaling the audio output of the DIDSS system with the shimmer signal. The shimmer signal is a measure of the estimated RMS value fluctuation at any instant. The property of RMS estimation that

$$RMS(\alpha * \mathbf{x}) = \alpha * RMS(\mathbf{x})$$

indicates that the audio signal should be multiplied by the shimmer signal somehow in order to reproduce the shimmer effect. First, the audio signal should have its RMS value normalized by dividing the signal by its RMS value (which is known ahead of time). Then, the signal should be scaled by the shimmered RMS value, which is produced by adding the shimmer signal to the signal's typical RMS value. Note that as the shimmer approaches zero, the output waveform approaches its default amplitude. The following equation outlines the process.

$$\mathbf{x}_{shimmered} = \mathbf{x} * \frac{RMS(\mathbf{x}) + shimmer}{RMS(\mathbf{x})}$$

5.4   Comparison of Jitter and Shimmer

After the jitter and shimmer signals had been characterized, it was desired to know if the two phenomena were correlated. It was hypothesized that they could be, because of
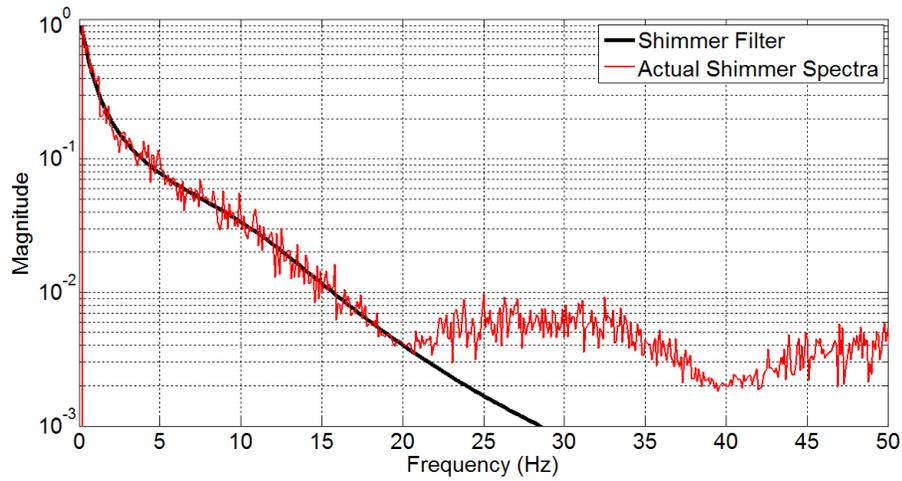
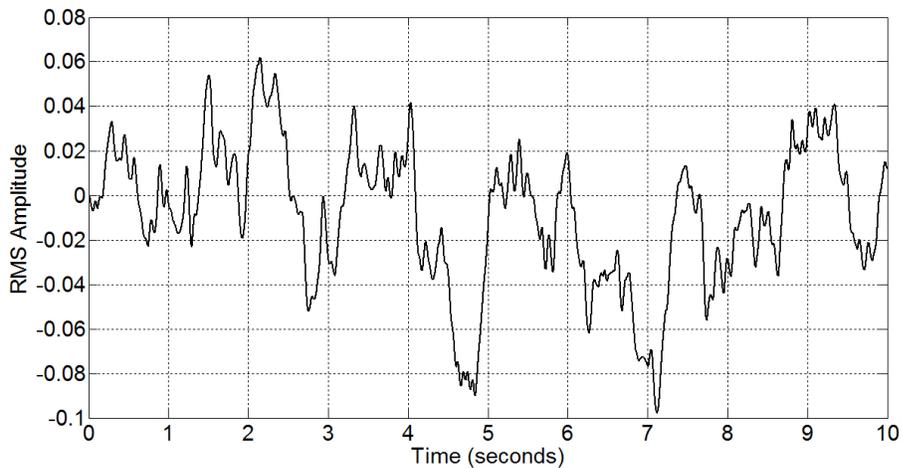Fig. 5.6: Shimmer filter frequency response vs. actual shimmer spectra.

Fig. 5.7: Synthetic shimmer signal (centered at 0).

the fact that they are caused by effects of the human body which may be related to one another. Because the jitter and shimmer properties are so similar, it was desired to create a simpler synthesis system consisting of a single noise signal passed through a single, low-order, low pass filter. If the jitter and shimmer signals are correlated, it is supposed that this correlation would be an important factor to consider when attempting to accurately synthesize these effects of the human system. A correlation between the two signals would be synthesized by making two copies of the filtered noise, one for jitter and one for shimmer. By delaying the jitter signal from the shimmer signal (or vice-versa) by a certain amount,

the desired amount of correlation can be achieved.

5.4.1 Correlations of Jitter and Shimmer

The data for the jitter and shimmer correlations were sampled at 100 Hz and normalized before the correlations were computed. Figure 5.8 shows the autocorrelation plots for the jitter and shimmer signals. The jitter autocorrelation signal shown is an average of the nine autocorrelation results derived from the nine jitter signals. The same is true for the shimmer autocorrelation plot. It was found that the maximum correlation value between the jitter and shimmer signals was roughly 21.53. This value occurred at a delay of nine samples, which corresponds to 90 ms of delay between the signals. It is supposed that in actuality the delay would not be so high, but is so in this case because the moving windows for jitter and shimmer analysis were of significantly different size. In any case, it does not matter what delay the signals were highest correlated at, but rather the maximum value of their correlation. The delay of the synthesized signals will be determined by this maximum correlation value.

Figure 5.9 shows the cross-correlation plots for the jitter and shimmer signals. The cross-correlation signal shown is an average of the nine cross-correlation results derived from the nine jitter and shimmer signals. The plot looks rather noisy, and it is thought that a better results would be achieved if more recordings were available to analyze and contribute to the average.

5.4.2 Creation of a Simple Jitter/Shimmer Filter

A simple low pass filter was created which was a compromise between the jitter and shimmer filters, with emphasis on simplicity. The filter is a second order Butterworth low pass filter with a passband frequency of 1.1 Hz, a stopband frequency of 200 Hz, and a stopband attenuation of 40 decibels. The filter coefficients are as follows:

$$\text{Numerator} = [1.206571689263108\text{e-}08, \ 2.413143378526217\text{e-}08, \ 1.206571689263108\text{e-}08],$$
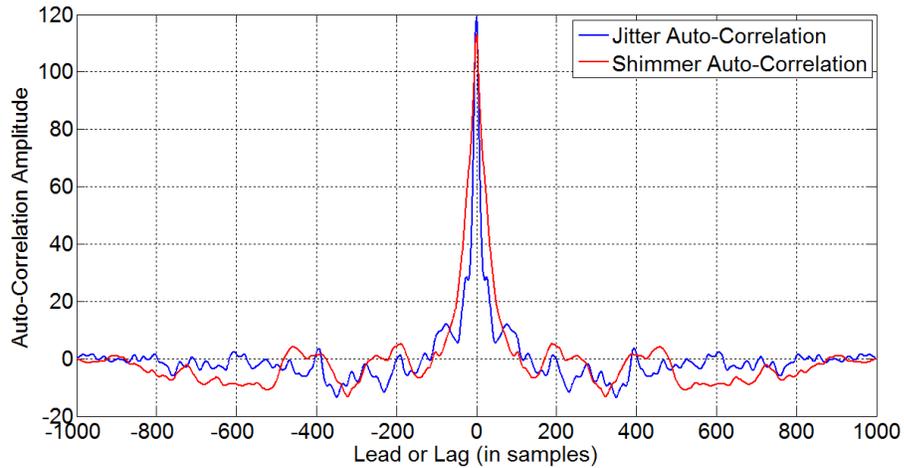$$\text{Denominator} = [1, \ -1.999689289958749, \ 0.999689338221617].$$

Fig. 5.8: Average auto-correlations of the jitter and shimmer signals.
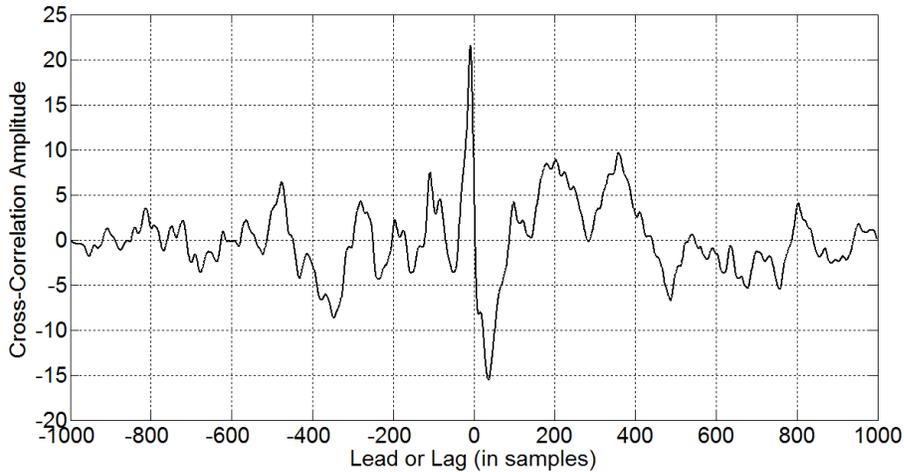

Fig. 5.9: Average cross-correlation of the jitter and shimmer signals.

Figure 5.10 shows the frequency response of this new filter along with the actual jitter and shimmer spectra. Figure 5.11 shows the output of this new filter.

5.4.3 Synthesizing Correlation with the New Filter

White noise was passed through the new filter and resampled at a rate of 100 Hz, and then the autocorrelation was computed. This process was repeated two hundred times, and the autocorrelation results were averaged together. Figure 5.12 shows the averaged autocorrelation for the output of the new filter.

Recall that the desired correlation value between the jitter and shimmer signals was
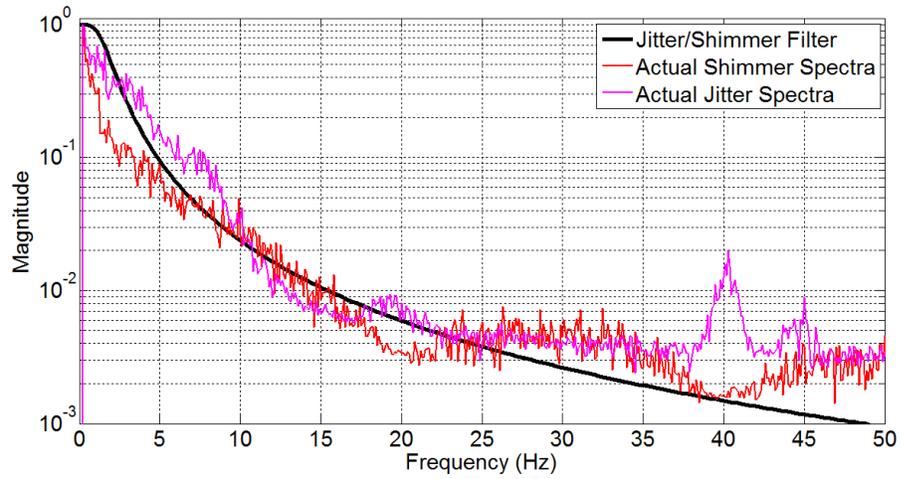
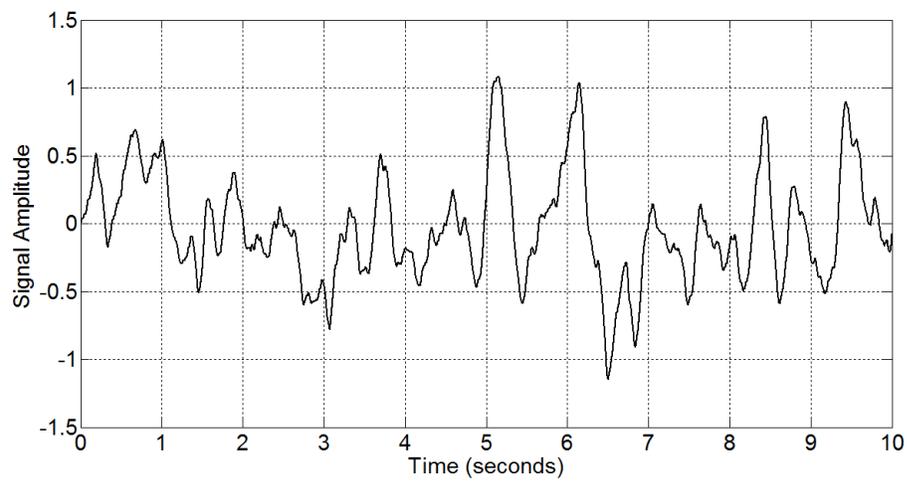Fig. 5.10: Jitter/shimmer filter vs. jitter and shimmer spectra.

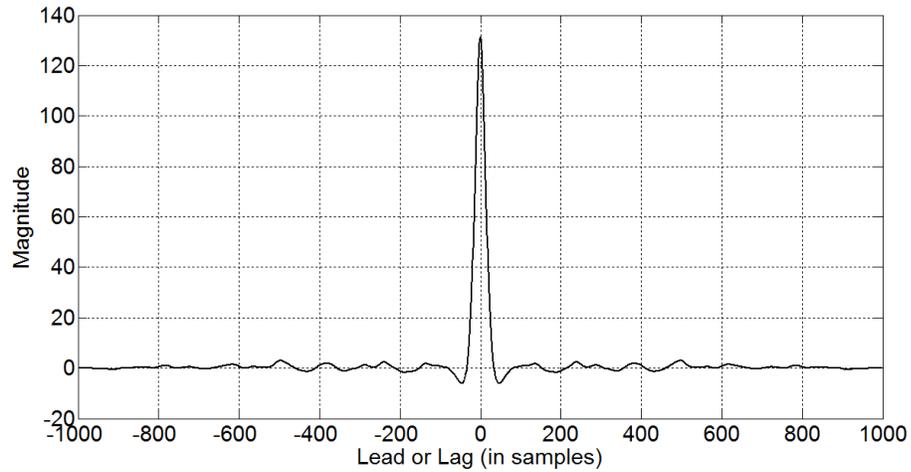Fig. 5.11: Synthesized jitter/shimmer signal.

Fig. 5.12: Autocorrelation for the new filter output.

21.53. On the autocorrelation plot of the new filter output, the correlation value of 21.53 was found to lie somewhere in between delays of 24 and 25 samples, corresponding to a time delay of roughly 245 milliseconds. So, in order to synthesize the proper correlation between the jitter and shimmer signals, one should be delayed by the other by 245 milliseconds. It does not matter which one is delayed, since the signals are identical and the autocorrelation signal is symmetric.

Chapter 6

Results

6.1 Glottal Waveform Extraction Using a White Noise Source

Figure 6.1 shows the spectrum of the /u/ phoneme vocalized at high vocal effort. Also shown is the estimate of the /u/ derived from the noise recording. It can be seen that the estimate from the noise recording does not follow closely to the actual formants of the vocalized phoneme. Figure 6.2 shows the spectrum of the vocalized /u/ phoneme after correction by its corresponding vowel inverse filter developed in Chapter 3. Large peaks and dips are still visible in the spectrum, indicating that an accurate glottal source has not been derived. Informal subjective listening tests confirm this, as the derived glottal source recording still sounds as if a vowel is being spoken.

In a previous study [12], M. Erickson and E. D'Alfonso introduced a periodic buzzing audio source into the oral cavity of an individual via a tube in order to characterize their vocal tract. Their results were also not favorable, and were not as accurate as traditional methods, even for high-pitched voices. They acknowledged the effects that the tube would have on the spectral estimates but did not attempt to correct for the effects of the tube.

The results of this report take the study one step further. Because a white noise source is used, the effects of the tube and other components were readily characterized and taken into account. However, like the study by Erickson and D'Alfonso, the final results were still not as good as traditional methods. While Erickson and D'Alfonso attributed the failure of their method largely to the effects of the tube, the same conclusion cannot be made for this report.

Two possible explanations remain obvious. The first possibility is that the formants of the vocal tract are extremely dependent on the placement of the articulators, and that, between the vocalized and the noise recordings, they moved enough to significantly alter
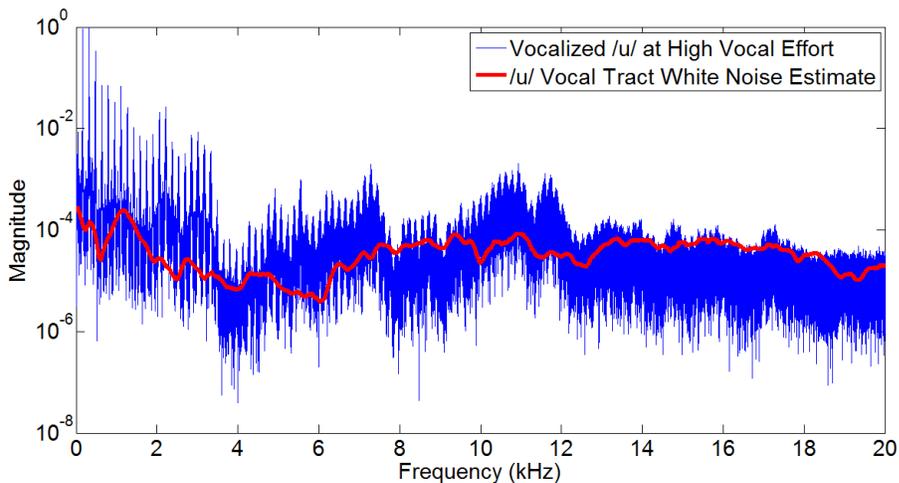
Fig. 6.1: Vocalized /u/ vs. /u/ white noise estimate.

the formants. However, much care was taken to ensure the same articulator position during all recordings for each phoneme, so it is thought that this is not the primary reason for failure of the noise characterization.

The second explanation is that the glottal source and the vocal tract are so strongly coupled that one cannot be accurately characterized if the other is removed from the process. It is already known that the glottal source and vocal tract are significantly coupled [1], and so it is this explanation which is assumed to be the primary reason behind the ineffectiveness of the white noise characterization method.

Cepstral analysis was chosen as a substitute for the noise method in order to extract a decent glottal source waveform from the vocalized recordings. Figure 6.3 shows the spectrum for the /u/ phoneme vocalized at high vocal effort along with the cepstral estimate of the formants of the vocal tract. We can see in this case that the formant estimates line up nicely with the actual formants, and in Figure 6.4 we see that the resulting spectrum for the glottal waveform looks very well-corrected from the effects of the formants.

6.2 Subjective Listening Test

The glottal waveform derived via cepstral analysis is not perfect. This is confirmed with the results from Chapter 4, where the three glottal waveforms derived for each level of vocal
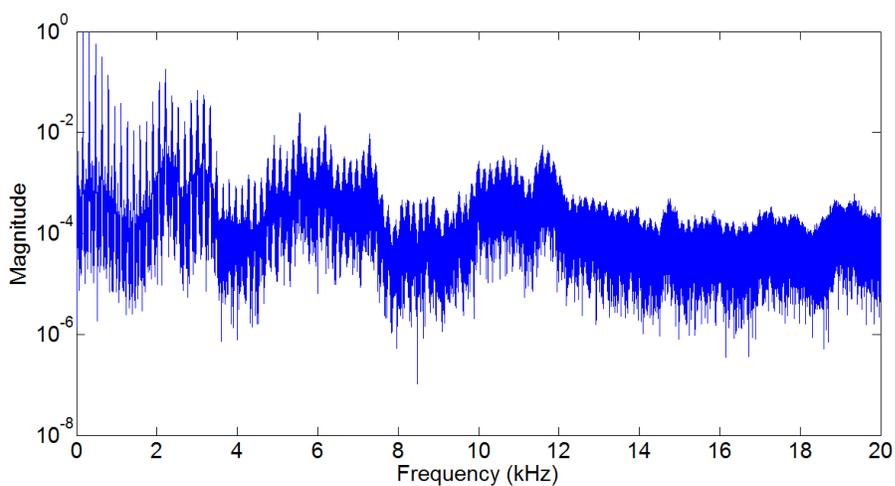
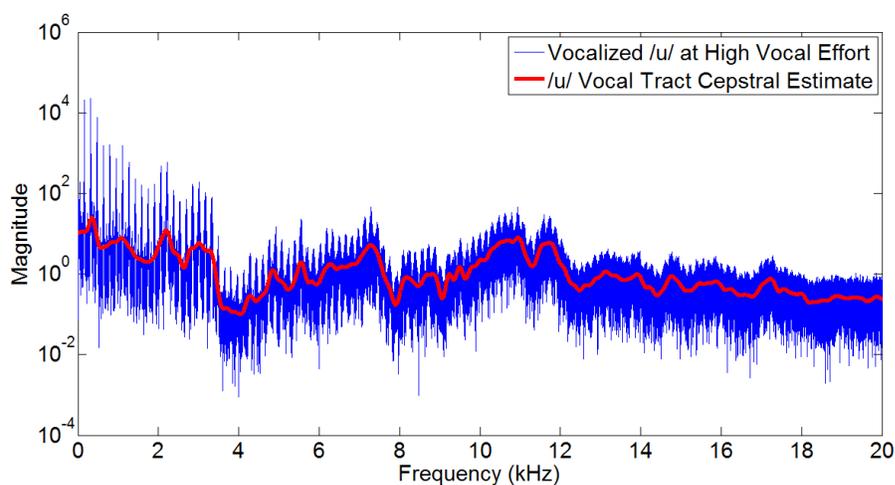Fig. 6.2: Vocalized /u/ after correction by the /u/ white noise estimate.

Fig. 6.3: Vocalized /u/ vs. /u/ cepstral estimate.

effort have quite different properties. However, the results are deemed to be good enough, and one of the glottal waveforms will later be chosen as the source for implementation in the DIDSS system. The effects of jitter and shimmer were successfully synthesized, and they will be discussed in the following sections.

6.2.1 Jitter

The inclusion of the jitter effect had a very significant impact on the naturalness of the synthesized voice. Without it, the voice seemed very unnaturally steady and machine-like.
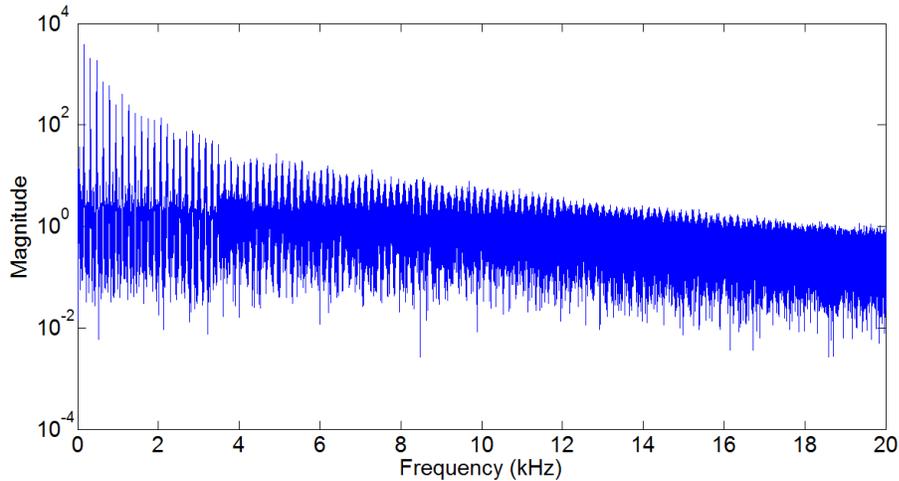
Fig. 6.4: Vocalized /u/ after correction by the /u/ cepstral estimate.

In comparison to actual sustained vocalizations, the synthesized frequency jitter seemed to happen more slowly than the natural amount of jitter. It is assumed that this may be caused by the chosen frequency estimation method, which could not estimate faster jitter very well. Before implementation into the DIDSS system, the jitter model will be modified to more accurately mimic the natural jitter of the human voice.

6.2.2 Shimmer

It was found that the effects of shimmer, when synthesized at a realistic level, are much more subtle than the effects of jitter. This confirms previous findings by other groups [15]. However, the incorporation of shimmer did allow for a synthesized source which sounded more natural. The synthetic amplitude shimmer is thought to be sufficiently characterized, and it is not thought that the model needs significant modification before incorporation into the DIDSS system.

6.2.3 Changes in Glottal Waveform Harmonic Content

Although a sufficient model of glottal harmonic change was not produced, these changes do have a significant effect on the quality of the synthesized voice. When the three recordings (one recording per level of vocal effort) for each vocalized phoneme are normalized to the

same level and compared, the difference in timbre is easily audible. In the future it is desired to synthesize this effect, because of the fact that it is a clearly distinguishable property of the natural voice.

Chapter 7

Conclusion

A white noise audio source was introduced into the oral cavity of an individual and used to characterize the frequency response of his vocal tract. That frequency response was smoothed, corrected for effects of the system, and then used in an attempt to filter vocalized recordings to extract an approximation of the glottal source waveform.

The results were poor in comparison to cepstral analysis, which was used instead. The best hypothesis as to the poor results is that the white noise method assumes that the properties of the vocal tract are time-invariant and independent of the glottal source, neither of which are completely true. These results can be instructive as to the limitations of the source-filter model for glottal waveform extraction.

Harmonic analysis was attempted on the extracted glottal waveforms, but a solid conclusion could not be reached due to differences in the results for the various vocalized recordings.

Jitter and shimmer analysis was performed on the vocalized recordings. Models were created and tested for jitter and shimmer synthesis, with very good results. They do not sound "noisy," as has been the problem with other jitter and shimmer synthesis attempts by other groups. The correlation between the jitter and shimmer signals was also analyzed. It is thought that this data could be significant, but that the number of recordings analyzed needs to be larger in order to make more accurate inferences from the result.

References

[1] P. Taylor, *Text-to-Speech Synthesis*. New York: Cambridge University Press, 2009.

[2] A. Hunt and J. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of the International Conference on Speech and Language Processing*, 1996.

[3] C. Ruhlemann, "Coming to terms with conversational grammar: 'dislocation' and 'disfluency'," *International Journal of Corpus Linguistics*, vol. 11, pp. 385–409, 2006.

[4] H. Dudley and T. H. Tarnoczy, "The speaking machine of Wolfgang von Kempelen," *Journal of the Acoustical Society of America*, vol. 22, pp. 151–166, 1950.

[5] H. Dudley, "Remaking speech," *Journal of the Acoustical Society of America*, vol. 11, pp. 169–177, 1939.

[6] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: Control and conversion," *Speech Communication*, vol. 16, pp. 165–173, 1995.

[7] G. Fant, *Acoustic Theory of Speech Production*. The Hague, Netherlands: Mouton & Co., 1960.

[8] B. Atal and S. Hanauer, "Speech analysis and synthesis by linear prediction of the speech waveform," *Journal of the Acoustical Society of America*, vol. 50, pp. 637–655, 1971.

[9] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, pp. 221–228, 1968.

[10] A. V. Oppenheim, *Discrete-Time Signal Processing (3rd Edition)*. New Jersey: Prentice Hall, 2009.

[11] Marshall Electronics. (2012) Mxl 993 pencil condenser microphone specifications. [Online]. Available: http://www.mxlmics.com/microphones/900-series/993/

[12] M. Erickson and A. D'Alfonso, "A comparison of two methods of formant frequency estimation for high-pitched voices," *Journal of Voice*, vol. 16, pp. 147–171, 2002.

[13] J. Holmes and W. Holmes, *Speech Synthesis and Recognition, 2nd ed.* New York: Taylor and Francis, 2001.

[14] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Berlin: Springer, 2008.

[15] A. J. Rozsypal and B. F. Millar, "Perception of jitter and shimmer in synthetic vowels," *Journal of Phonetics*, vol. 7, pp. 918–924, 1978.

[16] J. Schoentgen, "Stochastic models of jitter," *Journal of the Acoustical Society of America*, vol. 109, pp. 1631–1650, 2001.

[17] D. Ruinskiy and Y. Lavner, "Stochastic models of pitch jitter and amplitude shimmer for voice modification," *Electrical and Electronics Engineers in Israel, IEEE 25th Convention*, pp. 489–493, 2008.

[18] N. Aoki and T. Ifukube, "Two 1/f fluctuations in sustained phonation and their roles on naturalness of synthetic voice," *Electronics, Circuits, and Systems, Proceedings of the Third IEEE International Conference*, pp. 311–314, 1996.