Exploiting Pitch Dynamics for Speech Spectral Estimation Using a Two-Dimensional Processing Framework

by

Tianyu Tom Wang

B.S., Electrical Engineering
Georgia Institute of Technology, 2005

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2008

Signature of Author .....
Department of Electrical Engineering and Computer Science
February 1, 2008

Certified by .....
Thomas F. Quatieri
Senior Member of Technical Staff; MIT Lincoln Laboratory
Faculty of Speech and Hearing Bioscience and Technology Program;
Harvard-MIT Division of Health Sciences and Technology
Thesis Supervisor

Accepted by .....
Prof. Terry P. Orlando
Chair, Department Committee on Graduate Students
Department of Electrical Engineering and Computer Science

This work was supported by the Department of Defense under Air Force contract FA8721-05-C-0002. The opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government. The work of T.T. Wang was additionally supported by the National Institutes of Deafness and Other Communicative Disorders under grant 5 T 32 DC00038.

Exploiting Pitch Dynamics for Speech Spectral Estimation Using a Two-Dimensional Processing Framework

by

Tianyu Tom Wang

B.S., Electrical Engineering
Georgia Institute of Technology, 2005

Submitted to the Department of Electrical Engineering and Computer Science
on February 1, 2008 in partial fulfillment of the
Requirements for the Degree of Master of Science in
Electrical Engineering

Abstract

This thesis addresses the problem of obtaining an accurate spectral representation of speech formant structure when the voicing source exhibits a high fundamental frequency. Our work is inspired by auditory perception and physiological modeling studies implicating the use of temporal changes in speech by humans. Specifically, we develop and evaluate signal processing schemes that exploit temporal change of pitch as a basis for high-pitch formant estimation. As part of our development, we assess the source-filter separation capabilities of several two-dimensional processing schemes that utilize both standard spectrographic and auditory-based time-frequency representations. Our methods show quantitative improvements under certain conditions over representations derived from traditional and homomorphic linear prediction. We conclude by highlighting potential benefits of our framework in the particular application of speaker recognition with preliminary results indicating a performance gender-gap closure on subsets of the TIMIT corpus.

Thesis Supervisor: Thomas F. Quatieri

Title: Senior Member of Technical Staff; MIT Lincoln Laboratory

Faculty of Speech and Hearing Bioscience and Technology Program;

Harvard-MIT Division of Health Sciences and Technology

Acknowledgements

First and foremost I would like to thank my advisor, Tom Quatieri, without whom this thesis would not have been possible. In the last two years, I have greatly appreciated Tom's dedication to his students and his ability to bestow upon us a passion for the science behind the engineering. Thanks, Tom, for pushing me to think deeply about my research and for your patience as I worked through the "homework problems" you often posed that led to critical insights for this thesis.

I would like to thank Cliff Weinstein and all of the members of the Speech Group in Group 62 at MIT Lincoln Laboratory for providing a supportive and constructive environment for students to conduct thesis research. Many thanks to Doug Sturim for helping me understand the intricacies of the speaker recognition system and noting the gender performance gap. I would also like to thank James Glass of the Spoken Language Systems group at MIT for allowing me to use the group's computing resources during preliminary parts of this work.

To my official and unofficial labmates – Nicolas Malyska and Daryush Mehta for jazz_hour.pcm and valuable advice, Zahi Karam for many coffee walks and whiteboard discussions, Nancy Chen for the lengthy discussions on the shuttle ride regarding the future of speech research, and Nick Loomis for being a great friend and roommate. I would especially like to thank Zahi and Nicolas for taking time out of their schedules to give me a ride on the many occasions I stayed late in lab.

And to Mom, Dad, Lao Ye, and Jenn – your love makes everything I do worthwhile. Thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A major goal of speech analysis is to infer from a speech waveform the underlying anatomical characteristics and physiological mechanisms generating it. Existing analysis methods typically operate on short-time scales and on a frame-by-frame basis (e.g., the short-time Fourier transform). A major limitation of this framework is the inability of analysis techniques to exploit temporal changes of speech across time. In this thesis, we characterize the properties of a two-dimensional (2-D) processing framework that aims to explicitly exploit such changes and assess its value in addressing an existing problem in speech analysis. In addition, we present preliminary results of adopting this framework in the particular application of speaker recognition.

1.1 Problem Statement and Motivation

A well-known problem in speech analysis is that of estimating the formant frequencies of a speaker during vowel utterances, thereby inferring the speaker's vocal tract configuration. To this end, standard techniques such as linear prediction and cepstral analysis operate on short-time speech spectra and have been shown to provide reasonable estimates under certain conditions [1, 2]. Nonetheless, it has been shown that these methods are not generally robust to the condition of *high-pitch* speakers (e.g., females, children) [3, 4]. In this work, we are inspired by psychophysical evidence implicating the use of temporal changes in speech in human speech perception as a basis for addressing the high-pitch formant estimation problem. For instance, McAdams showed in a series of concurrent vowel segregation tasks that subjects reported an increased "prominence" percept for vowels whose pitch was modulated relative to those that were not modulated [5]. Diehl, et al. showed in vowel perception experiments that a linearly changing pitch improved subjects' vowel identification accuracy [6]. In both studies, the observed effects were greatest when the synthetic source was chosen to have a high pitch (e.g., ~250-400 Hz).

Our approach to address the high-pitch formant estimation problem may also have implications for the speaker recognition application. In speaker recognition, it has been observed that state-of-the-art systems exhibit a "gender gap": system performance is higher for male data sets than females [7]. A possible cause of this gap may be that the formant structure of the relatively higher-pitch females is poorly represented in short-time Fourier analysis. An improved spectral representation as characterized by its accuracy in formant estimation may therefore address this gap.

1.2 Approach

Concurrent with the psychophysical literature, recent physiological modeling efforts of the auditory system have implicated its use of acoustic information spanning longer durations than is typically used in traditional short-time speech analysis. In particular, Chi, et al. have proposed a comprehensive model of auditory processing [8] that views the low-level periphery as a means to generate a time-frequency distribution of an acoustic stimulus. While this aspect is similar to traditional short-time analysis, subsequent high-level cortical processing is modeled as analyzing spectrotemporal regions of this time-frequency distribution. This model is therefore capable of exploiting temporal changes in speech.

The auditory model proposed by Chi, et al. may be viewed as one realization of a generalized processing framework employing *any* input time-frequency distribution followed by *any* 2-D transform. In addition to the auditory model, this thesis characterizes several realizations of this more general framework including a simple method of harmonic projection and the Grating Compression Transform proposed by Quatieri [9] and later extended by Ezzat, et al. [10]. Our analysis motivates several approaches to addressing the high-pitch formant estimation problem which are then evaluated in relation to traditional techniques. In addition, we employ this generalized 2-D processing framework in preliminary speaker recognition experiments to assess its value in addressing the gender gap problem.

1.3 Summary of Contributions

This thesis provides a thorough phenomenological analysis of three realizations of a 2-D speech processing framework: (pitch) harmonic projection, the Grating Compression Transform (GCT) [9], and a comprehensive model of auditory signal processing ending in the auditory cortex [8]. Our aim in this analysis is to motivate an improved method of high-pitch formant estimation when pitch is changing and the formant envelope is assumed to be stationary. Under these conditions, we have observed that harmonic projection improves the spectral sampling of the formant envelope while the GCT invokes improved source-filter separability in a 2-D modulation space. These observations are also justified analytically. While the auditory model invokes a similar 2-D modulation space and highlights formant structure and pitch in several ways, we were unable to argue for the same type of source-filter separability as in the GCT. Our observations motivate several methods of deriving speech spectra for use in conjunction with linear prediction to improve high-pitch formant estimation. We present a methodology for evaluating these methods on synthesized speech in relation to standard methods. The results of our evaluation show that exploiting temporal change of pitch can provide improved estimates of formant frequencies, even under conditions of high pitch. Finally, this thesis illustrates the feasibility of employing the generalized 2-D framework for speaker recognition with promising results in relation to a gender performance gap in this particular application.

1.4 Thesis Outline

This thesis is organized as follows. Chapter 2 provides background on the high-pitch formant estimation problem and discusses existing analysis methods. Chapter 3 motivates the framework proposed for improving spectral estimation of formant structure by providing a rigorous phenomenological comparison of the projection of pitch harmonics, the GCT, and the auditory

cortex model. Chapter 4 gives the methodology used in assessing this framework for formant estimation while Chapter 5 discusses the results of our evaluation. In Chapter 6, we discuss our methodology and presents preliminary results in adopting our framework in speaker recognition. Finally, we provide conclusions and highlight areas for future work in Chapter 7.

Chapter 2

Undersampling in Speech Spectral Estimation

In this chapter, we provide background for deriving spectral estimates of high-pitch speech. We begin in Section 2.1 by discussing the source-filter model of speech production which we adopt for the entirety of this thesis. Using this model, we discuss in Section 2.2 short-time Fourier analysis of high-pitch speech waveforms. Specifically, we provide a spectral-domain view of high-pitch speech referred to as *spectral undersampling*. In Section 2.3, we discuss the effects of spectral undersampling on formant estimation via existing methods, thereby highlighting their limitations.

2.1 Source-filter model

The speech production system can be viewed as a pressure source (i.e., the lungs), vibrating mechanism (the larynx), and time-varying linear filter (the vocal tract) in series [11]. In this thesis, we are concerned with vowel sounds generated from this system at different "pitch" values. Specifically, the lungs first generate a pressure differential across the glottis of the larynx that results in periodic vocal fold vibrations. The periodicity T_0 of these vibrations is determined by the mechanical properties of the vocal folds themselves, with more (less) massive folds corresponding to longer (shorter) periods [11]. The vibrations of the larynx subsequently excite the acoustic resonances (i.e., formants) of the vocal tract that are determined by its configuration. Finally, the resulting vowel sound is radiated from the lips.

We adopt a discrete-time model of the speech signal, $s[n]$, based on the previously discussed physiological production mechanisms. For simplicity, we refer to the lungs and larynx as the "source" and the vocal tract as the "filter". For short durations (e.g., 20 ms), the linear filter can be assumed to be time invariant such that $s[n]$ is the time-domain convolution (denoted by *) between the source, $g[n]$, and the impulse response of the filter, $h[n]$:

$$s[n] = g[n] * h[n]. \tag{2.1}$$

$g[n]$ models the vibrations of the larynx and is a periodic impulse train with fundamental frequency $f_0 = 1/T_0$. For vowels, $h[n]$ characterizes the formant frequencies of the corresponding vowel throughout its duration. In the current time-domain framework, $s[n]$ can be viewed as overlapping copies of the impulse response $h[n]$ occurring with periodicity T_0. Figure 1 shows the convolution between $h[n]$ corresponding to the vowel */ae/* and a pure impulse train $g[n]$ with $T_0 = 8$ ms ($f_0 = 125$ Hz) (top) and $T_0 = 3.8$ ms ($f_0 = 260$ Hz)

(bottom) for a 25-ms duration. The amount of overlap between copies of $h[n]$ is determined by the value of T_0, with smaller T_0 leading to more overlap. As will be subsequently discussed, this increased overlap is one manifestation of high-pitch formants and leads to poor results in formant estimation using traditional methods.

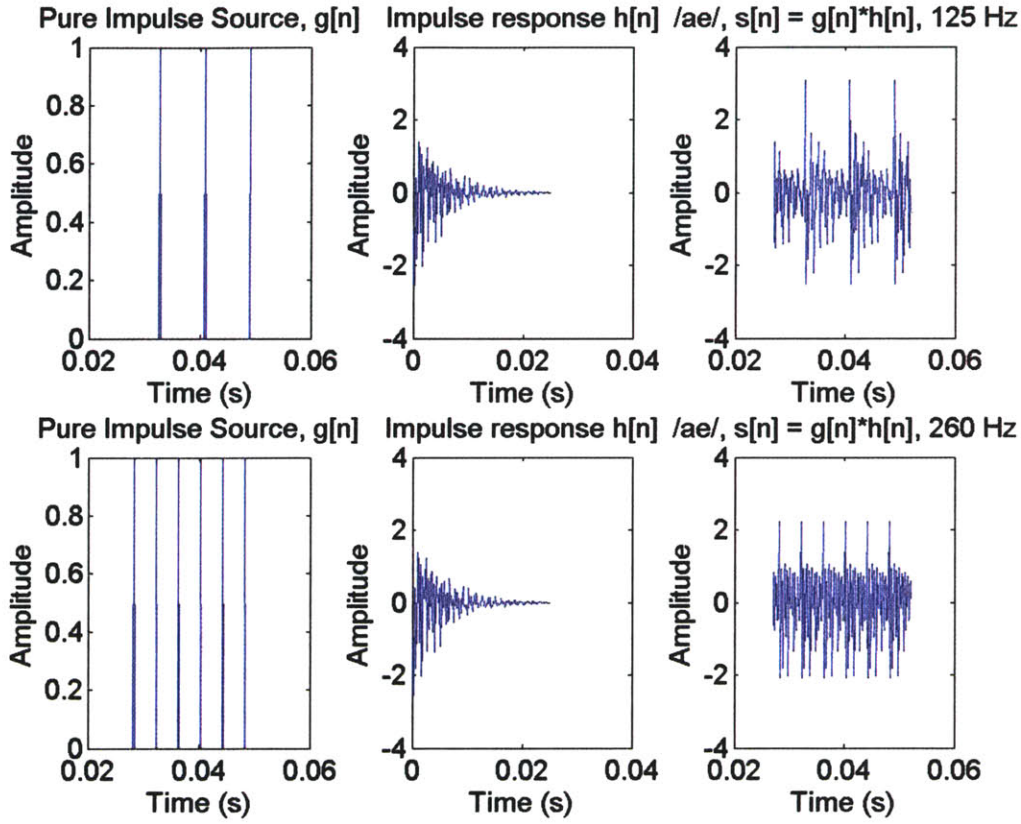

Figure 1 - Time-domain representation of source-filter model for the vowel */ae/*. For $f_0 = 260$ Hz (bottom), there is significantly more overlap of the underlying impulse response $h[n]$ than for $f_0 = 125$ Hz (top) in the resulting synthesized vowel $s[n]$.

2.2 Spectral Undersampling

We turn now to a spectral-domain view of the source-filter model. Speech is typically analyzed using short-time Fourier analysis. Specifically, the short-time Fourier transform (STFT) $X[n,\omega]$ is defined as [4]

$$X[n,\omega] = \sum_{m=0}^{L-1} x[n+m]w[m]e^{-j2\pi\omega m} \tag{2.2}$$

where $x[n]$ corresponds to the full-duration signal, $w[n]$ is a window spanning $0 \le n \le L-1$ and zero elsewhere, and n and ω correspond to the time and frequency indices, respectively.

Typically, $w[n]$ is chosen to have sufficiently short duration such that the assumption of a stationary vocal tract ($h[n]$) is valid (e.g., 20 ms).

For a single spectral slice, $X[n = n_0, \omega]$, $x[n_0 + m]$ may be substituted by $s[m]$ from (2.1) such that (2.2) is the Fourier transform of $(h[m] * g[m])w[m]$:

$$X[n_0, \omega] = \sum_{m=0}^{L-1} (h[m] * g[m]) w[m] e^{-j2\pi\omega m} . \tag{2.3}$$

In the Fourier domain, $X[n_0, \omega]$ can be rewritten as a frequency-domain convolution ($*_\omega$) between the Fourier transform of the window $W(\omega)$ and the product of Fourier transforms of $h[n]$ and $g[n]$ ($H(\omega)G(\omega)$):

$$X[n_0, \omega] = \frac{1}{2\pi} H(\omega)G(\omega) *_\omega W(\omega) . \tag{2.4}$$

Consider the case when $g[n]$ is a pure impulse train with periodicity $P = f_s / T_0$, where f_s corresponds to the sampling frequency such that:

$$\begin{aligned} g[n] &= \sum_{k=-\infty}^{\infty} \delta[n - kP] \\ G(\omega) &= \frac{1}{P} \sum_{k=-\infty}^{\infty} \delta(\omega - \omega_k) \end{aligned} \tag{2.5}$$

where $\omega_k = \frac{2\pi}{P}$. Substituting (2.5) into (2.4), we obtain

$$X[n_0, \omega] = \frac{1}{P} \sum_{k=-\infty}^{\infty} H(\omega_k) W(\omega - \omega_k, n_0) . \tag{2.6}$$

Assuming now that $w[n]$ is chosen such that there is little or no interaction between neighboring copies of $W(\omega)$, the periodic replicas of $W(\omega)$ can be viewed as sampling the underlying formant envelope $H(\omega)$ with sample spacing of $\frac{2\pi}{P}$ in the spectral domain. We then refer to Equation (2.6) as a narrow-band spectrum [3].

Figure 2 (top) shows the components of such a narrow-band spectrum of a pure impulse source with $f_0 = 125$ Hz convolved with $h[n]$ corresponding to the impulse response of the female vowel */ae/*. $w[n]$ is a Hamming window with length 25 ms. In this case, the closely-spaced (in frequency) replicas of $W(\omega)$ provide a reasonable spectral sampling of the underlying formant envelope $H(\omega)$. In contrast, (bottom) shows the narrow-band spectrum of a synthesized vowel with the same $H(\omega)$ and but $f_0 = 260$ Hz. In this case, the periodic replicas of $W(\omega)$ provide

a notably poorer sampling of $H(\omega)$, oftentimes missing the locations of $H(\omega)$'s formant peaks altogether. We refer to this condition as *spectral undersampling*, which corresponds to the previously observed overlap of $h[n]$ with high f_0 in the time domain (Figure 1).

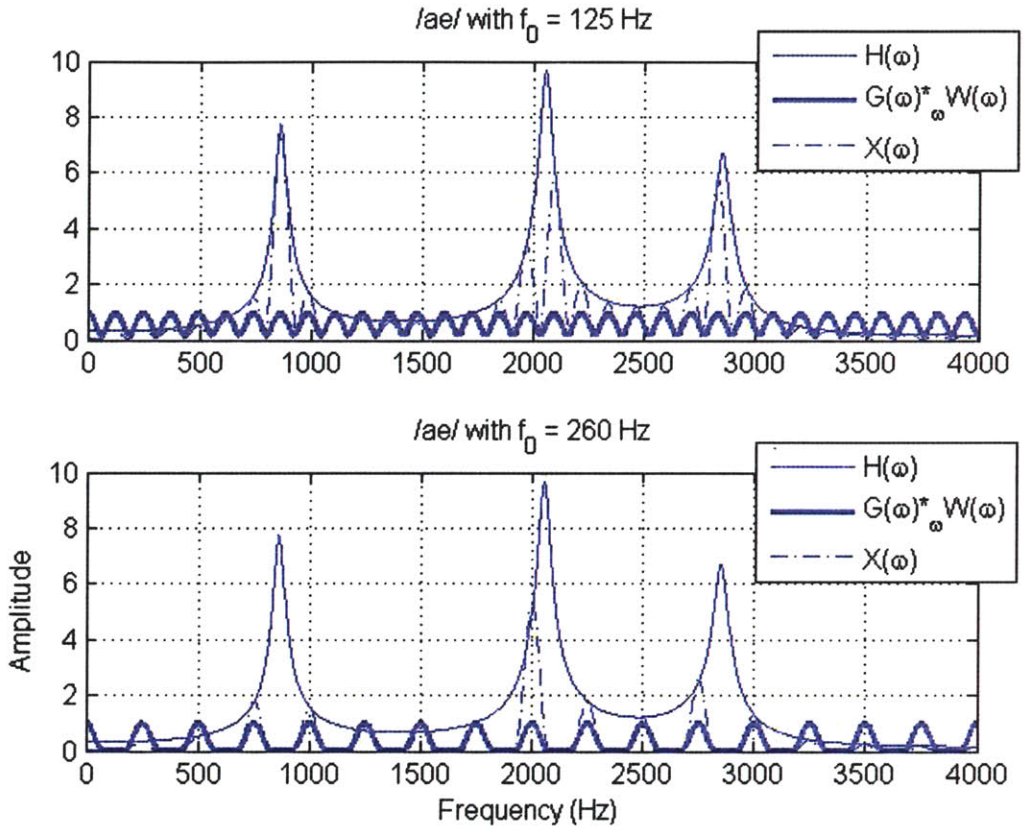

Figure 2 – Comparison of short-time spectra for low- (top) versus high-pitch (bottom). In the high-pitch spectra, the resulting speech spectrum exhibits harmonic peaks that do not align with the formant peaks as closely as that for low pitch. $X(\omega)$ denotes one short-time Fourier transform time slice.

2.3 Analysis

As previously noted, a major goal of speech analysis is to estimate characteristics of its production mechanisms from the waveform. Here, we discuss standard approaches for analysis in the context of the source-filter model. Specifically, Sections 2.3.1 and 2.3.2 discuss autocorrelation-based linear prediction and cepstral analysis, respectively, while Section 2.3.3 discusses their combined application for speech analysis.

2.3.1 Autocorrelation method of linear prediction

In the context of the source-filter model, accurate estimates of $h[n]$ and $g[n]$ are desired via analysis. In this thesis, we are concerned with estimates of $h[n]$, or equivalently, its Fourier transform $H(\omega)$. The vocal tract configuration for most vowel sounds can be reasonably modeled as the concatenation of uniform tubes that do not invoke significant coupling between each other or other parts of the vocal tract (e.g., nasal cavity). Consequently, $H(\omega)$ is an all-pole frequency response [3, 11]:

$$H(\omega) = \frac{A}{1 - \sum_{k=1}^{p} \alpha_k e^{-j\omega k}} \tag{2.7}$$

where A corresponds to the vocal tract gain, p denotes the order of the system, and α_k are the coefficients of the polynomial in $e^{-j\omega}$ that designate the pole locations. We interpret these pole locations as the distinct *formant frequencies* of a vowel. The time-domain representation of $s[n]$ when the input to $H(\omega)$ is $g[n]$ is

$$s[n] = \sum_{k=1}^{p} \alpha_k s[n-k] + Ag[n]. \tag{2.8}$$

For short-time intervals of $s[n]$, α_k can be estimated using the autocorrelation method of linear prediction from the set of normal equations [3] (Appendix A.1):

$$\sum_{k=1}^{p} \alpha_k r_n[i-k] = r_n[i] \qquad 1 \le i \le p. \tag{2.9}$$

Assuming that $g[n]$ is a pure impulse train with period P as in Section 2.2, it can be shown that $r_n[\tau]$ is an estimate of the autocorrelation function $r_h[\tau]$ corresponding to the all-pole system's impulse response ($h[n]$) (Appendix A.2):

$$r_n[\tau] = \sum_{k=-\infty}^{\infty} r_h[\tau - kP]. \tag{2.10}$$

Specifically, $r_n[\tau]$ is the summation of overlapping copies of $r_h[\tau]$ with period P. The discrepancy between $r_n[\tau]$ and $r_h[\tau]$ near the origin is directly related to the accuracy of the α_k estimates and is dependent on the period P of $g[n]$. In parallel with the observations of Section 2.1, smaller P values invoke more overlap of the copies of $r_h[\tau]$ such that worse estimates of α_k can be expected for higher-pitch speakers relative to lower-pitch speakers [3].

An equivalent interpretation of high-pitch formant estimation using the autocorrelation method of linear prediction can be made with regards to the spectral undersampling view discussed in Section 2.2. Define a spectral-domain function $Q(\omega)$ as

$$Q(\omega) = \log|X(\omega)|^2 - \log|H(\omega)|^2 \tag{2.11}$$

with $X(\omega)$ corresponding to the short-time speech spectra being analyzed and $H(\omega)$ as in Equation (2.7) [3]. As proposed by Itakura and Saito, an alternate error criterion I is defined as [12]

$$I = \int_{-\pi}^{\pi} [e^{Q(\omega)} - Q(\omega) - 1] \frac{d\omega}{2\pi} \tag{2.12}$$

Minimization of (2.12) with respect to the set of α_k can be shown to be equivalent to solving the normal equations of (2.9) (Appendix A.3). The autocorrelation method of linear prediction can therefore be viewed as finding a set of α_k (or equivalently, $H(\omega)$) that best satisfies the spectral matching condition I. As observed in Figure 2 (bottom), undersampled $X(\omega)$ are less likely to resemble the true formant envelope (e.g., when the pitch harmonics miss resonant peaks). We can therefore expect a poorer estimate of the true α_k's when minimizing I relative to a speech spectrum that better resembles the true formant envelope (e.g., Figure 2, top).

2.3.2 Cepstral analysis

An alternative method of analysis for speech is the cepstrum, which can provide separability of $h[n]$ from $g[n]$ in a transformed space under certain conditions. As in Section 2.2, short-time analysis of the source-filter model requires windowing of the speech waveform

$$x[n] = (h[n] * g[n])w[n]. \tag{2.13}$$

Denoting $h'[n]$ as a modified version of $h[n]$, $x[n]$ can be rewritten as $x[n] = h'[n] * g_w[n]$ where $g_w[n] = w[n]g[n]$ [13]. Denoting $X(\omega)$ as the short-time spectrum of the analyzed speech as in Section 2.2, the complex cepstrum $c[n]$ is defined as

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log[X(\omega)] e^{j\omega n} d\omega \tag{2.14}$$

For $X(\omega) = H'(\omega)G_w(\omega)$, Equation (2.14) becomes

$$c[n] = c_h[n] + c_g[n] \quad \text{where} \tag{2.15}$$

$$c_h[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log[H'(\omega)]e^{j\omega n}d\omega$$
$$c_g[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log[G_w(\omega)]e^{j\omega n}d\omega. \qquad (2.16)$$

In the cepstral domain, the source and vocal tract components are additive rather than multiplicative as in the spectral domain. For a pure impulse train source $g[n]$, $c_g[n]$ is [4]

$$\begin{aligned} c_g[n] &= c_w[n/P] \qquad && n = 0, \pm P, \pm 2P, \ldots \\ &= 0 && otherwise \end{aligned} \qquad (2.17)$$

where $c_w[n]$ is the complex cepstrum of $g_w[n]$ down-sampled by a factor of P (Appendix B). Similarly, $c_h[n]$ (i.e., the complex cepstrum of $h'[n]$) can be shown to be

$$c_h[n] = D[n]\sum_{k=-\infty}^{\infty} c_{hh}[n-kP] \qquad (2.18)$$

where $c_{hh}[n]$ corresponds to the complex cepstrum of $h[n]$ and $D[n]$ is a weighting function dependent on the choice of the window $w[n]$ [3, 13]. From (2.17) and (2.18), we observe that the periodicity of the source is manifested in the cepstrum for both the source and filter components. Nonetheless, for $0 < n < P$, $c[n]$ will be comprised primarily of a scaled (by $D[n]$) copy of $c_{hh}[n]$. Applying an appropriate lifter to the cepstrum followed by a Fourier transform presumably results in a spectral estimate with the harmonic structure of the source removed, thereby leaving the formant envelope. Rabiner and Schafer proposed a heuristic method of formant estimation using constraints on possible formant frequency values and peak-picking of this smoothed spectrum [2].

2.3.3 Homomorphic Linear Prediction

Kopec et, al. proposed speech analysis using linear prediction preceded by cepstral analysis [14], thereby combining the advantages of both techniques. In recent work, Rahman and Shimamura suggested that this framework can provide improvements over traditional linear prediction for the particular task of high-pitch formant estimation [15]. In their work, an ideal lifter whose cut-off was a function of an f_0 estimate was applied to the *real* cepstrum defined as:

$$c_{real}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(\omega)|e^{j\omega n}d\omega. \qquad (2.19)$$

The inverse transform was then applied to derive a magnitude spectrum $|X_{liftered}(\omega)|$ which was used for the autocorrelation method of linear prediction (Section 2.3.1). Though this method provides a more rigorous formant estimation framework than that of [2], the separability characteristics of the cepstrum are limited for the high-pitch case.

Specifically, recall that both methods employing the cepstrum assume that the region dominated by $c_{hh}[n]$ is in $0 < n < P$, where P denotes the pitch period as defined in Section 2.3.2. As P decreases (i.e., as f_0 increases), two effects can be observed analytically. First, more overlap will occur between the copy of $c_{hh}[n]$ centered at $n = 0$ and others located at $P, 2P...$, etc. Second, a lifter applied to remove the periodic components of $c_g[n]$ will necessarily have a smaller bandwidth in quefrency, thereby truncating more of the $c_{hh}[n]$ estimate and smoothing the resulting spectrum. Source-filter separability in the cepstrum therefore does not appear to be robust to high-pitch speech utterances in general.

2.4 Conclusions

In this chapter, we have observed the effects of high pitch on the source-filter model of speech production in time and for short-time spectral estimation. These effects were shown to correspond to the problems inherent in existing analysis methods for high-pitch formant estimation. Traditional linear prediction suffers from aliased autocorrelation coefficients in the time domain and spectral undersampling in the frequency domain. Though the cepstrum transforms the speech waveform into the quefrency domain and exhibits source-filter separability under low-pitch conditions, this separability does not generalize to high-pitch speech. In this thesis, we are motivated by observations from auditory psychophysics and physiological modeling work implicating the use of *temporal change* of pitch in humans as a basis for improving spectral estimation and formant estimation of high-pitch speech.

Chapter 3

Two-dimensional Processing Framework

In this chapter, we propose a two-dimensional (2-D) processing framework motivated from the work of Quatieri [9]. This framework has recently been shown to be consistent with the physiological modeling efforts of Chi, et al. in [8] of auditory signal processing. In their view, low-level peripheral mechanisms generate an auditory spectrogram from which spectrotemporally local regions are extracted and analyzed by a bank of 2-D filters located in the auditory cortex. The use of this framework is also consistent with psychophysical evidence implicating the use temporal changes in pitch for vowel perception [6]. Specifically, analysis of local spectrotemporal regions is done across longer durations in time than that of traditional short-time analysis, thereby exploiting temporal change of pitch. Figure 3 shows a block diagram of our framework which generalizes this view to performing *any* two-dimensional (2-D) transform on localized regions of *any* time-frequency distribution. Herein we discuss several realizations of this framework with the goal of improving spectral estimates of high-pitch speech to address the spectral undersampling problem.

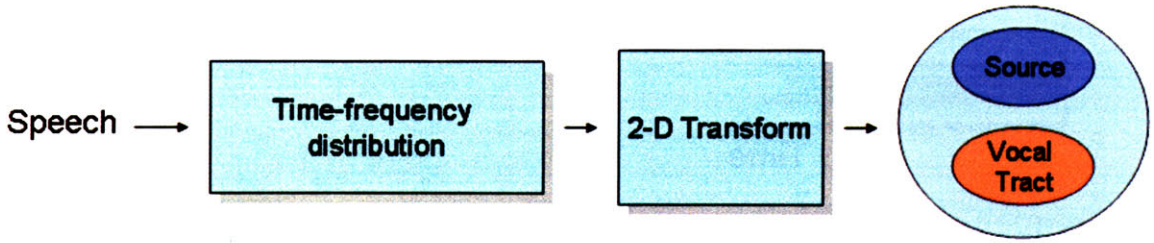

Figure 3 – 2-D processing framework.

This chapter is organized as follows. Section 3.1 presents a simple realization of our framework for exploiting temporal change of pitch; specifically, pitch harmonics across time are projected to a 1-D function to increase the spectral sampling of a stationary formant envelope. In Sections 3.2 and 3.3, we describe the analytical details of the Grating Compression Transform (GCT) [9] and the auditory cortex model of Chi, et al. [9], respectively; in both methods, a 2-D transform maps an input time-frequency distribution to a *rate-scale modulation* space. Section 3.4 presents simulation results between the GCT and the auditory model for synthesized vowels with stationary and changing pitch; the aim of these simulations is to assess the source-filter capabilities of each method under these conditions. We conclude in Section 3.5 by relating the proposed framework to improving high-pitch formant estimation.

3.1 Harmonic Projection

One simple interpretation of the proposed framework seeks to explicitly exploit temporal changes in pitch for high-pitch vowels. The schematic of a short-time Fourier transform (STFT) is shown in Figure 4a for a fixed vocal tract (shaded regions) with a fixed high pitch (solid horizontal

lines). As previously observed, a high f_0 in this case results in spectral undersampling (i.e., Figure 2, bottom); we highlight these spectral samples in Figure 4b. In contrast, consider now the STFT of the same vowel but with *changing* pitch from 235 Hz to 280 Hz, thereby including the 260-Hz pitch value of the fixed-pitch case (Figure 4c); under a multiplicative source-filter model, the pitch harmonics sweep through the spectral envelope over time in a fan-like structure. To show why this fan-like structure arises, consider a pitch f_0 with the n^{th} and $(n+1)^{th}$ harmonics as nf_0 and $(n+1)f_0$. Invoking a pitch change of Δf_0, the n^{th} and $(n+1)^{th}$ harmonics of the new pitch $f_0 + \Delta f_0$ become $n(f_0 + \Delta f_0)$ and $(n+1)(f_0 + \Delta f_0)$. Whereas the n^{th} harmonic is shifted by $n\Delta f_0$, the $(n+1)^{th}$ harmonic is shifted by $(n+1)\Delta f_0$. For a given pitch change, higher-order harmonics are therefore shifted more on an absolute frequency scale than lower-order harmonics. In the STFT, this invokes fanning of the harmonic line structure at higher frequency regions as illustrated in Figure 4c.

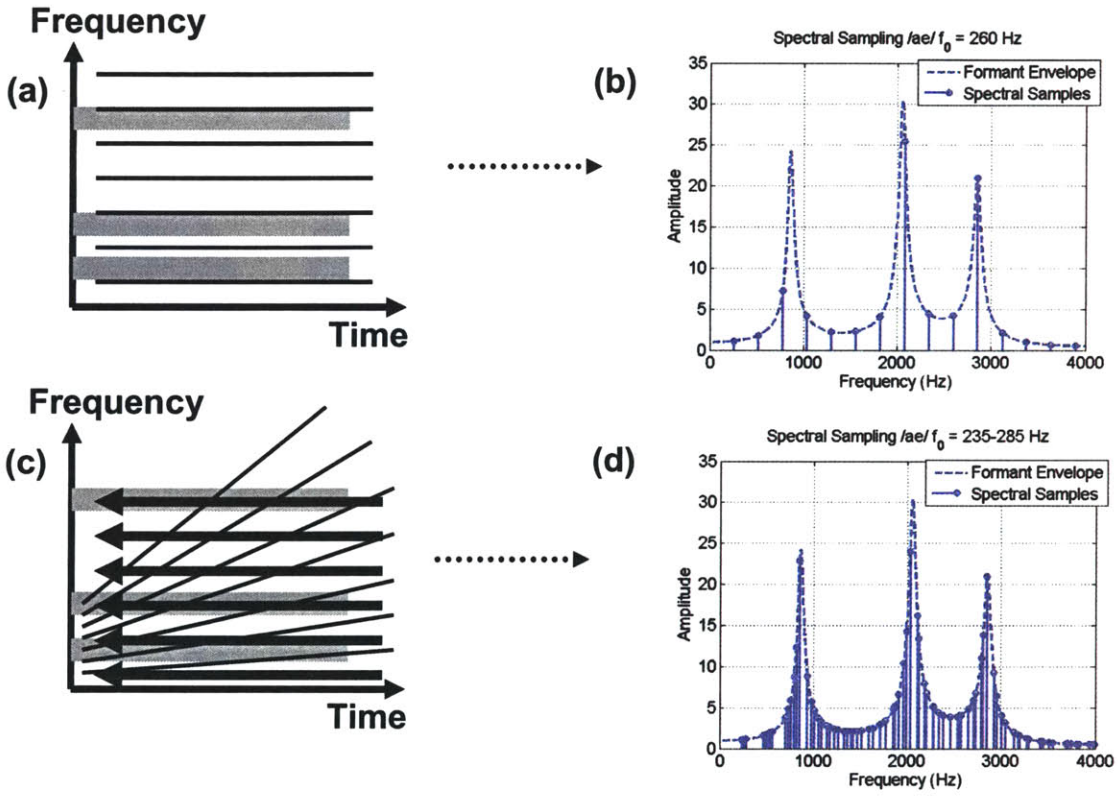

Figure 4 – Schematic illustrating projection of pitch harmonics; (a) stationary formant envelope (shaded regions) and fixed-pitch harmonic lines (horizontal lines); (b) spectral sampling of the stationary formant envelope invoked in (a); (b) stationary formant envelope (shaded); fanned harmonic line structure due to changing pitch (fanned lines); projection of harmonics across time (arrows); (d) spectral sampling of the stationary formant envelope invoked in (c).

Invoking again the spectral sampling of the harmonic lines, multiple short-time spectral slices computed across time can therefore be viewed as a collection of non-uniform samples of the spectral envelope for the condition of changing pitch. These samples can be projected to the vertical frequency axis, thereby providing improved sampling of the underlying formant

envelope. An example of this increased sampling is shown in Figure 4d, contrasting the uniform sampling in Figure 4b. A spectral estimate derived from this increased sampling could provide an improved representation of the underlying formant envelope even under conditions of high pitch relative to that of a single spectral slice. Observe also that low-frequency regions (e.g., near 500 Hz) exhibit narrower sampling than the broader sampling in high-frequency regions (e.g., near 1200 Hz) due to the fanning of harmonic lines in higher frequency regions (i.e., higher-order harmonics).

3.2 Grating Compression Transform

Another realization of our 2-D processing framework is based on the view that localized regions of time-frequency distributions are composed of spectrotemporal modulations. Figure 5 illustrates such a decomposition for a localized region of the short-time Fourier transform magnitude. Observe that purely temporal or purely spectral modulation components are special cases of the more general spectrotemporal case.

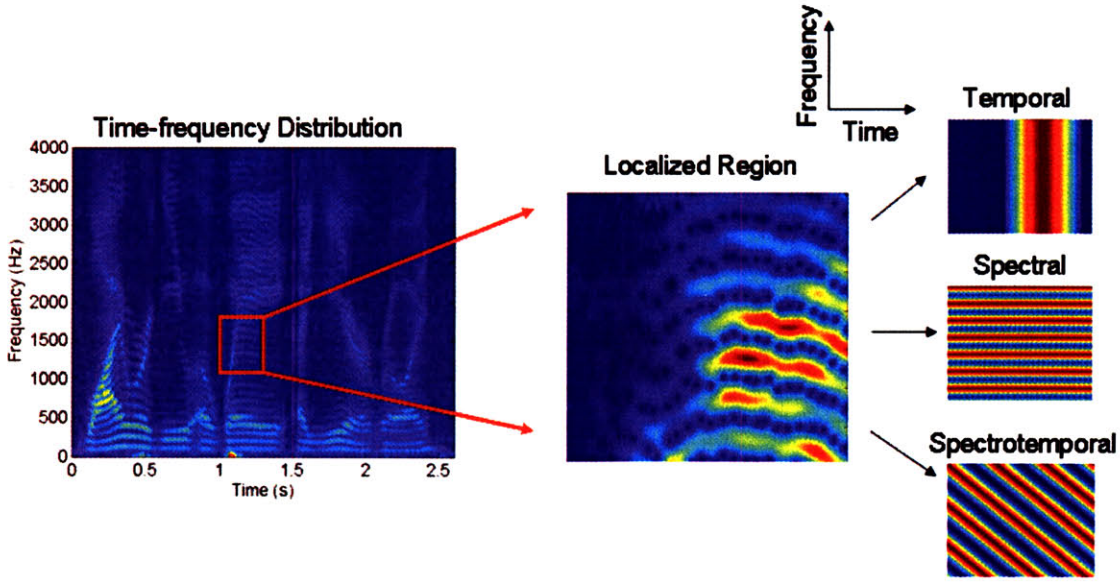

Figure 5[1] – Spectrotemporal modulations of a localized region of the short-time Fourier transform. Full STFT (left), localized region (center), spectrotemporal components (right). Vertical and horizontal axes are frequency and time, respectively.

One method for analysis of these spectrotemporal modulations in the magnitude STFT is to compute the short-space Fourier transform of the localized region, referred to as the *Grating Compression Transform* (GCT) as originally proposed by Quatieri [9]. Denoting n and m as the time and frequency indices of the STFT, respectively, each modulation component $c[n,m]$ can be modeled as a two-dimensional sinusoid resting on a DC pedestal, $c[n,m] = K + \cos(\hat{\omega}_0 \Phi[n,m])$. $\hat{\omega}_0$ corresponds to the spatial frequency of the modulation and K corresponds to the DC component. $\Phi[n,m]$ is defined as $\Phi[n,m] = m\cos(\hat{\theta}) + n\sin(\hat{\theta})$, where $\hat{\theta}$ is an angle describing the component orientation (to be subsequently discussed). In

[1] In this thesis, spectrogram figures are plotted on a linear scale.

practice, these components are analyzed over a localized region such that they are multiplied by a window which we denote as $w[n,m]$:

$$w[n,m]c[n,m] = w[n,m]\left(K + \cos(\hat{\omega}_0 \Phi[n,m])\right) \tag{3.1}$$

To illustrate the motivation for this model, we show in Figure 6 a spectrogram computed with a 20-ms Hamming window and 1-ms frame interval for the female vowel */ae/* with fixed pitch of 150 Hz[2]. Shown in Figure 6b and c is a full spectral slice of the spectrogram and a local portion of this slice, respectively. Observe that the local portion resembles a sinusoid resting on a DC pedestal along the frequency axis[3].

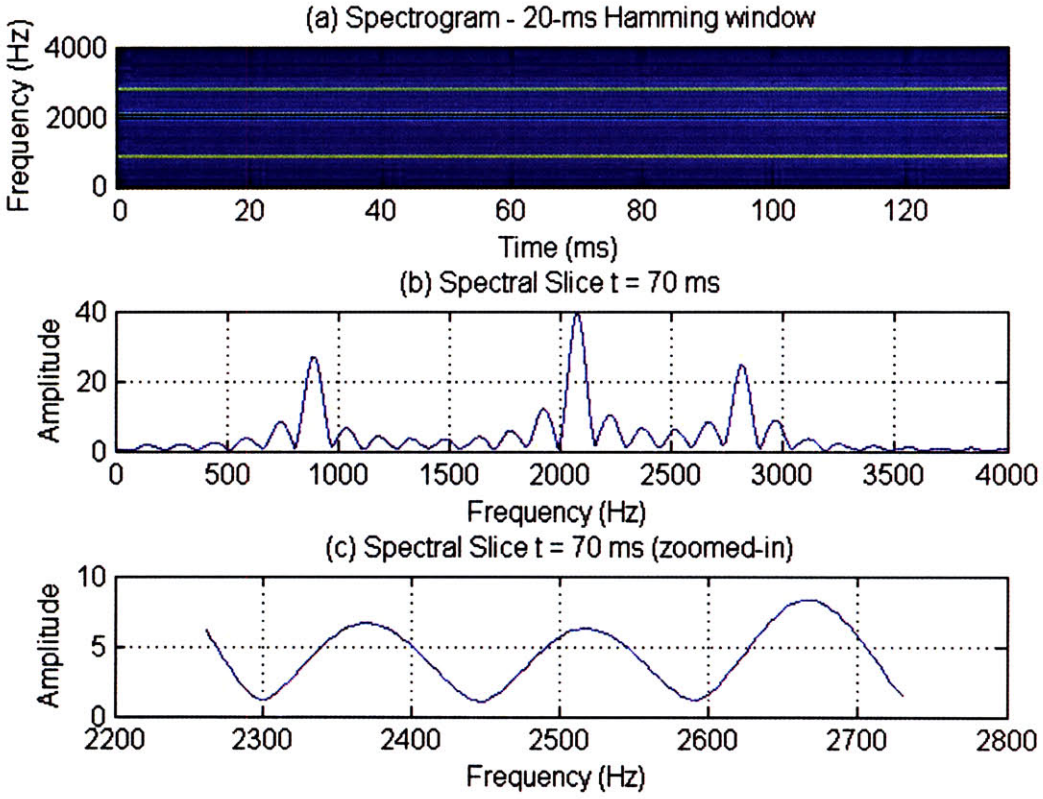

Figure 6 – (a) Spectrogram of the vowel */ae/* with fixed 150-Hz pitch; (b) Full spectral slice of (a); (c) Localized portion of the spectral slice.

Extending this local sinusoidal component to the 2-D case, the expression for $w[n,m]c[n,m]$ can be invoked for a local region of the STFT. This is illustrated in Figure 7 (left) with the harmonic lines representing the peaks of the sinusoid while n_W and m_W denote the length of the window in time and frequency, respectively. In the 2-D Fourier-transform domain, we then have:

---

[2] Refer to Section 3.4.1 for synthesis details.

[3] For now, we defer discussion of the multiplicative effect from the formant envelope on this sinusoidal component to Section 3.5, where we argue for improved source-filter separability in the GCT.

$$
\begin{aligned}
C(\hat{\omega},\hat{\Omega}) &= 2K\delta(\hat{\omega},\hat{\Omega}) + \delta(\hat{\omega}-\hat{\omega}_0\cos\hat{\theta},\hat{\Omega}+\hat{\omega}_0\sin\hat{\theta}) \\
&\quad +\delta(\hat{\omega}+\hat{\omega}_0\cos\hat{\theta},\hat{\Omega}-\hat{\omega}_0\sin\hat{\theta}) \\
W(\hat{\omega},\hat{\Omega}) *_{\hat{\omega},\hat{\Omega}} C(\hat{\omega},\hat{\Omega}) &= 2KW(\hat{\omega},\hat{\Omega}) + W(\hat{\omega}-\hat{\omega}_0\cos\hat{\theta},\hat{\Omega}+\hat{\omega}_0\sin\hat{\theta}) \\
&\quad +W(\hat{\omega}+\hat{\omega}_0\cos\hat{\theta},\hat{\Omega}-\hat{\omega}_0\sin\hat{\theta})
\end{aligned}
\tag{3.2}
$$

where $-\pi < \hat{\omega} \le \pi$ and $-\pi < \hat{\Omega} \le \pi$ and $*_{\hat{\omega},\hat{\Omega}}$ denotes convolution across the continuous-frequency variables $\hat{\omega}$ and $\hat{\Omega}$ of the 2-D continuous Fourier transform of the discrete STFT representation. We adopt the terminology proposed by Chi, et al. in referring to $\hat{\omega}$ as "rate" and $\hat{\Omega}$ as "scale" for modulation content in time and frequency, respectively. Figure 7 illustrates this mapping between $w[n,m]c[n,m]$ to the magnitude of $W(\hat{\omega},\hat{\Omega}) *_{\hat{\omega},\hat{\Omega}} C(\hat{\omega},\hat{\Omega})$. In implementation, this continuous rate-scale domain is sampled via the discrete-Fourier transform (DFT) (i.e., the GCT is a short-space DFT).

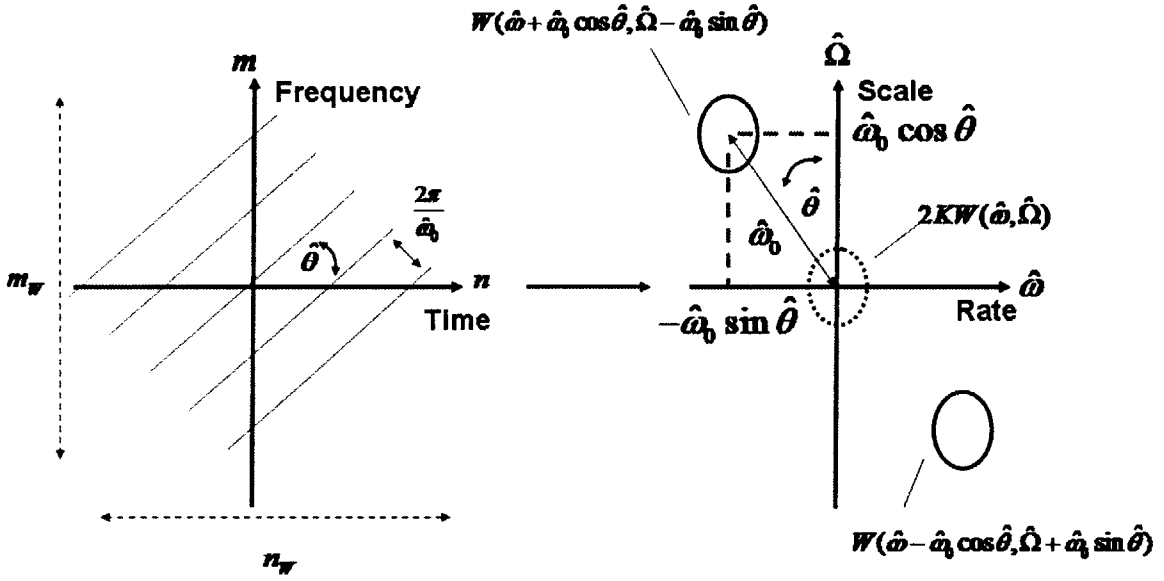

Figure 7 – GCT of a single spectrotemporal modulation component.

Recall that the STFT view of changing pitch invokes a fanned harmonic line structure (Figure 4). Nonetheless, within a *localized* region of the STFT, harmonic lines are approximately parallel (Figure 8) such that our model of modulation components can be invoked. Harmonic line structure can be expected to map to a pair coherent components *off* the scale axis with changing pitch. This is due to the rotational nature of transforming a *skewed* 2-D sinusoid (3.2). Across a localized region of interest with fixed duration in time, the amount of rotation off the scale axis (i.e., $\hat{\theta}$) will be related to the amount of pitch change invoked. Larger pitch changes will invoke larger $\hat{\theta}$ since the harmonic lines will have steeper slopes. Finally, at a fixed point in time for changing pitch, high-frequency regions of the STFT will exhibit larger $\hat{\theta}$ than low-frequency regions. This is due to the fanning of the harmonic line structure described in Section 3.1.

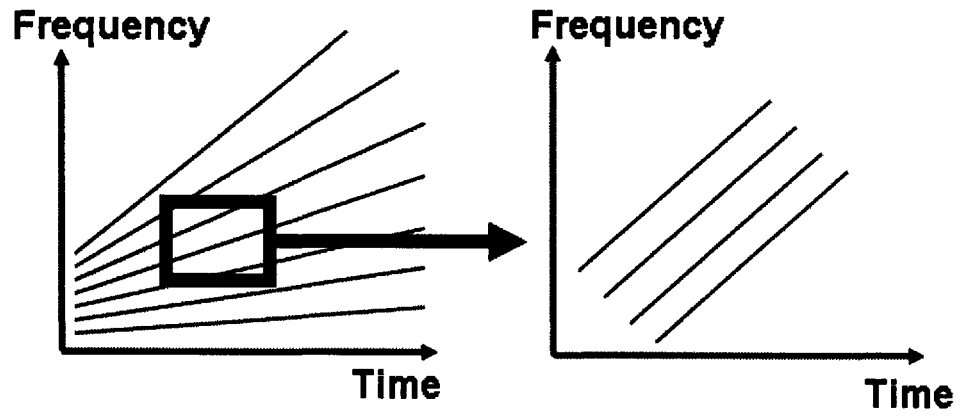

Figure 8 – Fanned line structure invoked by changing pitch (left) with localized region (left, rectangle) in which harmonic structure is approximately parallel (right).

Analogous to the uniform filterbank view of the STFT [4], the short-space DFT representation may alternatively be viewed as the output of a bank of uniformly spaced 2-D filters [10]. For reasons that will become clear in the subsequent discussion of the auditory cortex model, we derive this 2-D case here. Specifically, let $s[n,m]$ correspond to the entire magnitude STFT and $w[T,F]$ correspond to a two-dimensional window that is non-zero for $0 \le T \le n_W - 1, 0 \le F \le m_W - 1$ and zero elsewhere. Extracting a region of $s[n,m]$ centered at $n = T, m = F$, the GCT is:

$$S[k,l;n,m] = \frac{1}{n_W m_W} \sum_{T=0}^{n_W-1} \sum_{F=0}^{m_W-1} w[T,F]s[n+T,m+F] e^{\frac{-j2\pi kT}{N_{\hat{\omega}}}} e^{\frac{-j2\pi lF}{M_{\hat{\Omega}}}} \tag{3.3}$$

where $0 \le k < N_{\hat{\omega}}$ and $0 \le l < M_{\hat{\Omega}}$, and where $N_{\hat{\omega}}$ and $M_{\hat{\Omega}}$ correspond to the lengths of the DFTs in time and frequency, respectively. Letting $\tau = n + T$ and $\phi = m + F$ and fixing $k = k_{\hat{\omega}}$ and $l = l_{\hat{\Omega}}$, (3.3) can be rewritten as

$$\begin{aligned} S[n,m;k_{\hat{\omega}},l_{\hat{\Omega}}] &= \frac{1}{n_W m_W} \sum_{\tau=n+T}^{n+T+n_W-1} \sum_{\phi=m+F}^{n+F+m_W-1} w[-(n-\tau),-(m-\phi)]s[\tau,\phi] e^{\frac{-j2\pi k_{\hat{\omega}}(n-\tau)}{N_{\hat{\omega}}}} e^{\frac{-j2\pi l_{\hat{\Omega}}(m-\phi)}{M_{\hat{\Omega}}}} \\ &= \sum_{\tau=-\infty}^{\infty} \sum_{\phi=-\infty}^{\infty} s[\tau,\phi]h[n-\tau,\phi-m] \\ &= s[n,m] *_{n,m} h[n,m;k_{\hat{\omega}},l_{\hat{\Omega}}] \end{aligned} \tag{3.4}$$

where $h[n,m;k_{\hat{\omega}},l_{\hat{\Omega}}]$ *is defined as*

$$h[n,m;k_{\hat{\omega}},l_{\hat{\Omega}}] = \frac{1}{n_W m_W} w[-n,-m] e^{\frac{j2\pi k_{\hat{\omega}} n}{N_{\hat{\omega}}}} e^{\frac{j2\pi l_{\hat{\Omega}} m}{M_{\hat{\Omega}}}}$$

where $*_{n,m}$ denotes convolution in time and frequency. Equation (3.4) shows that for a single point in the time-frequency plane ($n = T, m = F$), the short-space DFT of $s[n,m]$ can be viewed as the output of filtering $s[n,m]$ with a set of $h[n,m;k_{\hat{\omega}},l_{\hat{\Omega}}]$ where $-\frac{N_{\hat{\omega}}}{2} < k_{\hat{\omega}} \le \frac{N_{\hat{\omega}}}{2}+1$ and $-\frac{M_{\hat{\Omega}}}{2} < l_{\hat{\Omega}} \le \frac{M_{\hat{\Omega}}}{2}+1$.

In the rate-scale $(\hat{\omega},\hat{\Omega})$ domain, the frequency response of $h[n,m;k_{\hat{\omega}},l_{\hat{\Omega}}]$ is $H(\hat{\omega},\hat{\Omega};k_{\hat{\omega}},l_{\hat{\Omega}}) = W(\hat{\omega}-\frac{2\pi k_{\hat{\omega}}}{N_{\hat{\omega}}},\hat{\Omega}-\frac{2\pi l_{\hat{\Omega}}}{M_{\hat{\Omega}}})$ such that the filter is centered at $\hat{\omega}=\frac{2\pi k_{\hat{\omega}}}{N_{\hat{\omega}}}$ and $\hat{\Omega}=\frac{2\pi l_{\hat{\Omega}}}{M_{\hat{\Omega}}}$. Observe that for the set of all $H(\hat{\omega},\hat{\Omega};k_{\hat{\omega}},l_{\hat{\Omega}})$, the filter bandwidths are the same (i.e., a uniform 2-D filterbank) and determined by the choice of the 2-D window. For reasons that will soon become clear, let us choose $k_{\hat{\omega}} < 0$ and $l_{\hat{\Omega}} > 0$ and denote $k_0 = |k_{\hat{\omega}}|$ and $l_0 = |l_{\hat{\Omega}}|$ such that $H(\hat{\omega},\hat{\Omega};k_{\hat{\omega}},l_{\hat{\Omega}}) = W(\hat{\omega}+\frac{2\pi k_0}{N_{\hat{\omega}}},\hat{\Omega}-\frac{2\pi l_0}{M_{\hat{\Omega}}})$ as shown in Figure 9 (left). Observe that the corresponding $h[n,m;k_{\hat{\omega}},l_{\hat{\Omega}}]$ is the analytic representation of a real 2-D impulse response $h_{real}[n,m;k_{\hat{\omega}},l_{\hat{\Omega}}]$:

$$\begin{aligned} h_{real}[n,m;k_{\hat{\omega}},l_{\hat{\Omega}}] &= \mathrm{Re}\left\{\frac{1}{n_W m_W} w[-n,-m] e^{\frac{j2\pi k_0 n}{N_{\hat{\omega}}}} e^{\frac{j2\pi l_0 m}{M_{\hat{\Omega}}}}\right\} \\ &= \frac{1}{n_W m_W} w[-n,-m]\cos\left(-\frac{2\pi k_0 n}{N_{\hat{\omega}}}+\frac{2\pi l_0 m}{M_{\hat{\Omega}}}\right) \end{aligned} \quad (3.5)$$

From (3.5), we see that $h_{real}[n,m;k_{\hat{\omega}},l_{\hat{\Omega}}]$ is a windowed 2-D sinusoid. By noting the similarity to (3.1), we can set

$$\hat{\omega}_0 = 2\pi\sqrt{\left(-\frac{k_0}{N_{\hat{\omega}}}\right)^2+\left(\frac{l_0}{M_{\hat{\Omega}}}\right)^2} \quad (3.6)$$

$$\cos\hat{\theta}_0 = \frac{\frac{l_0}{M_{\hat{\Omega}}}}{\sqrt{\left(-\frac{k_0}{N_{\hat{\omega}}}\right)^2+\left(\frac{l_0}{M_{\hat{\Omega}}}\right)^2}} \quad (3.7)$$

$$\sin\hat{\theta}_0 = \frac{-\frac{k_0}{N_{\hat{\omega}}}}{\sqrt{\left(-\frac{k_0}{N_{\hat{\omega}}}\right)^2 + \left(\frac{l_0}{M_{\hat{\Omega}}}\right)^2}} \tag{3.8}$$

such that (3.5) can be rewritten as

$$\begin{aligned} h_{real}[n,m;k_{\hat{\omega}},l_{\hat{\Omega}}] &= \frac{1}{n_W m_W} w[-n,-m]\cos\left(\hat{\omega}_0\Phi(n,m)\right) \\ \Phi(n,m) &= n\cos\hat{\theta}_0 + m\sin\hat{\theta}_0 \end{aligned} \tag{3.9}$$

Assuming that the 2-D window is symmetric across the origin (i.e., $w[-n,-m] = w[n,m]$), the relationship between (3.9) and (3.1) is now clear. Specifically, the uniform filterbank view of the GCT invokes filtering by analytic filters with corresponding real impulse responses $h_{real}[n,m;k_{\hat{\omega}},l_{\hat{\Omega}}]$ that *match* the windowed modulation components observed in the STFT (3.1). One way of interpreting the filtered outputs $S[n,m;k_{\hat{\omega}},l_{\hat{\Omega}}]$, which are inherently complex, is to take their magnitudes. This results in a set of 2-D Hilbert envelopes, which we define as the "strength" of this matching between a local region of the STFT surrounding $[n,k]$ and the set of $h_{real}[n,m;k_{\hat{\omega}},l_{\hat{\Omega}}]$.

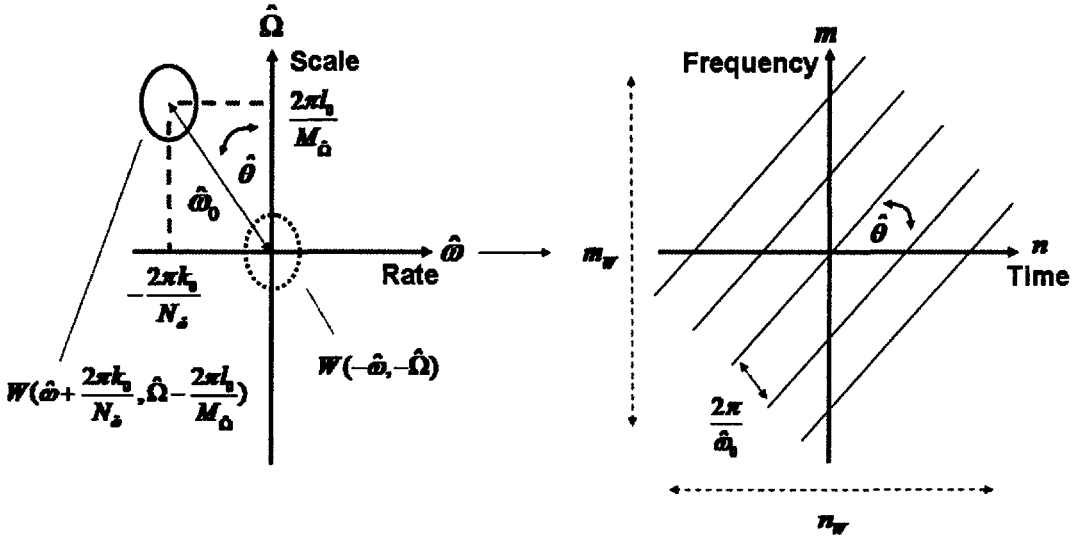

Figure 9 – Schematic showing the analytic-window filter of the GCT in the rate-scale domain (left) with mapping to its corresponding real impulse response in the time-frequency space (right).

To interpret the rate and scale axes of the GCT, $\hat{\omega}$ and $\hat{\Omega}$ may be scaled to represent the corresponding rate and scale axes (ω and Ω) of a *continuous-* time and frequency time-frequency distribution $s(t,f)$ rather than the discretized STFT. We interpret the frame interval

(Δ_t) used to compute $s[n,m]$ as the *sampling period* for temporal modulations (i.e., the rate axis). $\hat{\omega}$ is mapped to a temporal modulation frequency ω measured in Hz by

$$\omega = \frac{\hat{\omega}}{2\pi\Delta_t}. \tag{3.10}$$

Similarly, denoting N_{STFT} as the DFT length used to compute $s[n,m]$ and f_s as the sampling rate of the processed waveform, the sampling "period" along the frequency axis is $\frac{f_s}{N_{STFT}}$ Hz per DFT-sample. We define a unit of spectral modulation frequency (Ω) to be in *cycles per Hz* (cyc/Hz) derived from $\hat{\Omega}$ as

$$\Omega = \frac{\hat{\Omega} N_{STFT}}{2\pi f_s}. \tag{3.11}$$

The signs of the ω and Ω (or equivalently, $\hat{\omega}$ and $\hat{\Omega}$) values determine the orientation of the corresponding (real) impulse response in the time-frequency space. Observe that the component in Figure 9 has positive $\hat{\Omega}$ and negative $\hat{\omega}$ to invoke a positive value of $\hat{\theta}$ as measured counter-clockwise from the scale axis in the GCT and counter-clockwise in the STFT from the time axis. Conversely, a positive $\hat{\Omega}$ and positive $\hat{\omega}$ (with conjugate symmetric counterpart having negative $\hat{\Omega}$ and negative $\hat{\omega}$) invokes a negative value of $\hat{\theta}$. In Table 1, we summarize the variables used in this section.

Table 1 – Summary of variables used in GCT.

n and m	discrete- time and frequency indices of the STFT
$\hat{\omega},\hat{\Omega}$	rate and scale axes ranging from $-\pi$ to π
ω,Ω	continuous domain rate and scale (units of Hz and cyc/Hz)
$\hat{\theta}$	angle of filter orientation in rate-scale domain

3.3 Auditory Cortex Model

Similar to the GCT, another realization of the proposed framework is a comprehensive model of the auditory pathway proposed by Chi, et al [8]. The time-frequency distribution used is an *auditory spectrogram* in contrast to the STFT in the GCT. Localized regions of the auditory spectrogram are analyzed with a bank of non-uniform two-dimensional filters instead of the uniform filterbank of the GCT. Herein we discuss salient components of this model in mimicking auditory signal processing steps starting from the cochlea and ending at neurons of the auditory cortex and refer the reader to [8] for further details.

A discrete-time waveform $x[n]$ sampled at $f_s = 8000$ Hz is first filtered with cochlear-like filters $h[n;k]$ with the center frequencies f_k spaced to invoke 24 channels per octave:

$$f_k = 220 \cdot 2^{\frac{(k-31)}{24}}, \qquad k = 0,1...128$$
$$y[n;k] = x[n] *_n h[n;k] \tag{3.12}$$

The filters are designed to have increasing bandwidth with frequency such that high-frequency regions of $y[n;k]$ have relatively better (worse) temporal (frequency) resolution than lower frequency regions. Filtered outputs are next differentiated (velocity of stereocilia motion), passed through a saturating nonlinearity $g(\cdot)$ (operating characteristic of inner hair cell stereocilia), and low-pass filtered by $h_{RC}[n]$ (inner hair cell membrane time constant) to simulate the mechanoelectric transduction process generating the auditory nerve (AN) response $y_{AN}[n,k]$:

$$y_{AN}[n,k] = g(\frac{\partial y[n,k]}{dn}) *_n h_{RC}[n]. \tag{3.13}$$

$\frac{\partial}{dn}$ denotes a first-difference approximation to the differentiation operator with respect to time (n). Phase-locking of the auditory nerve is limited by the stop-band frequency of $h_{RC}[n]$ (in this case, 2 kHz). The auditory nerve synapses onto the cochlear nucleus (CN), which is known to contain a variety of neuron cell types capable of forming lateral inhibitory networks (LIN). This step is modeled as differentiation along the tonotopic (frequency) axis followed by half-wave rectification. Further reduction of phase-locking occurs at the CN via additional low-pass filtering by $h_{CN}[n]$ (in this case, 125 Hz) resulting in $y_{CN}[n,k]$. Finally, $y_{CN}[n,k]$ is sampled at a fixed frame interval $D = f_s \Delta t$ (here, $\Delta t = 1$ ms) to generate the auditory spectrogram $y_{AS}[n,k]$, thereby completing the low-level peripheral stage.

$$y_{LIN}[n,k] = \max(\frac{\partial y_{AN}[n,k]}{dk}, 0)$$
$$y_{CN}[n,k] = y_{LIN}[n,k] *_n h_{CN}[n] \tag{3.14}$$
$$y_{AS}[n,k] = y_{CN}[Dn,k]$$

To complete the auditory model, the auditory spectrogram is further processed by a high-level cortical stage. Specifically, $y_{AS}[n,k]$ is filtered with a bank of 2-D filters each tuned to a specific rate-scale modulation frequency pair and bandwidth. For any $\hat{\omega} = \hat{\omega}_1$ and $\hat{\Omega} = \hat{\Omega}_1$, the impulse response of the cortical filter $h_c[n,k;\hat{\omega},\hat{\Omega}]$ is derived from two continuous seed functions $h_r(t)$ and $h_s(f)$ corresponding to the real impulse responses of its rate- and scale-filter components:

$$h_r(t) = t^2 e^{-3.5t}\sin(2\pi t)$$
$$h_s(f) = (1-f^2)e^{\frac{-f^2}{2}} \quad . \tag{3.15}$$

Specifically, define $h_r[n;\hat{\omega}_1]$ and $h_s[k;\hat{\Omega}_1]$ as scaled and sampled versions of $h_r(t)$ and $h_s(f)$:

$$h_r[n;\hat{\omega}_1] = \hat{\omega}_1 h_r(\hat{\omega}_1 t)\big|_{t=nT}$$
$$h_s[k;\hat{\Omega}_1] = \hat{\Omega}_1 h_s(\hat{\Omega}_1 f)\Big|_{f=f_k} \tag{3.16}$$

with f_k defined as in (3.12). The analytic representations of $h_r[n;\hat{\omega}_1]$ and $h_s[k;\hat{\Omega}_1]$ are then:

$$h_R[n;\hat{\omega}_1] = h_r[n;\hat{\omega}_1] + j\hat{h}_r[n;\hat{\omega}_1]$$
$$h_S[k;\hat{\Omega}_1] = h_s[k;\hat{\Omega}_1] + j\hat{h}_s[k;\hat{\Omega}_1] \tag{3.17}$$

where $\hat{h}_r[n;\hat{\omega}_1]$ and $\hat{h}_s[k;\hat{\Omega}_1]$ denote the Hilbert transforms of $h_r[n;\hat{\omega}_1]$ and $h_s[k;\hat{\Omega}_1]$, respectively. Finally, the complex impulse response of the cortical filter $h_c[n,k;\hat{\omega}=\hat{\omega}_1,\hat{\Omega}=\hat{\Omega}_1]$ is defined as

$$h_c[n,k;\hat{\omega}=\hat{\omega}_1,\hat{\Omega}=\hat{\Omega}_1] = h_R[n;\hat{\omega}_1]h_S[k;\hat{\Omega}_1]. \tag{3.18}$$

2-D filtering of $y_{CN}[n,k]$ is done with a set of $h_c[n,k;\hat{\omega},\hat{\Omega}]$ and results in a four-dimensional (4-D) complex cortical response $y_{CORT}[n,k,\hat{\omega},\hat{\Omega}]$:

$$y_{CORT}[n,k,\hat{\omega},\hat{\Omega}] = y_{AS}[n,k] *_{n,k} h_c[n,k;\hat{\omega},\hat{\Omega}]. \tag{3.19}$$

Before discussing the interpretation of $y_{CORT}[n,k,\hat{\omega},\hat{\Omega}]$, we observe that since $h_R[n;\hat{\omega}_1]$ is independent of k and $h_S[k;\hat{\Omega}_1]$ of n, $h_c[n,k;\hat{\omega}=\hat{\omega}_1,\hat{\Omega}=\hat{\Omega}_1]$ in the rate-scale domain is the product of two separable filters. Therefore the frequency response of the cortical filter is given by

$$\begin{aligned} H_c(\hat{\omega},\hat{\Omega};\hat{\omega}_1,\hat{\Omega}_1) &= \sum_{n=0}^{N-1}\sum_{k=0}^{K-1} h_R[n;\hat{\omega}_1]h_S[k;\hat{\Omega}_1]e^{-(j\hat{\omega}n + j\hat{\Omega}k)} \\ &= \sum_{n=0}^{N-1} h_R[n;\hat{\omega}_1]e^{-j\hat{\omega}n}\sum_{k=0}^{K-1} h_S[k;\hat{\Omega}_1]e^{-j\hat{\Omega}k} \quad . \\ &= H_r(\hat{\omega};\hat{\omega}_1)H_s(\hat{\Omega};\hat{\Omega}_1) \end{aligned} \tag{3.20}$$

Since $H_r(\hat{\omega};\hat{\omega}_1)$ and $H_s(\hat{\Omega};\hat{\Omega}_1)$ correspond to analytic signals (3.17), the cortical filters are said to be "quadrant-separable"; specifically, for a single quadrant of the rate-scale domain, the filter is the product of two independent functions of $\hat{\omega}$ and $\hat{\Omega}$. This separability is consistent

with physiological response characteristics of neurons in the ferret auditory cortex [8]. Choosing $\hat{\omega}_1 < 0$ and $\hat{\Omega}_1 > 0$ and denoting $\hat{\omega}_m = |\hat{\omega}_1|$, $\hat{\Omega}_m = |\hat{\Omega}_1|$, we show in Figure 10 (left) the construction of this separable filter; we define $\hat{\omega}'$ as the radial distance of the filter's rate-scale frequency pair from the origin computed as $\hat{\omega}' = \sqrt{\hat{\omega}_1^2 + \hat{\Omega}_1^2}$. In addition, we show schematically in Figure 10 (right) this filter's corresponding *real* impulse response derived from taking the real part of $h_c[n,k;\hat{\omega} = \hat{\omega}_1, \hat{\Omega} = \hat{\Omega}_1]$:

$$\begin{aligned} \mathrm{Re}\left\{h_c[n,k;\hat{\omega},\hat{\Omega}]\right\} &= \mathrm{Re}\left\{h_c[n,k;\hat{\omega} = \hat{\omega}_1, \hat{\Omega} = \hat{\Omega}_1]\right\} \\ &= \mathrm{Re}\left\{h_R[n;\hat{\omega}_1]h_S[k;\hat{\Omega}_1]\right\} \\ &= \mathrm{Re}\left\{\left(h_r[n;\hat{\omega}_1] + j\hat{h}_r[n;\hat{\omega}_1]\right)\left(h_s[k;\hat{\Omega}_1] + j\hat{h}_s[k;\hat{\Omega}_1]\right)\right\} \\ &= h_r[n;\hat{\omega}_1]h_s[k;\hat{\Omega}_1] - \hat{h}_r[n;\hat{\omega}_1]\hat{h}_s[k;\hat{\Omega}_1] \end{aligned} \quad (3.21)$$

From Equations (3.15) and (3.16), observe that $\mathrm{Re}\left\{h_c[n,k;\hat{\omega},\hat{\Omega}]\right\}$ has (in general) infinite length in both the time and frequency directions. In practice, it is truncated based on the DFT lengths used in generating $H_r(\hat{\omega};\hat{\omega}_1)$ and $H_s(\hat{\Omega};\hat{\Omega}_1)$. We denote these values as $N_{\hat{\omega}}$ and $N_{\hat{\Omega}}$ for rate and scale, respectively, as in the GCT.

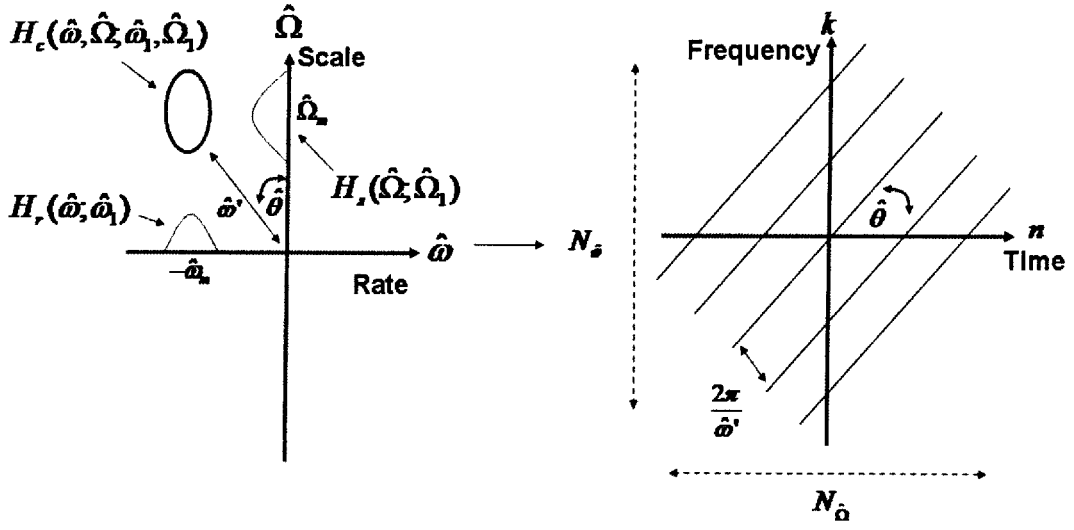

Figure 10 – Schematic of the design of an analytic cortical filter (left) and its corresponding real impulse response (right).

By comparing Figure 9 and Figure 10, observe that the GCT and the auditory model are similar. In both methods, analytic filters centered at spectrotemporal modulation frequencies are used to filter a time-frequency distribution. As in the GCT, $\mathrm{Re}\left\{h_c[n,k;\hat{\omega},\hat{\Omega}]\right\}$ are presumably *matched* to distinct components of the auditory spectrogram. We can therefore interpret the cortical response magnitude at a fixed point in time and frequency $\left|y_{CORT}[n,k,\hat{\omega},\hat{\Omega}]\right|$ (i.e., the 2-D

Hilbert envelope) as the strength of the match between $\mathrm{Re}\{h_c[n,k;\hat{\omega},\hat{\Omega}]\}$ and a localized region of the auditory spectrogram near $[n,k]$.

As in the GCT, temporal modulation frequencies (rate) in the continuous domain have units of Hz that depend on the frame interval used in computing the auditory spectrogram. This invokes a mapping from $\hat{\omega}$ to ω as in (3.10). Due to the logarithmic spacing of the cochlear filterbank, the continuous spectral modulation axis (scale) has units of cycles per *octave* instead of cycles per Hz. This mapping is defined as:

$$\Omega' = \frac{\hat{\Omega}}{2\pi\left(\frac{1}{24}\right)} \tag{3.22}$$

where the $\left(\frac{1}{24}\right)$ factor arises from sampling 24 channels per octave (3.12). We distinguish the continuous scale axes between the GCT and the cortical model with distinctive variables Ω (GCT) vs. Ω' (cortical model). In Table 2, we summarize the variables used in the auditory cortex model.

Table 2 – Summary of variables used in auditory cortex model.

n and k	discrete- time and frequency indices of the auditory spectrogram
$\hat{\omega},\hat{\Omega}$	rate and scale axes ranging from $-\pi$ to π
ω,Ω'	continuous domain rate and scale (units of Hz and cyc/oct)
$\hat{\theta}$	angle of filter orientation in rate-scale domain

To interpret the signs of the modulation frequency values, Chi, et al. denote an "upward" impulse response as one that has negative ω such that $\hat{\theta}$ is positive as measured counter-clockwise from the scale axis (e.g., Figure 10). Conversely, a "downward" impulse response has positive ω and negative $\hat{\theta}$. Recall that this is the same interpretation as in the GCT for the signs of ω and Ω. Finally, k maps to f by (3.12) and n maps to continuous time by the frame interval Δ_t used in computing the auditory spectrogram ($t = n\Delta_t$) such that:

$$y_{CORT}[n,k,\hat{\omega},\hat{\Omega}] \rightarrow y_{CORT}(t,f,\omega,\Omega') \tag{3.23}$$

A key distinction between the GCT and the cortical model lies in their 2-D filter characteristics. Recall that in the GCT, filter bandwidths in the rate-scale space are *fixed* based on the choice of $W[n,m]$; in contrast, the bandwidths of cortical filters are scaled by $\hat{\omega}$ and $\hat{\Omega}$ such that filters with high center frequencies have larger bandwidths than lower center frequencies (3.15). This is manifested in the time-frequency plane as differences in the decay of the impulse responses. We illustrate this with examples of $\mathrm{Re}\{h_c[n,k;\hat{\omega}\rightarrow\omega,\hat{\Omega}\rightarrow\Omega]\}$ shown in Figure 11. These plots are generated with ($\omega = \pm 8$ Hz, $\Omega' = 0.1$ cyc/oct - Figure 11a,b) and ($\omega = \pm 32$ Hz, $\Omega' = 0.5$ cyc/oct - Figure 11c,d). To illustrate the fact that all four impulse responses have the same

absolute lengths in time and frequency based on the DFT lengths (here, 512) used to generate them, we plot them along the sampled axes of time (n) and frequency (k). To map k to absolute frequency, Equation (3.12) can be invoked while n maps to absolute time by the sampling frequency (in this case, 1 ms). Observe that the filters with center frequencies of $\omega = \pm 32$ Hz and $\Omega' = 0.5$ cyc/oct decay faster along both the time and frequency axes than those with center frequencies of $\omega = \pm 8$ Hz and $\Omega' = 0.1$ cyc/oct. This is consistent with the wider bandwidths along rate and scale directions for the filters centered at $\omega = \pm 32$ Hz and $\Omega' = 0.5$ cyc/oct filters shown in Figure 12.

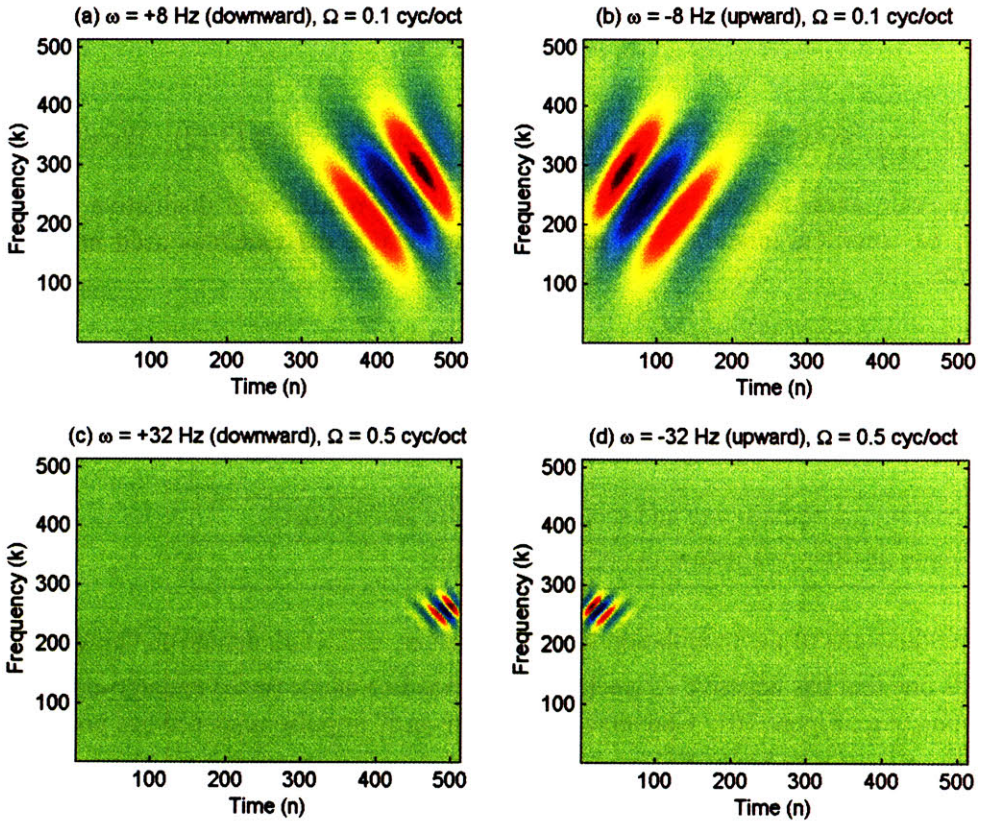

Figure 11 – Real impulse responses of cortical filters centered at [4] (a) 8 Hz, 0.1 cyc/oct (b) -8 Hz, 0.1 cyc/oct, (c) 32 Hz, 0.5 cyc/oct, (d) -32 Hz, 0.5 cyc/oct . Observe the difference in decay along the time and frequency axes between (a, b) and (c, d).

[4] Filter impulse responses have been shifted in time and frequency for display purposes. In the 2-D filtering step, impulse responses are effectively centered at the time and frequency origins.

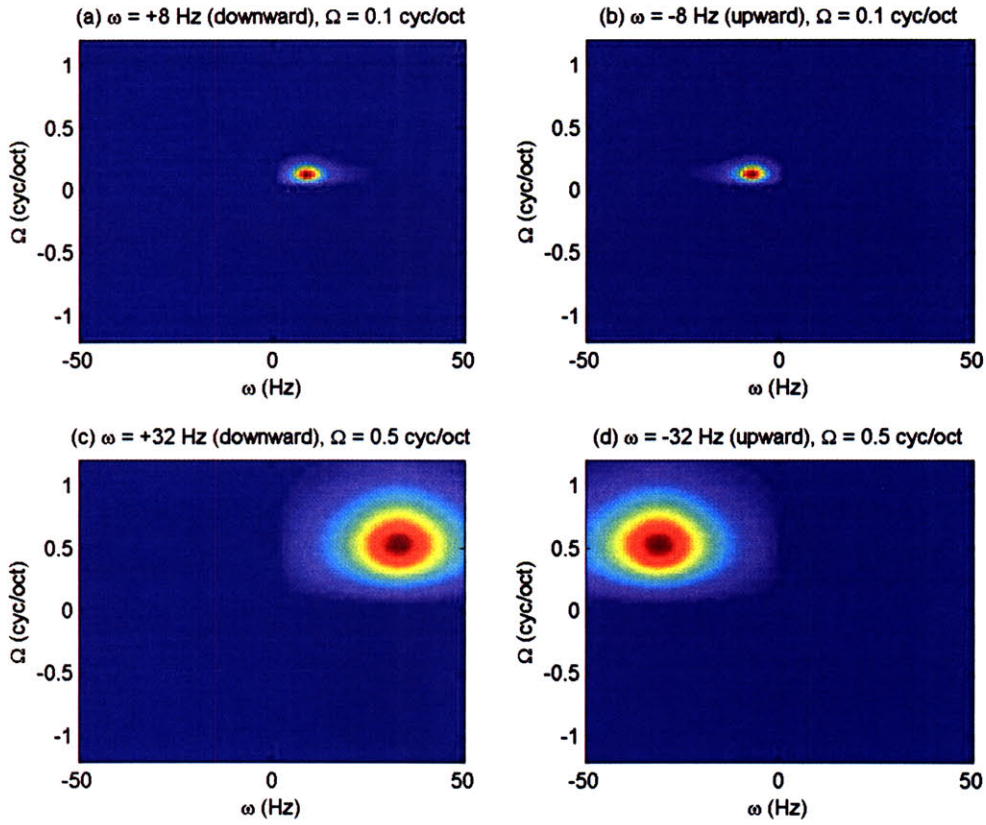

Figure 12 – Magnitude of analytic cortical filters centered at (a) 8 Hz, 0.1 cyc/oct (b) -8 Hz, 0.1 cyc/oct, (c) 32 Hz, 0.5 cyc/oct, (d) -32 Hz, 0.5 cyc/oct. Observe the difference in bandwidth between (a, b) and (c, d).

In summary, we have shown that both the Grating Compression Transform (GCT) and the auditory cortex model analyze spectrotemporal modulations of an input time-frequency distribution. In the GCT, this input is a narrow-band short-time Fourier transform while for the cortical model, it is the auditory spectrogram. At a fixed point in time and frequency, we view the 2-D transform for both methods as the collected outputs of filtering the time-frequency distribution with analytic filters centered at various spectrotemporal modulation frequencies. The magnitudes of these (complex) filtered outputs are 2-D Hilbert envelopes. The envelope amplitude can be interpreted as the strength of matching between the real impulse responses of the filters and a localized region of the time-frequency distribution. Whereas filter bandwidths of the GCT are fixed and determined by the window used in computing 2-D DFT, cortical filters exhibit varying bandwidths based on their center spectrotemporal modulation frequencies. Figure 13 summarizes these two realizations of the proposed framework.

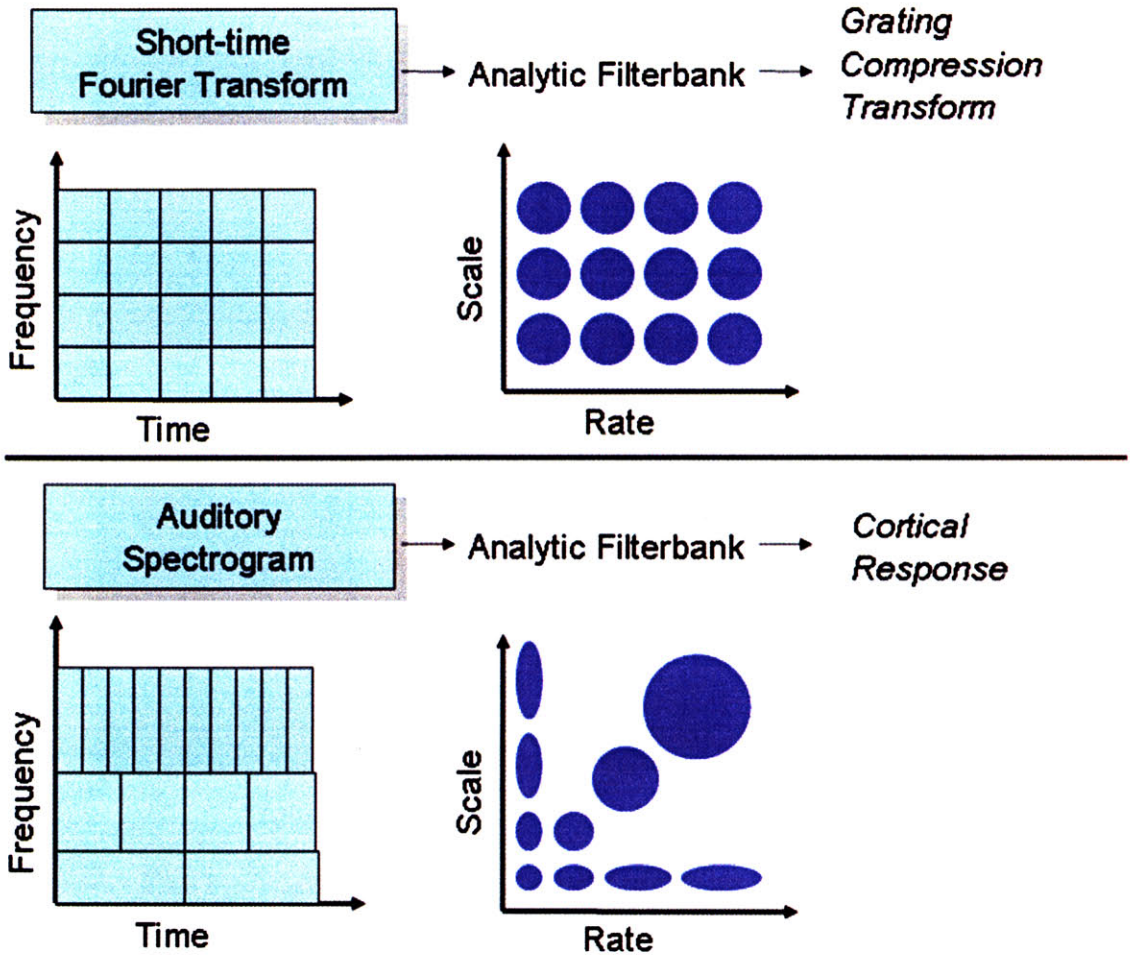

Figure 13 – Schematic comparing the GCT with the auditory cortex model.

3.4 Phenomenological Comparison

In this section, we compare the GCT and auditory cortex model through simulations. Specifically, we aim to assess the value of each realization of the 2-D processing framework in addressing the spectral undersampling problem for high-pitch vowels. One way in which this may be achieved is through improving source-filter *separability* (i.e., Figure 3) in the rate-scale space under conditions of high-pitch speech. We describe our experimental setup in Section 3.4.1 and present results and discussion in Section 3.4.2.

3.4.1 Experimental Setup

We used for a test signal a synthesized female vowel */ae/* with first three formant frequencies F1 = 860 Hz, F2 = 2050 Hz, and F3 = 2850 and formant bandwidths B1 = 54 Hz, B2 = 65 Hz, and B3 = 70 Hz as reported by Stevens in [11]. An all-pole model implementation was used to generate the vocal tract filter coefficients [16]. The vocal tract filter is excited with a pure impulse train source generated at 16 kHz and downsampled to 8 kHz with varying fundamental

frequencies. The duration of the vowel was fixed to 135 ms. Four synthetic vowels were generated with different settings for pitch values:

1. Fixed pitch of 125 Hz.
2. Fixed pitch of 260 Hz.
3. Linearly changing pitch from 235 Hz to 285 Hz.
4. Linearly changing pitch from 210 Hz to 310 Hz.

Conditions 1 and 2 were used to simulate a stationary low- and high-pitch condition. Conditions 3 and 4 were used to simulate a high-pitch condition with moderate to extensive shifts in pitch.

To generate the GCT, a STFT was computed using a 20-ms Hamming window, 1-ms frame interval, and a 2048-point discrete-Fourier transform (DFT). Localized regions were extracted of width 100 ms by 700 Hz, and multiplied by a separable 2-D window consisting of a rectangular window in time and Hamming window in frequency to generate a spectrotemporal "patch"; the patch was then zero-padded to a DFT length of 512 in both directions. As suggested by Quatieri and Ezzat, et al. the DC component of the patch was also removed prior to computing the 2-D DFT so that it would not dominate the GCT response for display purposes [9, 17]. Denoting t_{center} and f_{center} as the center of the patch extracted for the GCT, t_{center} is set to 58 ms corresponding to the middle of the utterance while f_{center} is set to 2100 Hz for analyzing a high-frequency region and 500 Hz for analyzing a low-frequency region.

To generate the cortical response, we used the publicly available NSL MATLAB toolbox developed by Ru and colleagues [18]. The auditory spectrogram was generated by performing the steps outlined in the previous section on the cochlear filterbank outputs $y[n,k]$ (3.12) . The inner hair cell $h_{RC}[n]$ (3.13) and cochlear nucleus $h_{CN}[n]$ (3.14) time constants were set to 0.5-ms and 8-ms time constants, respectively, consistent with physiological observations [19]. The resulting $y_{CN}[n,k]$ was sampled at a 1-ms frame interval to match the STFT computation as described in Equation (3.14). The size of the auditory spectrogram in all simulations was 135 samples in time (i.e., a 135-ms duration and 1-ms frame interval) by 128 in frequency (129 cochlear channels followed by lateral inhibition (3.14)). Cortical filters were designed with linear spacing of 0.25 in the log-modulation-frequency domain. Specifically, rate- and scale-filter center frequencies were computed as

$$\begin{aligned} \omega(p_{rate}) &= \pm 2^{|p_{rate}|}, & p_{rate} &= \pm 1, 1.25, 1.5, \ldots 7 \\ &= 0, & p_{rate} &= 0 \end{aligned} \quad . \tag{3.24}$$

$$\Omega'(p_{scale}) = 2^{p_{scale}}, \qquad p_{scale} = -4, -1.75, -1.5 \ldots 3.5$$

$\omega(p_{rate})$ has a maximum of 128 Hz, to account for the highest possible temporal modulation components present in the auditory spectrogram since an 8-ms time constant was chosen for $h_{CN}[n]$. A-0 Hz rate was also invoked to generate a set of filters located along the scale axis. Given that the filterbank contains 24 channels per octave, we assume a cycle requires more than

two channels to be observed (i.e., a local maximum followed a minimum or vice versa) such that the highest spectral modulation frequency would less than 12 cyc/oct. $\Omega'(p_{scale})$ was chosen to invoke a maximum at ~11 cyc/oct. The rate and scale filter lengths (i.e., $N_{\hat{\omega}}$ and $N_{\hat{\Omega}}$) were set to 512 as in ; filtering was done using the DFT of length 1024 in both time and frequency to avoid aliasing[5]. Observe that the rate component ($H_r(\hat{\omega})$) for the set of filters situated along the scale axis (0 Hz) is

$$\begin{aligned} H_r(\hat{\omega}) &= 1 \qquad \hat{\omega} = 0 \\ &= 0 \qquad \textit{otherwise} \end{aligned} \tag{3.25}$$

which therefore extracts the average[6] of all spectral slices of the auditory spectrogram since the DFT length ($N_{\hat{\omega}} = 512$) is greater than the auditory spectrogram in time. This set of scale-only filters operates along the frequency axis of this averaged spectral slice. In the subsequent plots, the magnitude of the cortical response is shown.

With t_{center} and f_{center} denoting the center of the patch extracted for the GCT, the corresponding cortical response is $y_{CORT}(t_{center}, f_{center}, \omega, \Omega')$ to provide as fair of a comparison as possible. t_{center} is set to 58 ms corresponding to the middle of the utterance while f_{center} is set to 2100 Hz for analyzing a high-frequency region and 500 Hz for analyzing a low-frequency region. In addition, note that the cortical response does not contain negative spectral modulation frequencies whereas the GCT displays both positive and negative scale. Finally, we observe that an even more fair comparison than that presented here would involve both frameworks performing analysis on their respective time-frequency distributions with the *same* number of 2-D filters. The GCT's 512 point DFT results in a bank of ~130,000 filters for the positive-scale region; because ~130,000 cortical filters would be computationally prohibitive, our simulations of the cortical response invoked instead ~780 cortical filters.

3.4.2 Results

Fixed 125-Hz pitch case: For the fixed $f_0 = 125$ Hz condition, Figure 14 compares the output of the GCT (top) with that of the auditory model (bottom) in a high-frequency region of the time-frequency distribution at $f_{center} = 2102$ Hz; similarly, Figure 15 shows this comparison for a low-frequency region at $f_{center} = 500$ Hz. In both cases, analysis is performed at $t_{center} = 58$ ms.

As would be expected for the localized region in both high- and low-frequency regions of the STFT (b), a harmonic line structure is present with periodicity (along the frequency axis) of 125-Hz corresponding to the pitch of the synthetic source. In the corresponding GCT (d), we observe a distinct pair of peaks at $\Omega \approx \pm\frac{1}{125} = \pm 0.008$ cyc/Hz (dashed arrow), consistent with this scale component. Observe also in the GCT a pair of slower spectral modulation components at $\Omega \approx \pm 0.0025$ cyc/Hz (solid arrow, Figure 14d) and $\Omega \approx \pm 0.0015$ cyc/Hz (solid arrow, Figure

[6] To within a scale factor since 512 > 135.

15d). These may correspond to the windowed portions of the stationary formant envelope. Under this assumption, the GCT appears then to provide separability of the pitch from formant structure in the modulation space. The difference in location along the scale axis for the slower modulation components presumably reflects the portion of the formant envelope extracted within each patch. The high-frequency patch is centered ($f_{center} = 2102$ Hz) close to a formant peak (F3 = 2050 Hz) and would exhibit sharp transitions in the time-frequency space. In contrast, the patch of the low-frequency region ($f_{center} = 500$ Hz) is located 360 Hz away from the first formant peak (F1 = 860 Hz) would expected to be smoother.

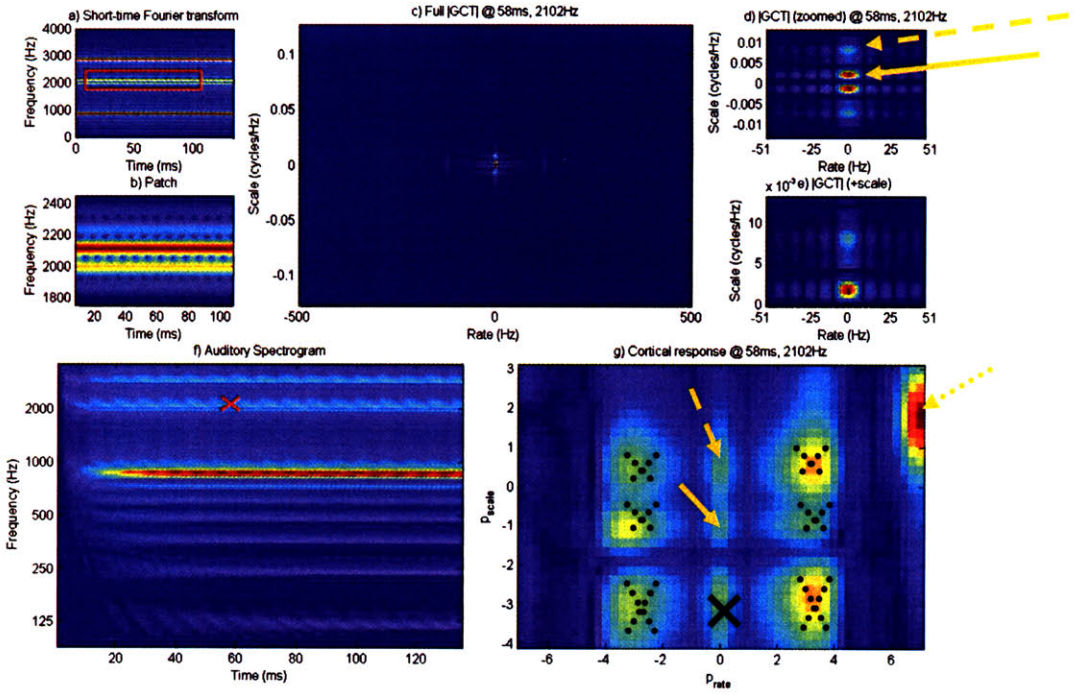

Figure 14[7] – */ae/* with 125-Hz pitch; a) STFT with localized region (rectangle) centered at $t_{center} = 58$ ms and $f_{center} = 2102$ Hz; b) Zoomed-in view of localized region of a); c) Full GCT magnitude; d) Zoomed-in version of c) to show dominant spectrotemporal components; dashed arrow – 0.008 cyc/Hz; solid arrow – 0.0025 cyc/oct; e) As in c) but with negative-scale removed for comparison to cortical response; f) Auditory spectrogram with 'X' denoting $t_{center} = 58$ ms and $f_{center} = 2102$ Hz of cortical response; g) Full cortical response magnitude; dashed arrow – 1.4 cyc/oct; solid arrow – 0.5 cyc/oct; dotted arrow – 125 Hz, 3.4 cyc/oct; solid X – 0.125 cyc/oct; dotted X's – $\pm$ 8 Hz.

[7] The rate and scale axes shown are plotted on a linear axis for display purposes. The mapping to absolute rate (in Hz) and scale (cyc/oct) was previously described in Equation (3.24).

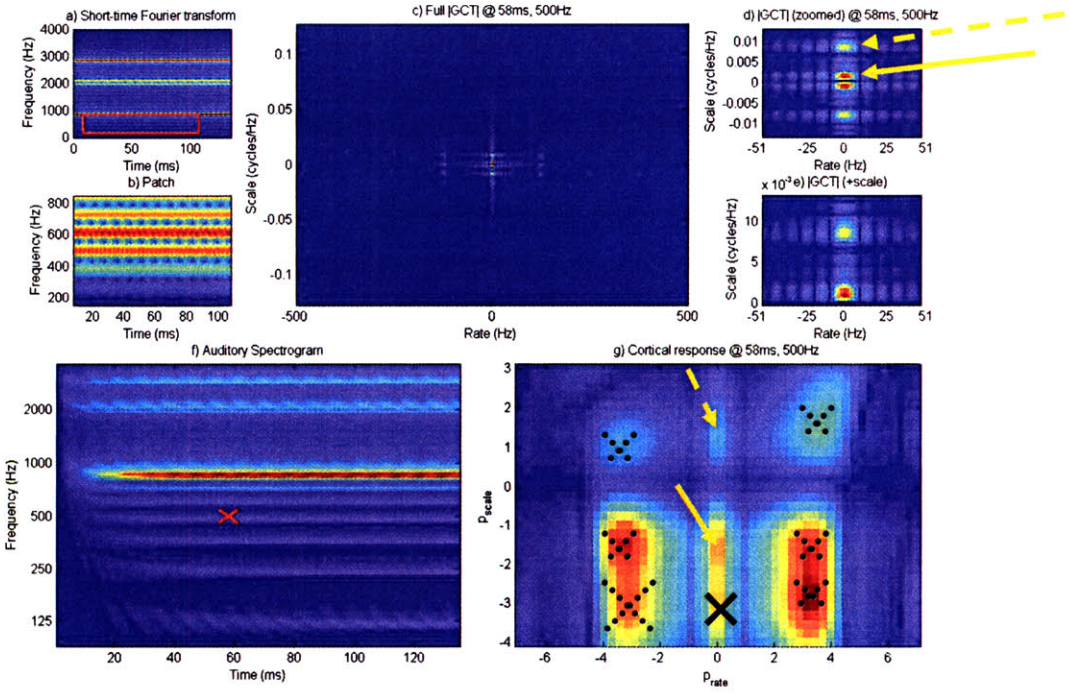

Figure 15 – */ae/* with 125-Hz pitch as in Figure 14 but with $f_{center} = 500$ Hz; (d) dashed arrow – 0.008 cyc/Hz; solid arrow – 0.0015 cyc/Hz; (g) dashed arrow – 2.4 cyc/oct; solid arrow – 0.35 cyc/oct; dotted X's – ± 8 Hz; solid X – 0.125 cyc/oct.

In contrast to the GCT, the octave-spacing of the cochlear filterbank invokes a non-uniform spacing of the 125-Hz pitch harmonics in the auditory spectrogram (f). In addition, observe that frequency (temporal) resolution decreases (increases) towards high-frequency regions of the auditory spectrogram due to the increase in bandwidth of the cochlear filters as *k* increases. We can therefore think of low-frequency regions as resembling a narrow-band STFT while high-frequency regions resemble a wide-band STFT.

In (f) 'X' denotes $t_{center} = 58$ ms, $f_{center} = 2102$ Hz in Figure 14 and $t_{center} = 58$ ms, $f_{center} = 2102$ Hz in Figure 15 such that the corresponding cortical response magnitude in (g) is $\left| y_{CORT}(t_{center}, f_{center}, \omega, \Omega') \right|$. The cortical response corresponding to the high-frequency region (Figure 14g) contains three dominant components along the scale axis (0 Hz) at $p_{scale} \approx -3 \approx 0.125$ cyc/oct (X), $p_{scale} \approx -1 \approx 0.5$ cyc/oct (solid arrow), and $p_{scale} \approx 0.5 \approx 1.4$ cyc/oct (dashed arrow). In addition, a distinctive peak is located at $p_{rate} \approx +6.9 \approx 125$ Hz, $p_{scale} \approx 1.75 \approx 3.4$ cyc/oct (dotted arrow). Similarly, the low-frequency region's cortical response (Figure 15g) contains three components along the scale axis at $p_{scale} \approx -3 \approx 0.125$ cyc/oct (X), $p_{scale} \approx -1.5 \approx 0.35$ cyc/oct (solid arrow), and $p_{scale} \approx 1.25 \approx 2.4$ cyc/oct (dashed arrow). Finally, observe that for both frequency regions, the cortical response exhibits several moderate-rate (e.g., $p_{rate} \approx +3 \approx 8$ Hz) components located at the same scales as those components along the scale axis (dotted X's).

To interpret the component located at 125 Hz and 3.4 cyc/oct (dotted arrow, Figure 14g), we show in Figure 16a zoomed-in portion of the auditory spectrogram near $t_{center} = 58$ ms, $f_{center} = 2102$ Hz. In Figure 16b, we show the real impulse response corresponding to the filter centered at $\omega = 128$ Hz, $\Omega' = 3.4$ cyc/oct. Recall that the magnitude cortical response shown in Figure 14g denotes the strength of the match between this real impulse response at the point $t_{center} = 58$ ms, $f_{center} = 2102$ Hz of the auditory spectrogram. For plotting purposes, we have mapped the absolute frequency (f) scale to octaves (oct) by inverting Equation (3.12):

$$oct = \log_2 \frac{f}{220}. \tag{3.26}$$

From Figure 16a, it is clear that a spectrotemporal modulation component is present in the auditory spectrogram with periodicity ~8 ms, consistent with the temporal modulation frequency of the real impulse response shown in Figure 16b. This also corresponds to the pitch value used in synthesis. Recall that high-frequency regions of the auditory spectrogram provide improved temporal resolution due to the broadening of cochlear filters; this improved resolution is presumably necessary to observe this "phase-locked" response of the pitch. In particular, observe that the cortical response corresponding to the low-frequency region (with better frequency resolution but poorer temporal resolution) does *not* exhibit this high-rate component.

We observe that the orientation of the impulse response and modulation component in Figure 16a and Figure 16b are skewed. From Figure 10, we can estimate $\hat{\theta}$ (as measured counter-clockwise from the scale-axis) by mapping ω and Ω' back to $\hat{\omega}$ and $\hat{\Omega}$ such that:

$$\begin{aligned} \hat{\Omega} &\approx \frac{(3.4)(2\pi)}{24} \approx 0.89 \\ \hat{\omega} &\approx \frac{(125)(2\pi)}{1000} \approx 0.79 \\ \hat{\theta} &\approx -\tan^{-1}\left(\frac{\hat{\omega}}{\hat{\Omega}}\right) \approx -41.6^{o} \end{aligned} \tag{3.27}$$

From Figure 16c, we can also measure (using samples) the corresponding angle in the auditory spectrogram of this component to be:

$$\begin{aligned} \Delta n &\approx 7 \\ \Delta k &\approx 6.5 \\ \hat{\theta} &\approx -\tan^{-1}\left(\frac{\Delta k}{\Delta n}\right) \approx -42.9^{o} \end{aligned} \tag{3.28}$$

The estimates of $\hat{\theta}$ are similar, thereby confirming the correspondence between the peak in the cortical response and the component observed in the auditory spectrogram.

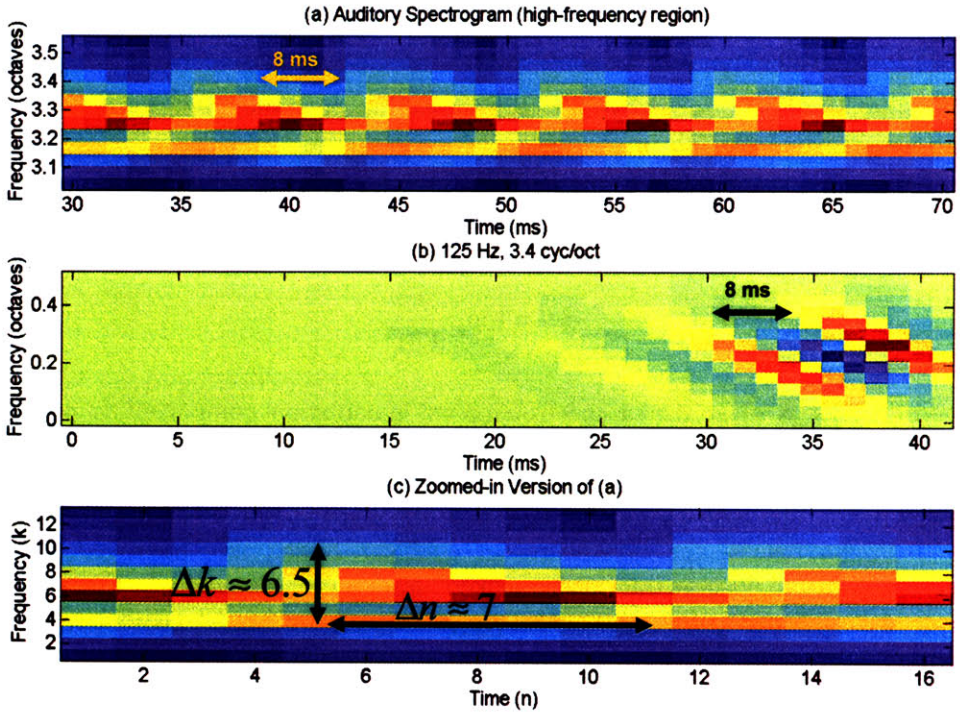

Figure 16 - (a) Zoomed in portion of auditory spectrogram near $t_{center} = 58$ ms and $f_{center} = 2102$ Hz (oct = 3.25) with (b) zoomed-in real impulse response in time-frequency plane of cortical filter 125 Hz; (c) Zoomed-in version of (a) for estimating $\hat{\theta}$.

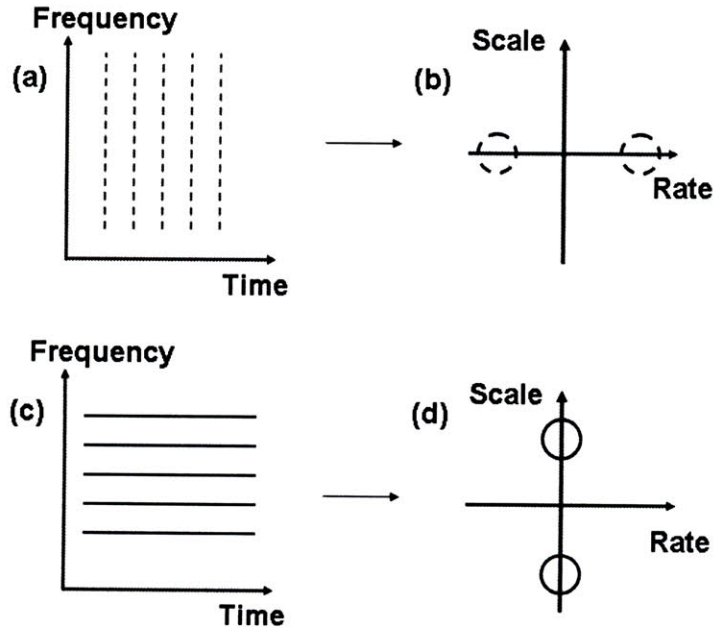

Figure 17 – Comparison of wide-band time-frequency distribution (a) and its corresponding 2-D transform (b) to a narrow-band time-frequency distribution and its 2-D transform (d).

Recall that high-frequency regions of the auditory spectrogram can be thought of resembling a wide-band STFT while low-frequency regions are similar to a narrow-band STFT. For pitch, a typical wide-band STFT emphasizes temporal resolution such that individual pitch pulses would be oriented vertically (in time). In contrast, a narrow-band STFT emphasizes frequency resolution such that pitch harmonics are resolved and oriented horizontally. If we consider a 2-D transform of these regions, a wide-band-like region will then map pitch component along the rate axis while a narrow-band-like region will map it along the scale axis (Figure 17). We can therefore interpret the estimate of as indicating the trade-off between temporal and frequency resolution; the auditory spectrogram in the high-frequency region shown therefore resembles a time-frequency distribution "between" a broad- and narrow-band since $\hat{\theta}$ is neither 0 nor $-\frac{\pi}{2}$ (as measured counter-clockwise from the scale axis).

To interpret components located along the scale axis in the cortical response, recall that the scale filters operate on the average of all spectral slices from the auditory spectrogram. Figure 18 shows this averaged spectral slice where we have again invoked the frequency-to-octave mapping of (3.26) for the frequency axis. Concurrently, we also show both the formant envelope and locations of the 125-Hz pitch harmonics. Observe that the locations of the harmonics are no longer uniformly spaced and that the formant envelope appears warped (compare to the formant envelope shown in Figure 4); both effects are due to the logarithmic spacing of the octave scale. In low-frequency regions of the spectral slice, local maxima correspond to pitch harmonic locations; in contrast, high-frequency regions do not exhibit such a correspondence. This reflects the loss of frequency resolution in high-frequency regions due to the widening of cochlear filter bandwidths.

In Figure 19, we plot the real impulse responses of filters with center (scale) frequencies of the observed local maxima in Figure 14g and Figure 15g. Recall from our previous discussion that these responses can be viewed as matched filters for components along the frequency axis of the auditory spectral slice. For interpretation purposes, note that filtering in the scale-domain is implemented *non-causally*. The filtered output (i.e., cortical response) at f_{center} is generated when the *peak* of the impulse responses shown in Figure 19 are aligned with f_{center} (i.e., H and L in Figure 18) [18].

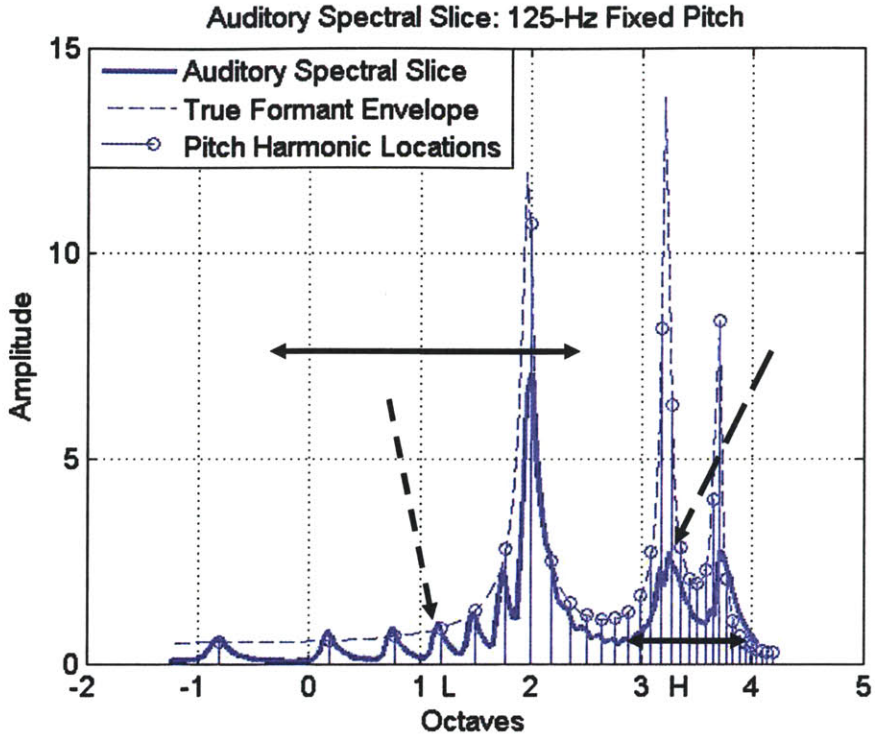

Figure 18 – Average of spectral slices of auditory spectrogram. H and L denote 2100 Hz (3.25 oct) and 500 Hz (1.18 oct) mapped to octaves, respectively. Dashed arrow (near L) – Peak of the 2.4 cyc/oct component; horizontal solid arrow (near L) – 0.35 cyc/oct component; solid arrow (near H) – peak of 1.4 cyc/oct component; horizontal solid arrow (near H) – peak region of 0.5 cyc/oct component.

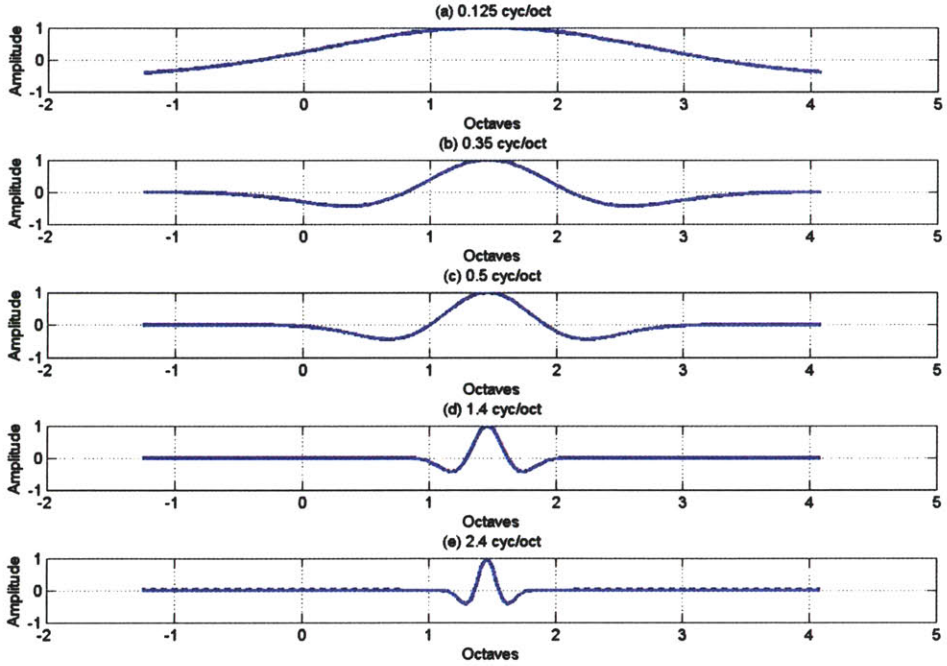

Figure 19 – Real impulse responses corresponding to filters along the scale axis.

For the low-frequency region of analysis (L, Figure 18), the 2.4 cyc/oct component (Figure 19e) corresponds to the harmonic peak located at L (dashed arrow (near L), Figure 18). Observe that because the impulse response consists of a single cycle, this component in the cortical response does not represent a coherent mapping of the surrounding harmonic peaks in this region. This is in contrast to the mapping of the harmonic line structure in the localized region of STFT to the coherent components observed in the GCT. The 0.35 cyc/oct (Figure 19b) likely reflects matching of the local portion of the formant envelope swept out by the pitch harmonics ranging from 0 – 2 (horizontal solid arrow (near L), Figure 18). For the high-frequency region (H, Figure 18) the 1.4 cyc/oct component corresponds to the peak located between 2.5 – 3.5 octaves (dashed arrow (near H), Figure 18) and resembles the resonant peak of the formant envelope in this local region. We can also view the region from 3 – 4 octaves as the peak region of a single cycle (horizontal solid arrow (near H), Figure 18) corresponding to the 0.5 cyc/oct component (Figure 19c). This component can be interpreted as "coarse" formant structure in comparison to the "fine" formant structure of the 1.4 cyc/oct component. The 0.125 cyc/oct component present in both the low- and high-frequency regions of analysis has a notably wide extent in frequency (Figure 19a) and is therefore DC-like in representing the overall amplitude across frequency of the entire spectral slice.

The remaining components in the cortical responses for both the high- and low-frequency regions have moderate rates (e.g., 8 Hz) (Figure 14g and Figure 15g, dotted X) and can be interpreted similarly as the 0.125 cyc/oct component along the scale axis. From Figure 11, recall that the impulse responses of the 8 Hz cortical filters has duration ~170 ms, thereby exceeding the duration of the entire utterance. We therefore interpret components with rates ~8 Hz or smaller as DC-like representations of the entire auditory spectrogram across time. This interpretation is consistent with the components' corresponding scale frequencies matching those along the scale axis.

In summary, for the fixed 125-pitch condition, the GCT appears to afford separability of pitch harmonics from localized formant structure along the scale axis. The primary evidence for separability is that the set of pitch harmonics in a localized region are mapped *coherently* to a set of impulses in the GCT domain and are separated from another set of components presumably corresponding to formant structure. The cortical response also decomposes the auditory spectrogram into distinct components; this decomposition is dependent on the frequency region analyzed. In the high-frequency region, due to improved temporal resolution, pitch can be mapped to a high-rate component *off* the scale axis leaving components along the scale axis resembling both coarse and fine formant structure, thereby suggesting source-filter separability. A caveat of this interpretation is that the high-frequency region is not strictly wide-band ($\hat{\theta} \approx -42^{o}$); therefore, we cannot rule out contributions from pitch harmonics to components along the scale axis. In the low-frequency region, single harmonic peaks are mapped to high-scale components along the scale axis while low-scale components resemble formant structure swept out by several harmonic peaks. Because we did not observe a coherent mapping of multiple pitch harmonics to the cortical response, it is more difficult to rule out their contribution to the low-scale component matched to formant structure as was argued in the case of the GCT.

Fixed 260-Hz pitch case: In Figure 20 and Figure 21, we show a comparison between the GCT and the magnitude cortical response for the vowel with fixed $f_0 = 260$ Hz. As in the low-pitch condition, we observe a pair of local maxima present in the GCT (d) at $\Omega \approx \pm\frac{1}{260} = \pm 0.00385$ cyc/Hz (solid arrows) that correspond to the harmonic line structure. Another set of peaks (dashed arrows) are located at lower scales of $\Omega \approx \pm 0.0025$ cyc/Hz for the high-frequency

region (Figure 20d) and $\Omega \approx \pm 0.0015$ cyc/Hz in the low-frequency region (Figure 21d); recall that these peaks were also observed in the low-pitch case and may correspond to portions of the formant envelope. The distance between those peaks corresponding to the harmonic line structure and this second set has been reduced due to the increase of pitch and appear merged. The GCT therefore appears to suffer a similar problem as the cepstrum in source-filter separability: the components in the transformed space corresponding to pitch are moved closer to those that presumably represent formant structure.

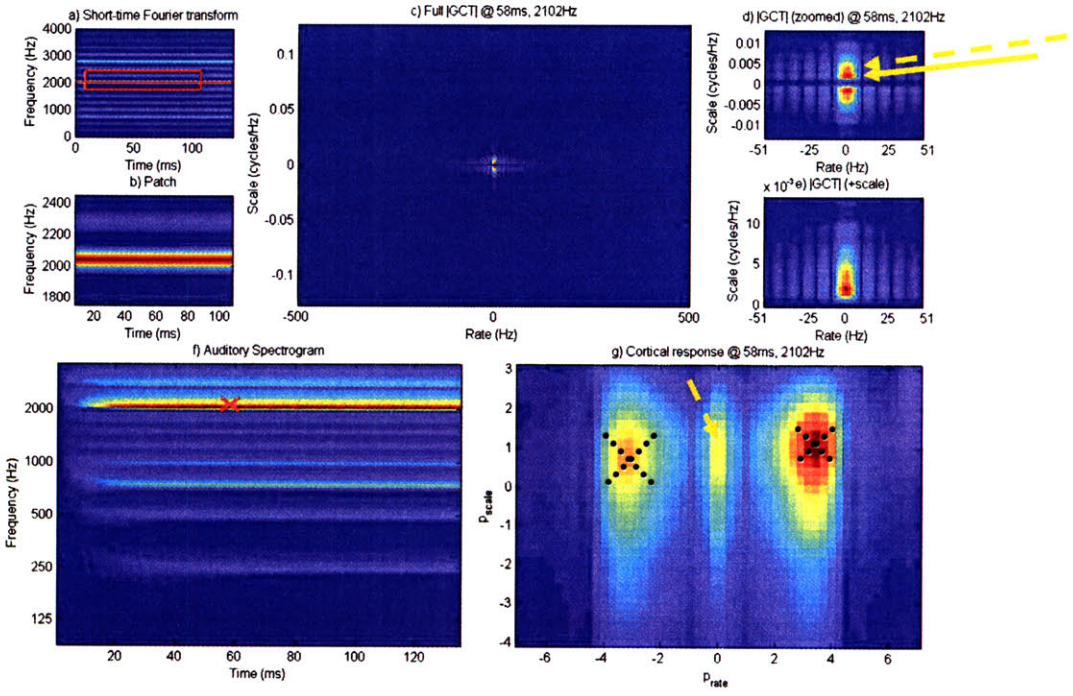

Figure 20 – /*ae*/ with 260-Hz pitch; a) STFT with localized region (rectangle) centered at $t_{center} = 58$ ms and $f_{center} = 2102$ Hz; b) Zoomed-in view of localized region of a); c) Full GCT magnitude; d) Zoomed-in version of c) to show dominant spectrotemporal components; dashed arrow – 0.00385 cyc/Hz, solid arrow 0.0025 cyc/Hz; e) As in c) but with negative-scale removed for comparison to cortical response; f) Auditory spectrogram with 'X' denoting $t_{center} = 58$ ms and $f_{center} = 2102$ Hz of cortical response; g) Full cortical response magnitude; dashed arrow – 2 cyc/oct; dotted X's – $\pm$ 8 Hz.

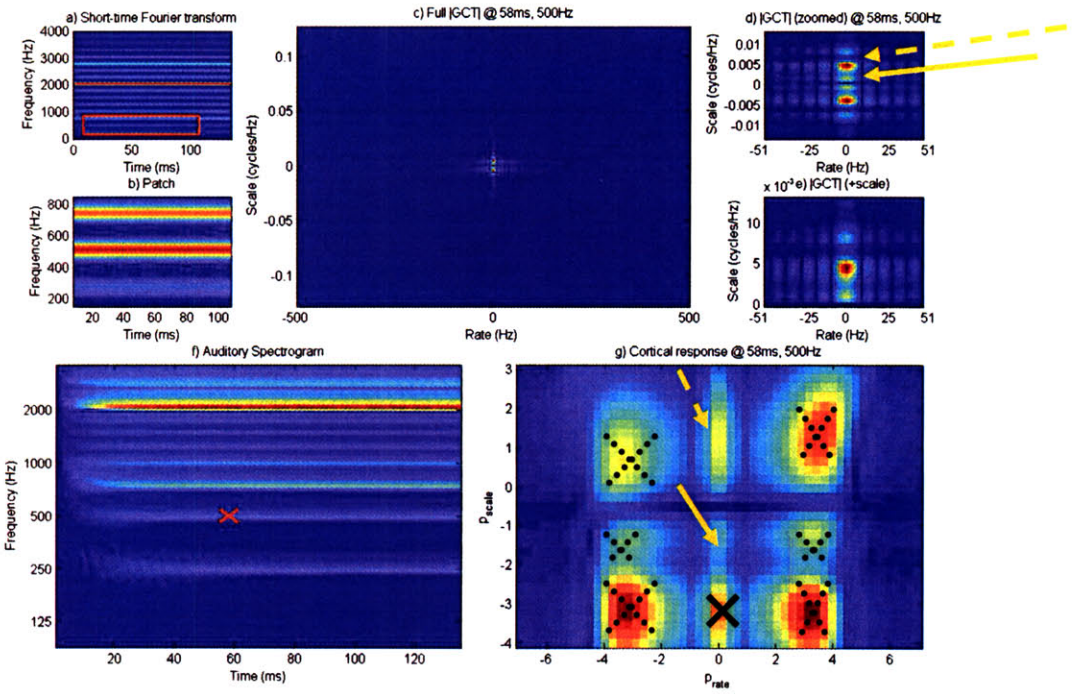

Figure 21 – /*ae*/ with 260-Hz pitch as in Figure 20 but with $f_{center} = 500$ Hz; (d) dashed arrow – 0.00385 cyc/Hz; solid arrow – 0.0015 cyc/oct; (g) Full cortical response magnitude; dashed arrow – 2 cyc/oct; solid arrow – 0.4 cyc/oct; solid X – 0.125 cyc/oct; dotted X's – ± 8 Hz; solid x – 0.125 cyc/oct.

In the auditory spectrogram (f), the higher pitch results in more widely spaced harmonics as in the STFT. In the corresponding cortical responses for the high- and low-frequency regions (g), we again observe DC-like components at moderate rates (i.e., dotted X's) and at 0.125 cyc/oct along the scale axis in Figure 21 (solid X). For the high-frequency region (Figure 20), the cortical response (g) exhibits a single component along the scale axis at $p_{scale} \approx 1 \approx 2$ cyc/oct. For the low-frequency region (Figure 21g), two components are observed along the scale axis at $p_{scale} \approx -1.25 \approx 0.4$ cyc/oct (solid arrow) and $p_{scale} \approx 1.25 \approx 2.4$ cyc/oct (dashed arrow).

In Figure 22a, we show a plot of the averaged (across time) auditory spectrogram, pitch harmonic locations, and true formant envelope as in Figure 18 but for the 260-Hz pitch case. Figure 22b and c show the real impulse responses for scale filters centered at 0.4 and 2 cyc/oct, respectively; we refer the reader to Figure 19 for the 2.4 cyc/oct impulse response. Observe that the pitch harmonics are more widely spaced along frequency as can be expected for a higher pitch. The 2 cyc/oct component of the high-frequency region corresponds to peak located at H (dashed arrow (near H), Figure 22a) and appears to match the resonant peak of the true formant envelope. In the low-frequency region, the 2.4 cyc/oct component corresponds to the single harmonic peak at L spanning 1 – 1.5 octaves (dashed arrow (near L), Figure 22a). As in the 125-Hz case, we interpret the low-scale component of 0.4 cyc/oct (Figure 22c) as matching the portion of the formant envelope spanned by the pitch harmonics from 0.5 – 3 octaves (horizontal solid arrow (near L), Figure 22a).

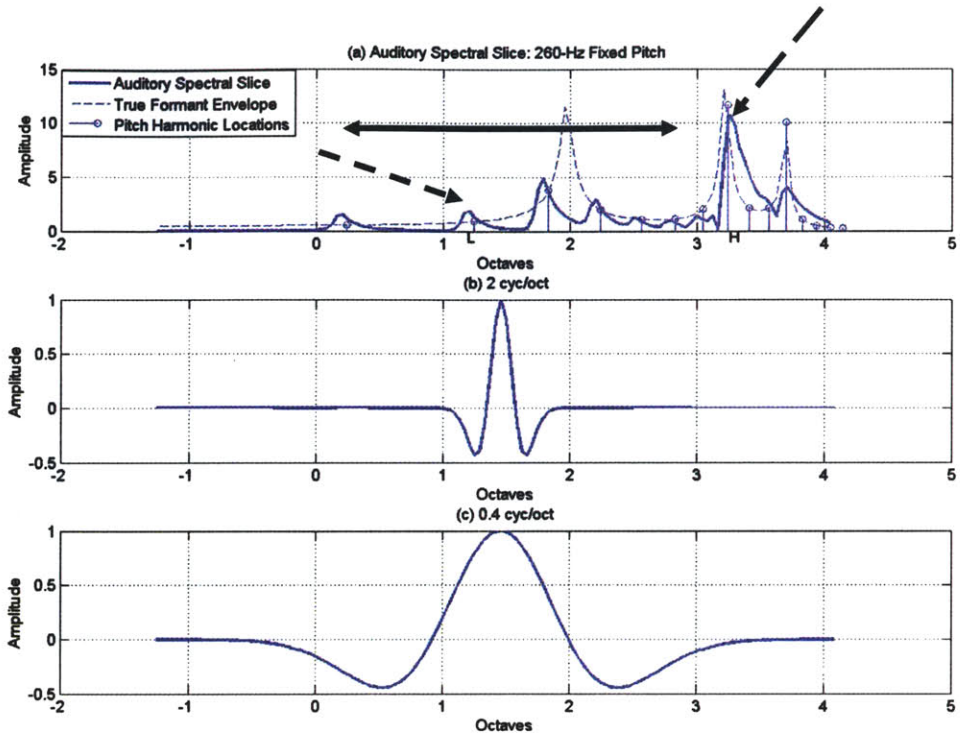

Figure 22 – (a) Auditory spectral slice for the 260-Hz pitch */ae/*; L and H denote the low- and high-frequency regions, respectively; dashed arrow (near L) – peak of 2.4 cyc/oct component; horizontal solid arrow (near L) – 0.4 cyc/oct component; dashed arrow (near H) – peak of 2 cyc/oct component; (b, c) Real impulse responses of scale filters with 2 and 0.4 cyc/oct, respectively.

In summary, for the fixed 260-Hz condition, the GCT appears to suffer a similar problem as the cepstrum. Specifically, harmonic structure is mapped closer to the presumed formant structure along the scale axis, independent of frequency region. In the cortical response, a single component was located along the scale axis for the high-frequency region and coincided closely to a resonant peak of the formant envelope; however, recall from our previous discussion that we cannot rule out contributions to this component from pitch harmonics. Note that for this region, the phase-locked pitch response observed in the 125-Hz case was not observed because 260 Hz is greater than the low-pass cut-off (125 Hz) of the cochlear nucleus ($h_{CN}[n]$). A smaller time constant for $h_{CN}[n]$ would have invoked a temporal modulation component in the cortical response at 260 Hz, thereby affording the type of separability observed in the 125-Hz pitch case. For the low-frequency region, as in the 125-Hz pitch case, we did not find evidence for source-filter separability due to the lack of a coherent mapping of harmonic structure.

Changing pitch, 235-285 Hz and 210-310 Hz: Figure 23 and Figure 24 show the results of analysis for the high- and low-frequency regions for a pitch shift of 235-285 Hz for the vowel, respectively while Figure 25 and Figure 26 present the same analysis for the 210-310 Hz condition. The skewed harmonic line structure corresponding to pitch observed in (b) is mapped to coherent components *off* the scale axis in the GCT (dashed arrows, d). For a fixed frequency region, harmonic structure is mapped to a larger angle ($\hat{\theta}$) off the scale axis for a larger pitch change (e.g., compare Figure 24d with Figure 26d). Observe also that for a fixed pitch change, harmonic structure in the high-frequency regions is rotated further off the scale axis than in low-

frequency regions (compare Figure 23d with Figure 24d). These observations are consistent with our discussion of the GCT for changing pitch (Section 3.2). As in fixed pitch conditions observe that a set of modulation components remain *along* the scale axis at $\Omega \approx \pm 0.0025$ cyc/Hz (Figure 23d, Figure 25d) and $\Omega \approx \pm 0.0015$ cyc/Hz (Figure 24d, Figure 26d) (solid arrows). Since the formant envelope remains stationary across time, this observation provides further evidence that they correspond to a local portion of the stationary formant. Overall then, it appears that invoking pitch change can improve source-filter separability in the GCT even under conditions of high pitch. Nonetheless, observe that this separation is not entirely complete under certain conditions. For instance, in the low-frequency region of the 235 – 285 Hz case, the two sets of peaks appear merged.

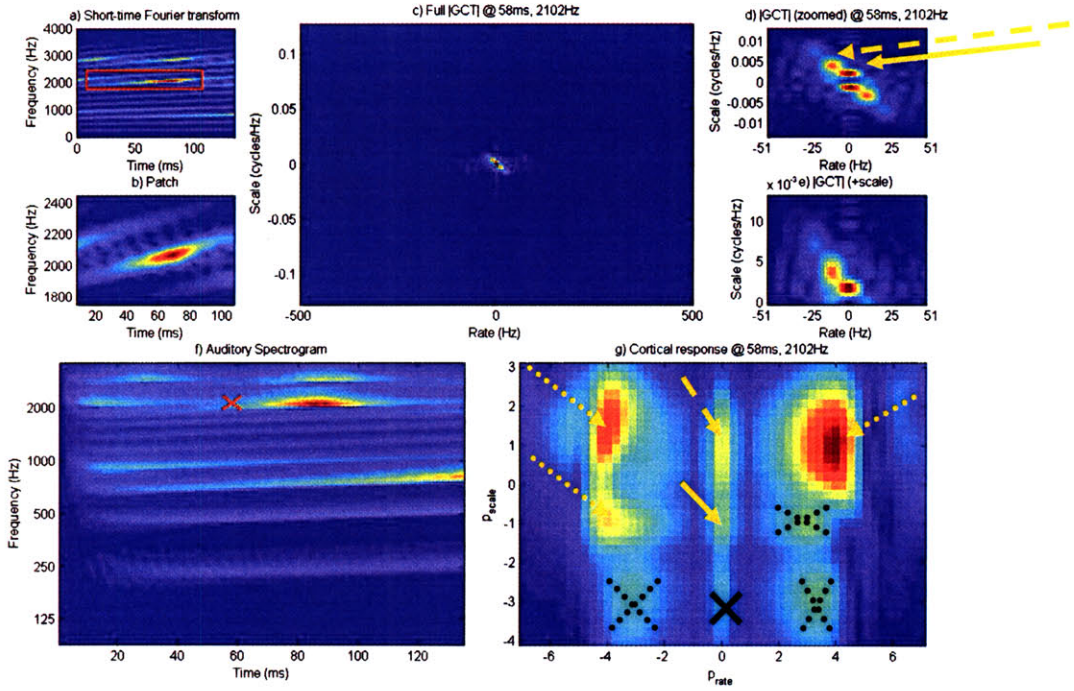

Figure 23 – /*ae*/with changing pitch from 235 – 285 Hz; a) STFT with localized region (rectangle) centered at $t_{center} = 58$ ms and $f_{center} = 2102$ Hz; b) Zoomed-in view of localized region of a); c) Full GCT magnitude; d) Zoomed-in version of c) to show dominant spectrotemporal components; solid arrow – 0.025 cyc/Hz; dashed arrow – rotated component corresponding to skewed harmonic structure; e) As in c) but with negative-scale removed for comparison to cortical response; f) Auditory spectrogram with 'X' denoting $t_{center} = 58$ ms and $f_{center} = 2102$ Hz of cortical response; g) Full cortical response magnitude; dashed arrow – 2 cyc/oct; solid arrow – 0.5 cyc/oct; dotted arrows – $\pm$ 16 Hz; solid X – 0.125 cyc/oct; dotted X's – $\pm$ 8 Hz.

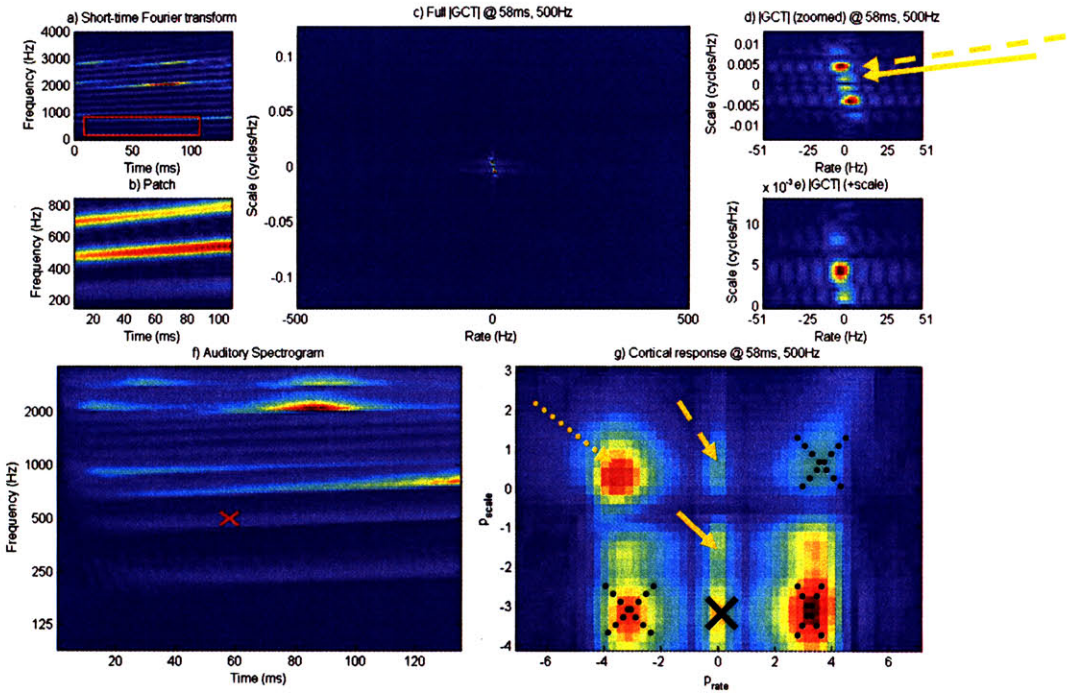

Figure 24 – */ae/*with changing pitch from 235 – 285 Hz as in Figure 23 but with $f_{center} = 500$ Hz; d) solid arrow – 0.015 cyc/Hz; dashed arrow – rotated component corresponding to skewed harmonic structure; g) dashed arrow – 1.4 cyc/oct; solid arrow – 0.4 cyc/oct; dotted arrow – $\pm$ 16 Hz; solid X – 0.125 cyc/oct; dotted X's - $\pm$ 8Hz.

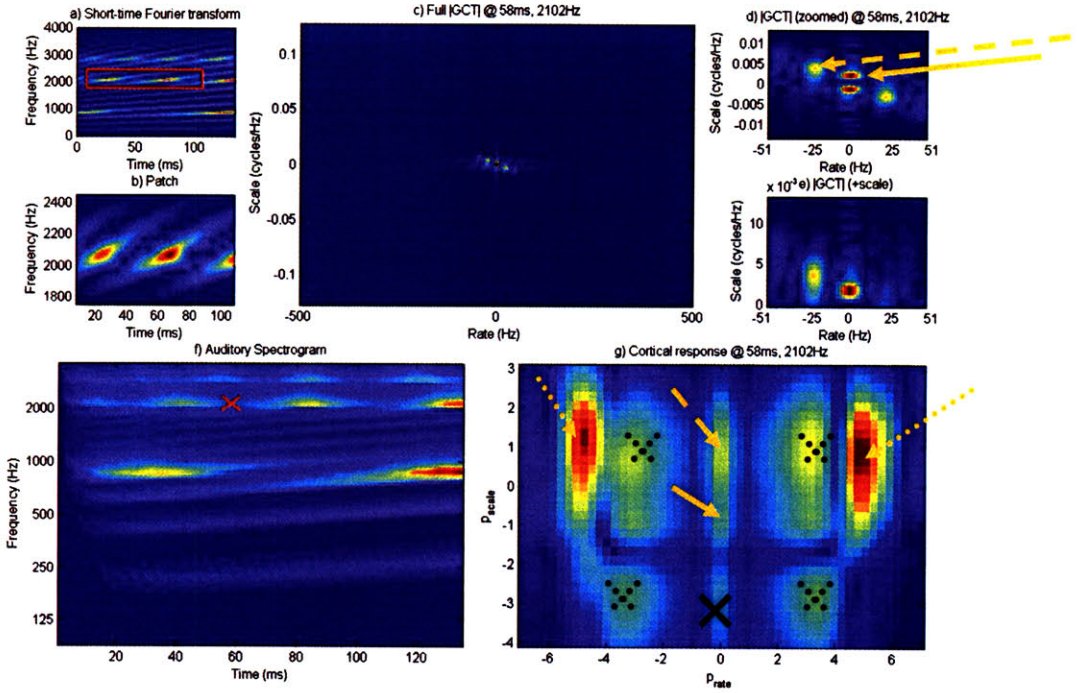

Figure 25 – */ae/* with changing pitch from 210 – 310 Hz; a) STFT with localized region (rectangle) centered at $t_{center} = 58$ ms and $f_{center} = 2102$ Hz; b) Zoomed-in view of localized region of a); c) Full GCT magnitude; d) Zoomed-in version of c) to show dominant spectrotemporal components; solid arrow – 0.025 cyc/Hz; dashed arrow – rotated component corresponding to skewed harmonic structure; e) As in c) but with negative-scale removed for comparison to cortical response; f) Auditory spectrogram with 'X' denoting $t_{center} = 58$ ms and $f_{center} = 2102$ Hz of cortical response; g) Full cortical response magnitude; dashed arrow – 2 cyc/oct; solid arrow – 0.7 cyc/oct; dotted arrows – $\pm$ 32 Hz; solid X – 0.125 cyc/oct; dotted X's - $\pm$ 8 Hz.

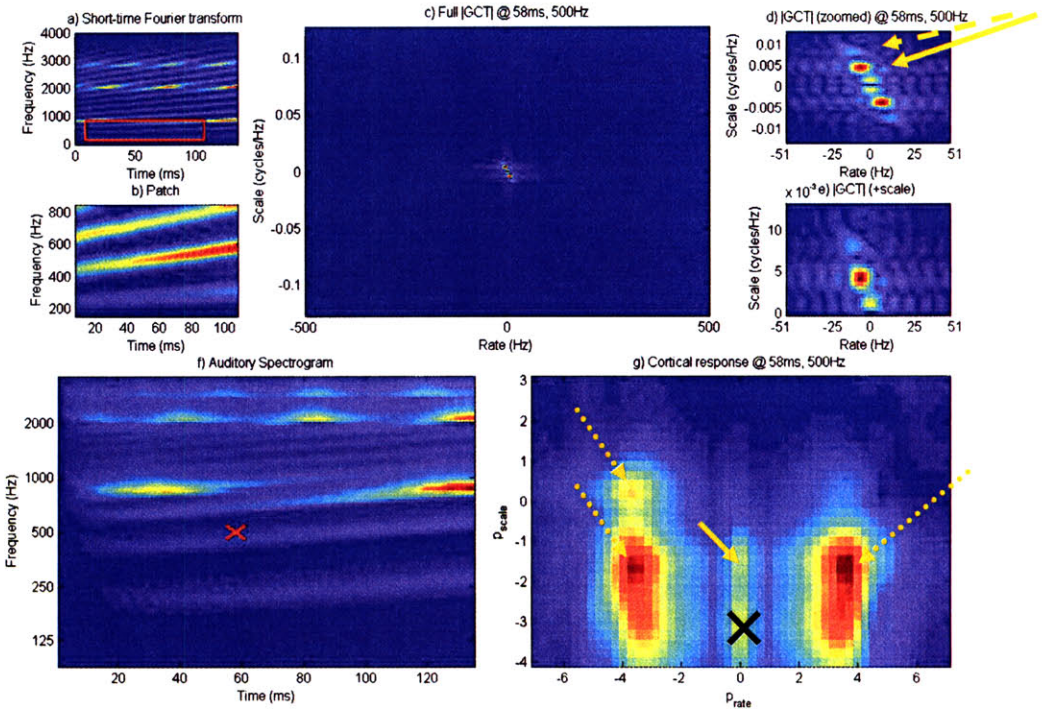

Figure 26 – */ae/*with changing pitch from 210 – 310 Hz as in Figure 25 but with $f_{center} = 500$ Hz; d) solid arrow – 0.015 cyc/Hz; dashed arrow – rotated component corresponding to skewed harmonic structure; g) solid arrow – 0.4 cyc/oct; dotted arrows – $\pm$ 16 Hz; solid X – 0.125 cyc/oct.

In the auditory spectrogram (f), changing pitch invokes a similar skew of the harmonic line structure. In the corresponding cortical responses, this is manifested by the higher-rate components off the scale axis (dotted arrows, (g)). For the 235 – 285 Hz condition, these components are at $p_{rate} \approx \pm 4 \approx \pm 16$ Hz for the low-and high-frequency regions; for the 210 – 310 Hz case, they are at $p_{rate} \approx \pm 5 \approx \pm 32$ Hz in the high-frequency region and $p_{rate} \approx \pm 4 \approx \pm 16$ Hz in the low-frequency region. Since the duration of the vowel is fixed, the 210 – 310 Hz case will invoke a faster rate of pitch change than the 235 – 285 Hz condition. This is consistent with the (faster) 32 Hz rate components in the 210 – 310 Hz case while the 16 Hz components correspond to the 235 – 285 Hz case for the high-frequency regions of analysis. Observe, however, that these components are not purely temporal and have spectral modulation components as well. We show in Figure 27 a representative manifestation of pitch modulation for the 32 Hz, 2 cyc/oct component of the 210 – 310 Hz (high-frequency region). Figure 27b shows the impulse response of a cortical filter centered at this rate-scale pair with a periodicity of ~30 ms in time and 0.5 octaves in frequency, as expected. This component likely corresponds to the spectrotemporal modulation present in the zoomed-in region auditory spectrogram denoted by the dotted arrow in Figure 27a. The cortical response therefore appears similar to the GCT in invoking components *off* the scale axis when pitch is changing.

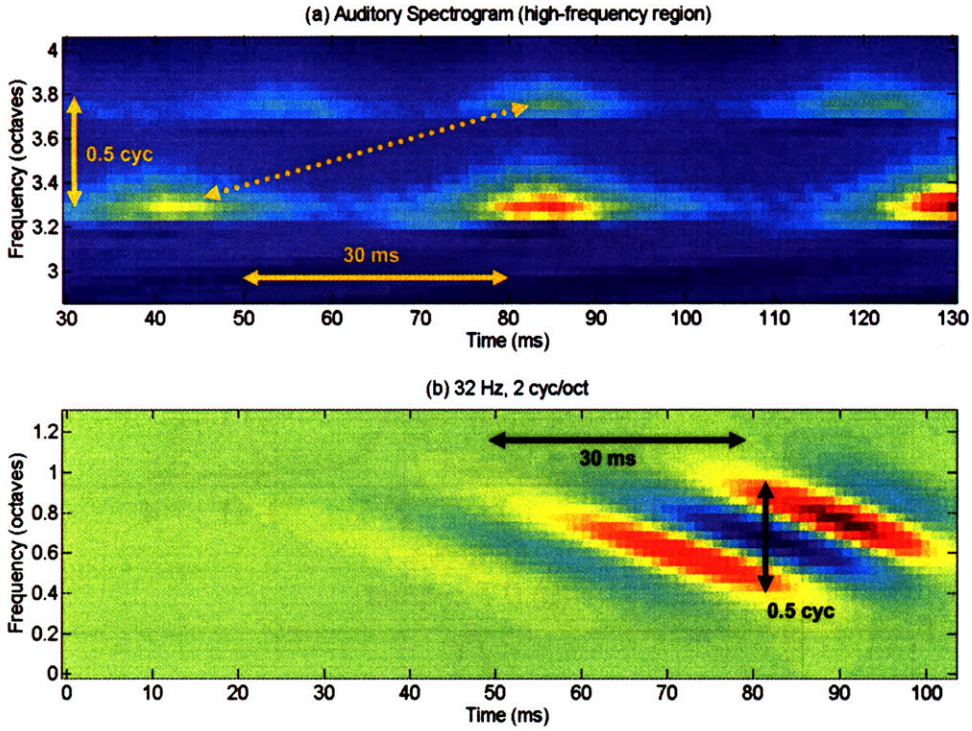

Figure 27 – Interpretation of the 32 Hz, 2 cyc/oct component in the high-frequency region of the auditory spectrogram for /ae/ with pitch change 210 – 310 Hz; (a) Zoomed-in auditory spectrogram in the high-frequency region; solid arrows indicate periodicity in time and frequency of the spectrotemporal modulation component indicated by the dotted arrow; (b) Real impulse response of 32 Hz, 2 cyc/oct cortical filter.

To interpret components along the scale axis of the cortical response for changing pitch, we consider here the 210 – 310 Hz case; overall, similar observations were made for the 235 – 280 Hz case though with one exception to be highlighted in the subsequent discussion. We show in Figure 28a the average all spectral slices from Figure 25f along with the true formant envelope. We refer the reader to Figure 22 for the impulse responses of the 0.4 and 2 cyc/oct components and show in Figure 28b the response for the 0.7 cyc/oct component, all observed in the cortical response along the scale axis (Figure 25g and Figure 26g). In the high-frequency region, the 0.7 cyc/oct component likely corresponds to the coarse formant structure with peak region between 3 – 4 octaves (horizontal solid arrow (near H), Figure 28a); the 2 cyc/oct component corresponds to the peak at H near the true formant peak (dashed arrow (near H), Figure 28a).

In the low-frequency region, we do not observe a high-scale component in the cortical response corresponding to individual harmonic peaks in contrast to the fixed pitch conditions. Presumably, the harmonic peaks have been smoothed across time such that the low-scale component (0.4 cyc/oct) corresponding to a local portion (horizontal solid arrow (near L), Figure 28a) of the formant envelope dominates the cortical response. Nonetheless, low-amplitude maxima are present in the spectrum near L (dotted arrows (near L), Figure 28a) indicative of this smoothing effect. Turning now to the cortical response corresponding to the low-frequency region for the 235 – 280 Hz case, a component at 1.4 cyc/oct (Figure 24g, dashed arrow) and another at 0.4 cyc/oct (Figure 24g, solid arrow) can be observed. The 0.4 cyc/oct component has the same interpretation as in the 210 – 310 Hz condition. The 1.4 cyc/oct component will correspond to a single peak in the auditory spectral slice representing the result of harmonic peaks at different

frequencies smoothed across time. It appears in this case that smoothing was not as significant as in the 210 – 310 Hz condition such that this peak is reflected in the corresponding cortical response.

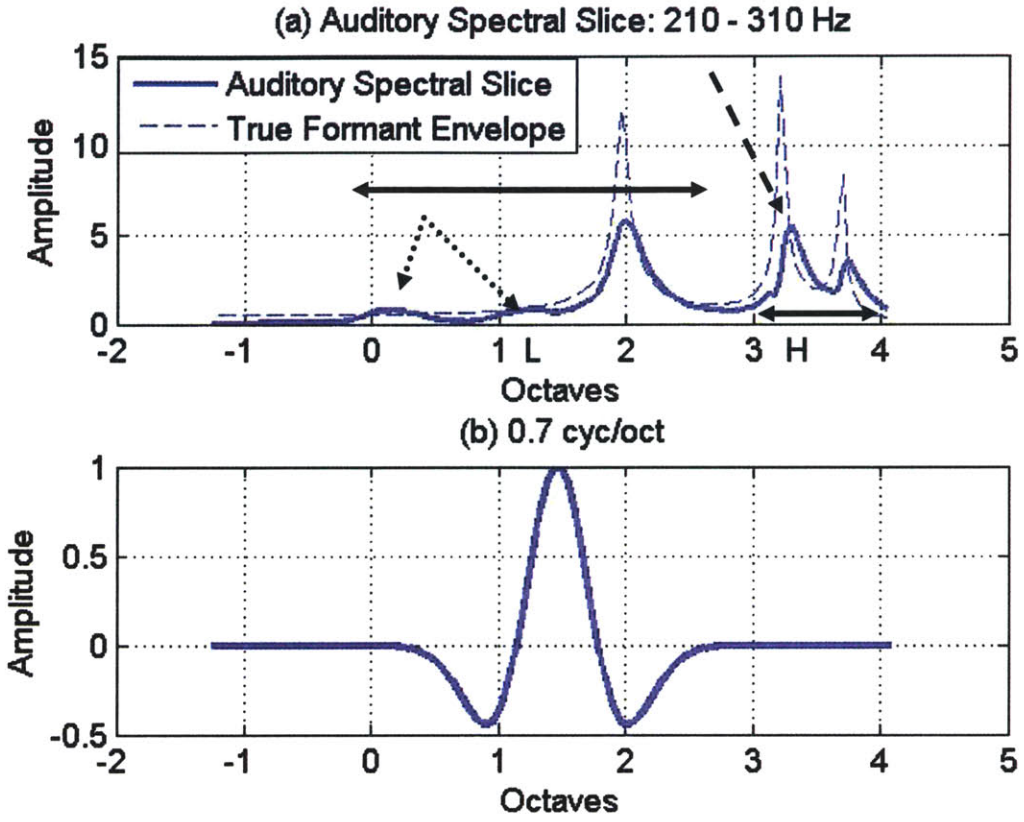

Figure 28 – Interpretation of components along the scale axis for the 210 – 310 Hz condition; (a) averaged auditory spectral slice and true formant envelope; horizontal solid arrow (near H) – peak of 0.4 cyc/oct component; dashed arrow (near H) – peak of 2.4 cyc/oct component; horizontal solid arrow (near L) – 0.7 cyc/oct component; dotted arrows (near L) – low-amplitude peaks indicative of smoothing effect of multiple pitch harmonics; (b) impulse response of 0.7 cyc/oct component.

In summary, for the condition of changing pitch, we have observed that the GCT may provide improved separability even under conditions of high-pitch with harmonic structure being rotated off the scale axis. The auditory model is similar to the GCT in invoking components off the scale axis with rates > 8 Hz for changing pitch. In addition, for the 210 – 310 Hz condition, components on the scale axis appear to primarily resemble formant structure. Nonetheless, we cannot argue source-filter separability as in the GCT; specifically, those components off the scale axis do not necessarily represent a coherent mapping of changing harmonic lines as in the GCT.

3.5 Relation to formant estimation

From Section 3.4, we have obtained evidence suggesting that harmonic projection and the GCT may provides means of addressing the spectral undersampling problem under the condition of changing pitch. The former method improves the spectral sampling of an underlying stationary

formant envelope. The latter method appears to invoke source-filter separability by mapping harmonic line structure (corresponding to pitch) to a coherent component in a 2-D modulation space. Because this type of coherent mapping was not observed for the auditory cortex model, separability in the cortical model could not be as easily argued as in the GCT. In this thesis, we therefore focus on the harmonic projection and GCT methods, deferring use of the cortical model for future work. Herein we discuss methods of spectral estimation via these two methods with the end goal of improving high-pitch formant estimation.

We have observed that the projection of harmonics under conditions of changing pitch is able to increase the sampling of an underlying formant envelope. It is conceivable that with an appropriate interpolation method, an improved estimate of the underlying spectral envelope can be obtained using such a collection of harmonic samples relative to that derived from a single spectral slice. One simple method of interpolation is that of averaging short-time spectra across time. In relation to the GCT, we can think of this averaging as extracting the rate axis of GCTs computed for localized regions of the STFT followed by an inverse transform to generate a spectrum. Denoting N and M as the length in time and frequency of a localized STFT region $s[n,m]$ and k, l, and $N_{\hat{\omega}}$ and $M_{\hat{\Omega}}$ as in Section 3.2, the GCT is then

$$S[k,l]=\frac{1}{NM}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}s[n,m]e^{\frac{-j2\pi k}{N_{\hat{\omega}}}}e^{\frac{-j2\pi l}{M_{\hat{\Omega}}}}. \tag{3.29}$$

Extracting the rate axis only ($k = 0$), we have

$$S[0,l]=\frac{1}{M}\sum_{m=0}^{M-1}\left(\frac{1}{N}\sum_{n=0}^{N-1}s[n,m]\right)e^{\frac{-j2\pi l}{M_{\hat{\Omega}}}} \tag{3.30}$$

which is the equation for the (forward) DFT of $\frac{1}{N}\sum_{n=0}^{N-1}s[n,m]$.

Analogous to the improved spectral sampling of harmonic projection, we have observed that the GCT under conditions of changing pitch appears to separate harmonic line structure and formant envelope components of the STFT. From our observations, we model the harmonic line structure as in (3.1); in addition, we denote the portion of the formant envelope within the localized patch as a relatively *slowly-varying* $a[n,m]$ with 2-D DFT expressed as $A(\hat{\omega},\hat{\Omega})$. Under a multiplicative spectral source-filter model, the windowed localized region $s_w[n,m]$ is

$$\begin{aligned}s_w[n,m]&=w[n,m](1+\cos(\hat{\omega}_0\Phi[n,m]))a[n,m]\\ s_w[n,m]&=w[n,m]a[n,m]+w[n,m]a[n,m]\cos(\hat{\omega}_0\Phi[n,m])\end{aligned} \tag{3.31}$$

In the rate-scale domain $S_w(\hat{\omega},\hat{\Omega})$ becomes

$$\begin{aligned}S_w(\hat{\omega},\hat{\Omega})&=H(\hat{\omega},\hat{\Omega})+H(\hat{\omega}+\hat{\omega}_0\sin\hat{\theta},\hat{\Omega}-\hat{\omega}_0\cos\hat{\theta})+H(\hat{\omega}-\hat{\omega}_0\sin\hat{\theta},\hat{\Omega}+\hat{\omega}_0\sin\hat{\theta})\\ H(\hat{\omega},\hat{\Omega})&=W(\hat{\omega},\hat{\Omega})*_{\hat{\omega},\hat{\Omega}}A(\hat{\omega},\hat{\Omega})\end{aligned} \tag{3.32}$$

where $*_{\hat{\omega},\hat{\Omega}}$ denotes convolution in the rate-scale space (Figure 29).

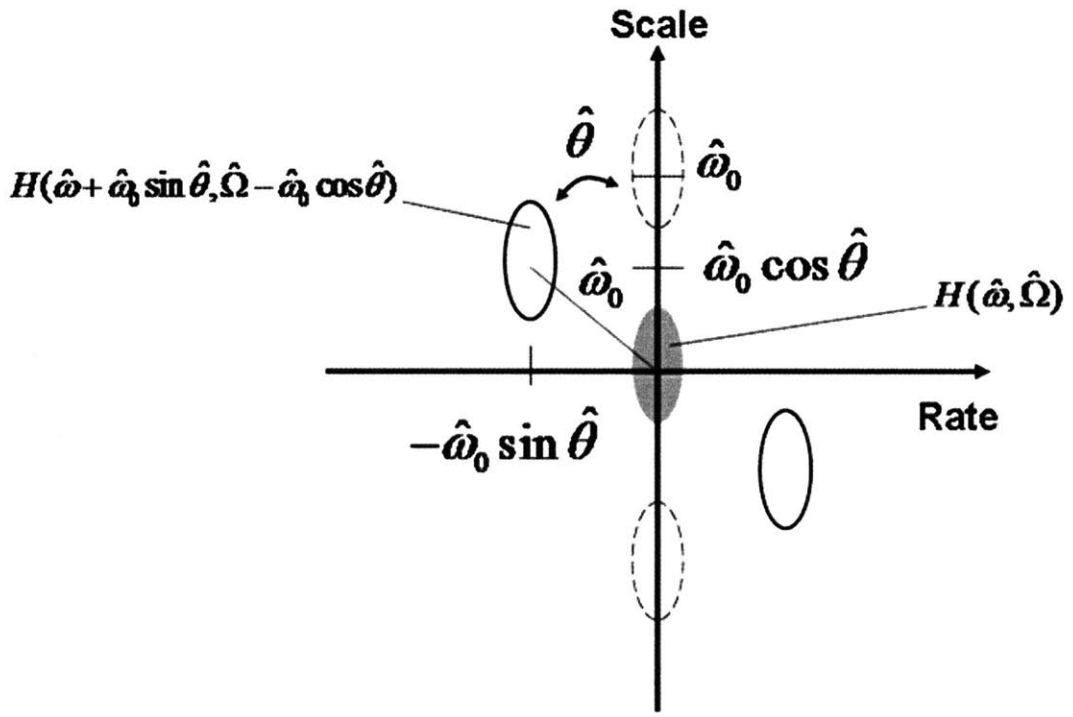

Figure 29 – Rate-scale schematic showing fixed pitch (dotted) versus changing pitch (solid). $H(\hat{\omega},\hat{\Omega})$ is represented by ellipses.

We see that $w[n,m]a[n,m]$ maps to a function concentrated at the origin ($H(\hat{\omega},\hat{\Omega})$) as represented by the shaded ellipse. In contrast, the fast-varying component $w[n,m]a[n,m]\cos(\hat{\omega}_0\Phi[n,m])$ maps to a pair of smeared impulses located $\hat{\omega}_0$ from the GCT origin as measured radially. For changing f_0, the rotational nature in transforming skewed harmonic lines maps them off the scale axis at a non-zero angle $\hat{\theta}$ (represented by solid ellipses), thereby invoking previously observed separability of the source from vocal-tract spectra. For fixed f_0, the smeared impulses lie along the scale axis (dotted). Note that for fixed f_0, $\hat{\theta} = 0$, and source-filter separability can be achieved only if $\hat{\omega}_0$ is greater than the scale-frequency bandwidth of $H(\hat{\omega},\hat{\Omega})$. This model motivates a simple method of spectral estimation; specifically, 2-D processing across localized regions of a STFT via low-pass filtering could isolate $w[n,m]a[n,m]$ from $w[n,m]a[n,m]\cos(\hat{\omega}_0\Phi[n,m])$. A reconstructed STFT could then be used to generate a spectral estimate.

In relation to formant estimation, harmonic projection and the GCT can afford improved spectral representations (i.e., estimates) of formant envelopes. Any magnitude spectrum $|X(\hat{\omega})|$ derived from these methods may then be used to estimate formant frequencies via the autocorrelation-method of linear prediction since the inverse DFT of $|X(\hat{\omega})|^2$ is an estimate of the autocorrelation function $r_x[n]$.

3.6 Conclusions

In this chapter, we have characterized several realizations of a 2-D speech processing framework through simulations. The harmonic projection method appears to improve spectral sampling of a stationary formant envelope under the condition of changing pitch. Similarly, the GCT appears to afford improved source-filter separability in a 2-D modulation space when pitch is changing; this is done through a coherent mapping of harmonic structure. For these two realizations, we have provided analytical justifications for our observations and have proposed their use in deriving spectra as a basis for improving high-pitch formant estimation. Though the auditory cortex model also highlights distinctive aspects of speech, the lack of a coherent mapping prevented us from arguing for source-filter separability as was done for the GCT. Further work is necessary to assess the potential benefits of this particular realization.

Chapter 4

Formant Estimation Methods

In this chapter, we discuss the methodology of using the proposed 2-D framework for formant frequency estimation. Spectral estimation methods motivated from our observations and analysis presented in Chapter 3 are applied to a series of synthesized vowels with degrees levels of pitch shifts. All spectra are then used in linear prediction to estimate formant frequencies. These results are compared to the results of traditional and homomorphic linear prediction baselines.

This chapter is organized as follows. Section 4.1 describes our vowel synthesis methodology. In Section 4.2, we discuss a variety spectral estimation methods aiming to exploit temporal change of pitch and their application to formant estimation using linear prediction. We conclude in Section 4.3 by discussing our baseline methods of traditional and homomorphic linear prediction and summarizing the array of methods used.

4.1 Synthesis

Pure impulse-train source signals with starting f_0 (f_{0s}) ranging from 80-200 Hz (males), 150-350 Hz (females), and 200-450 (children) Hz were synthesized with linear pitch increases (df_0) ranging from 10 to 50 Hz. f_{0s} and df_0 varied in 5-Hz steps. Specifically, starting and ending pitch values were used in linear interpolation across the duration of each synthesized utterance. To generate the impulse train with time-varying pitch, impulses were spaced according to points of this interpolation (denoted as $f_0[n]$) across the desired duration. Initially, we synthesized the source at 8 kHz; however, we observed that the temporal resolution at this sampling rate invoked a piece-wise linear pitch change. Consequently, sharp transitions were observed in the spectrogram at pitch transition points (Figure 30, top). To generate a smoother pitch track, we used instead a 16-kHz sampling rate for generating the impulse train and downsampled the result to 8 kHz using a 100^{th} order least-squares, linear-phase, finite-impulse response filter. A representative spectrogram of an input source generated using this method is shown in Figure 30 (bottom); in this case, energy transitions between pitch transitions are notably reduced in magnitude.

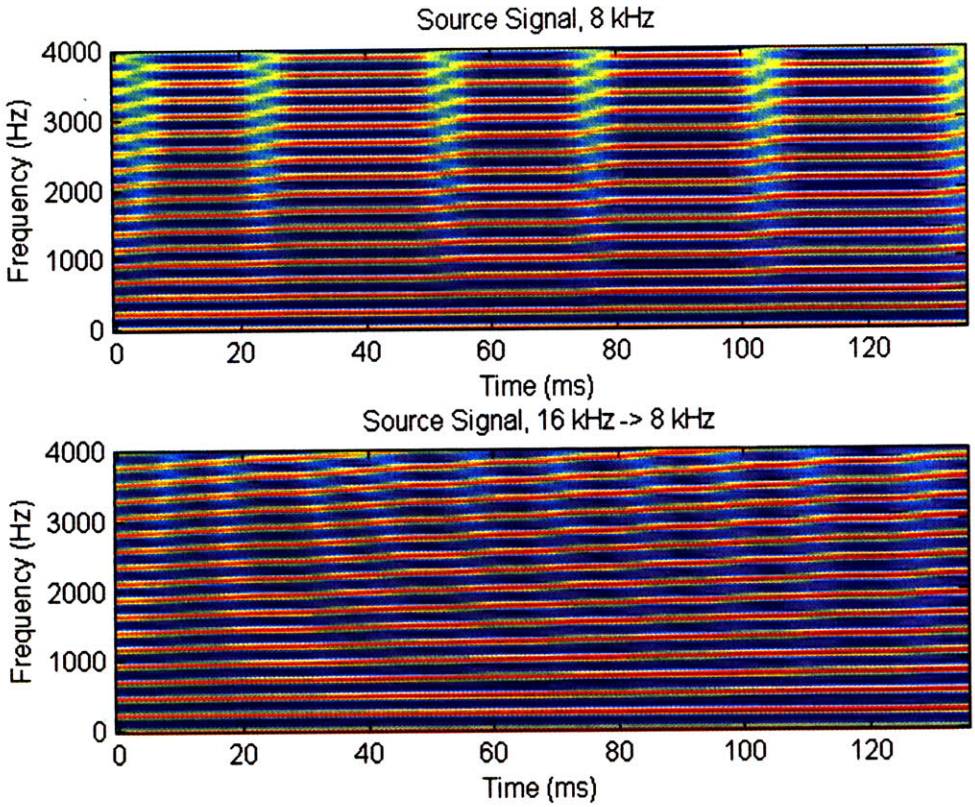

Figure 30 – Comparison of source generated at 8 kHz (top) versus source generated at 16 kHz and downsampled to 8 kHz (bottom) for 235 Hz – 285 Hz. The large energy transitions in the signal generated at 8 kHz are reduced in the signal generated at a higher sampling rate and downsampled. Pitch changes from 235 – 285 Hz across 135-ms duration.

Table 3 – Table of formant frequency values (in Hz) used in synthesis.

	/ah/ (v=1)	/iy/ (v=2)	/ey/ (v=3)	/ae/ (v=4)	/oh/ (v=5)	/oo/ (v =6)	/uh/ (v=5)
F1 (males)	730	270	460	660	450	300	-
F2 (males)	1090	2290	1890	1720	1050	870	-
F3 (males)	2440	3010	2670	2410	2610	2240	-
F1 (females)	850	310	560	860	600	370	-
F2 (females)	1220	2790	2320	2050	1200	950	-
F3 (females)	2810	3310	2950	2850	2540	2670	-
F1 children)	680	370	690	1010	-	430	850
F2 children)	1060	3200	2610	2320	-	1170	1590
F3 children)	3180	3730	3570	3320	-	3260	3360

Source signals were filtered with 6th order all-pole models corresponding to the vowels */ah/* (v=1), */iy/* (2), */ey/* (3), */ae/* (4), */oh/* (5), and */oo/* (6); for children, */oh/* was replaced by */uh/* (5) because we were unable to find formant data in the literature for */oh/*. Formant bandwidths (B1, B2, B3) were set to 54, 65, and 70 Hz, respectively and correspond to average measurements in vowels for males and females reported in Stevens [11]. Formant frequencies (F1, F2, F3) were set to average measurements reported by Stevens [11] and Peterson and Barney [20] for males, females, and children (Table 3). Synthesized utterance durations were vowel-specific and set to

average durations (across adult females and males) measured from the Switchboard corpus by Greenberg and Hitchcock (Table 4) [21]; we were unable to find data in the literature regarding child vowel durations.

Table 4 – Average vowel durations across adult females and males (in ms) from [21].

	ah (v=1)	iy (v=2)	ey (v=3)	ae (v=4)	oh (v=5)	oo (v =6)	uh (v=5)
Duration	95	100	125	135	135	105	55

All-pole filter coefficients were generated via a cascade of formant resonators proposed by Klatt [22] and implemented in MATLAB by Mehta [23]. Specifically, the vocal-tract transfer function $H(\omega)$ is defined as [23]

$$
\begin{aligned}
H(\omega) &= \prod_{i=1}^{3} \frac{A_i}{1 - B_i e^{-j\omega} - C_i e^{-j2\omega}} \\
A_i &= 1 - B_i - C_i \\
B_i &= 2e^{-\pi \frac{B_i}{fs}} \cos(2\pi \frac{F_i}{fs}) \\
C_i &= -e^{-2\pi \frac{B_i}{fs}}
\end{aligned}
\tag{4.1}
$$

where i denotes the ith formant (F) and bandwidth (B) and fs is sampling frequency (here, $fs = 8$ kHz). Figure 31 shows a representative example of a synthesized vowel */ae/* with pitch 235 – 285 Hz based on the above method.

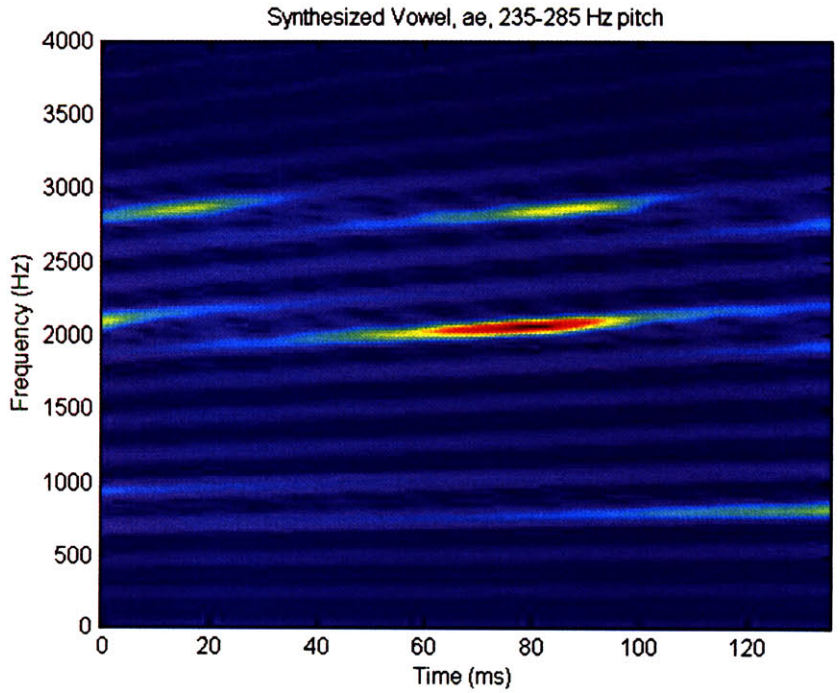

Figure 31 – Synthesized vowel */ae/* with pitch 235 – 285 Hz.

4.2 Formant Estimation Exploiting Pitch Dynamics

In this section, we discuss several spectral estimation methods aiming to exploit temporal change of pitch as a basis for formant estimation. For formant estimation, the magnitude spectrum resulting from each of these methods is used to obtain an autocorrelation estimate for use in linear prediction. Specifically, for a DFT length of N, a one-sided magnitude spectrum $|X[k]|$ of length $\frac{N}{2}+1$ is appended by a frequency-reversed version, thereby resulting in a two-sided zero-phase spectrum $X_2[k]$. Denoting the inverse DFT computation of $X_2^2[k]$ as $x[n]$, this sequence is circularly shifted by half of the DFT length to generate the autocorrelation estimate $r_x[n]$:

$$\begin{aligned} X_2[k] = \quad & |X[k]| \quad k = 0,1\ldots\frac{N}{2} \\ & |X[N-k]| \quad k = \frac{N}{2}+1, \frac{N}{2}+1,\ldots N \\ x[n] = & \sum_{k=0}^{N-1} X_2^2[k] e^{j\frac{2\pi nk}{N}} \\ r_x[n] = & \, x[((n-\frac{N}{2}))_N] \end{aligned} \quad . \tag{4.2}$$

We observed that $N = 2048$ was sufficient to avoid aliasing of $r_x[n]$ due to the inverse DFT. $r_x[n]$ is used to generate the normal equations with the order set to 6 (corresponding to the three synthesized formants) (7.2A.1) which are then solved using the Levinson-Durbin recursion [1]. Finally, the roots of the resulting coefficients are solved to obtain the formant frequency estimates. Herein we describe several methods motivated from our discussion in Section Chapter 3 for obtaining $|X[k]|$.

4.2.1 Harmonic Projection and Interpolation

To implement the method of harmonic projection suggested by Figure 4, the f_0 pitch contour of the synthesizer was used to collect spectral samples across the full duration of each vowel. Specifically, the following steps were taken to generate a spectral estimate:

1. A STFT was computed with a 1-ms frame interval and a pitch-adaptive Blackman window , $w[n]$ (4.3), with duration corresponding to four times the true f_0 for each frame (denoted as *STFT₁*).

$$w[n] = \begin{cases} 0.42 - 0.5\cos(\frac{2\pi n}{M}) + 0.08\cos(\frac{4\pi n}{M}), & 0 \le n < M = \frac{4}{f_0} \\ 0 & otherwise \end{cases} \tag{4.3}$$

2. Each spectral slice of $STFT_l$ was normalized by $W(0) = \sum_{n=0}^{M-1} w[n]$.

3. Using the true f_0 contour, peak-picking of the harmonics of f_0 was done across all normalized spectral slices via the SEEVOC algorithm [24].

4. All harmonic peaks were collapsed into a single function linearly interpolated across frequency.

The pitch estimate for each frame was determined from decimating the interpolated pitch contour used in the synthesis by a factor of 2 to account for the downsampling from 16 kHz to 8 kHz (i.e., $f_0[2n]$). In selecting a strategy for obtaining harmonic peaks for interpolation, we initially performed short-time analysis using a fixed 20-ms Hamming window and 1-ms frame interval. Figure 32 (top) illustrates short-time spectra for the female vowel */ae/* across the minimum, average, and maximum pitch values used in synthesis. Observe that while the harmonic peaks are present for the 290- and 500-Hz condition, the 80-Hz spectrum lacks harmonic structure. To understand this effect, recall from Equation (2.6) that the short-time spectrum of a stationary vowel can be viewed as scaled versions of the window with spacing $\omega_k = \frac{2\pi}{P}$ with $P = \frac{1}{f_0}$. As f_0 decreases, the replicas of $W(\omega)$ become more closely spaced such that there is significant interaction between neighboring windows' main- and side-lobes. To address this, we used instead short-time analysis with window size dependent on the local pitch value (i.e., Step 1); specifically, larger pitch periods were analyzed with longer windows. In addition, to reduce side-lobe interactions, we chose the Blackman window with a peak side-lobe amplitude of -57 dB (relative to the main-lobe) [4]. Figure 32 (bottom) shows spectra derived in this manner; we observed empirically that a window length of four times the pitch period provided reasonable harmonic peaks across all pitch conditions.

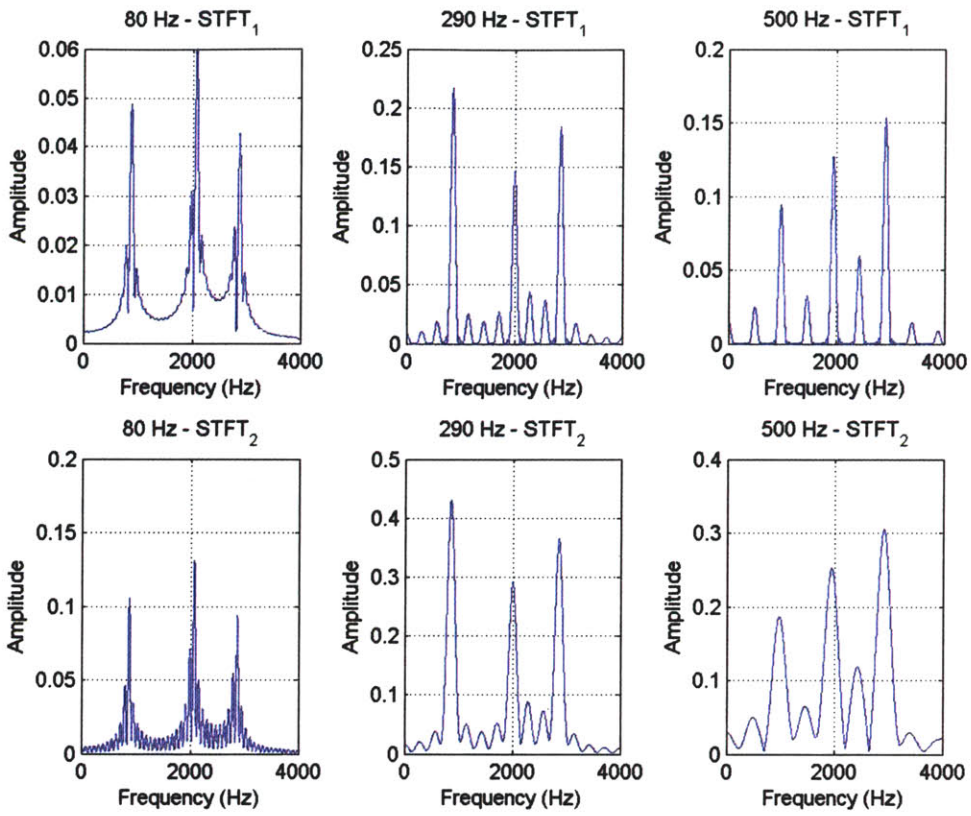

Figure 32 – Comparison of short-time spectra using a fixed 20-ms Hamming window (top) or a pitch-adaptive Blackman window (bottom) for the vowel */ae/* across representative pitch values used in synthesis.

Due to the pitch-adaptive window, $W(\omega = 0)$ is different across frames and is dependent on the window length. The harmonic peaks corresponding to portions of the formant envelope are therefore scaled differently across frames from Equation (2.6). Step 2 is done invoke the same *absolute* magnitudes across spectral slices, independent of window length. In Figure 33a, we show the pitch-adaptive spectrogram while Figure 33b illustrates peak-picking using the SEEVOC algorithm. Figure 33c shows the collection of harmonic samples obtained across the entire vowel and the linearly interpolated spectral estimate (denoted as m = 3).

To remove confounds of the interpolation method itself, interpolation was also performed on harmonic spectral samples from a single spectral slice of $STFT_1$ extracted from the middle of the utterance (m = 4). The resulting single-slice interpolation is shown in Figure 33b. Finally, as previously noted, a simpler method of interpolation is spectral slice averaging which we also implemented and denote as m = 5. Figure 33d shows the result of averaging all magnitude spectral slices in $STFT_1$ (solid) along with two sample spectral slices (dashed) used in computing this average.

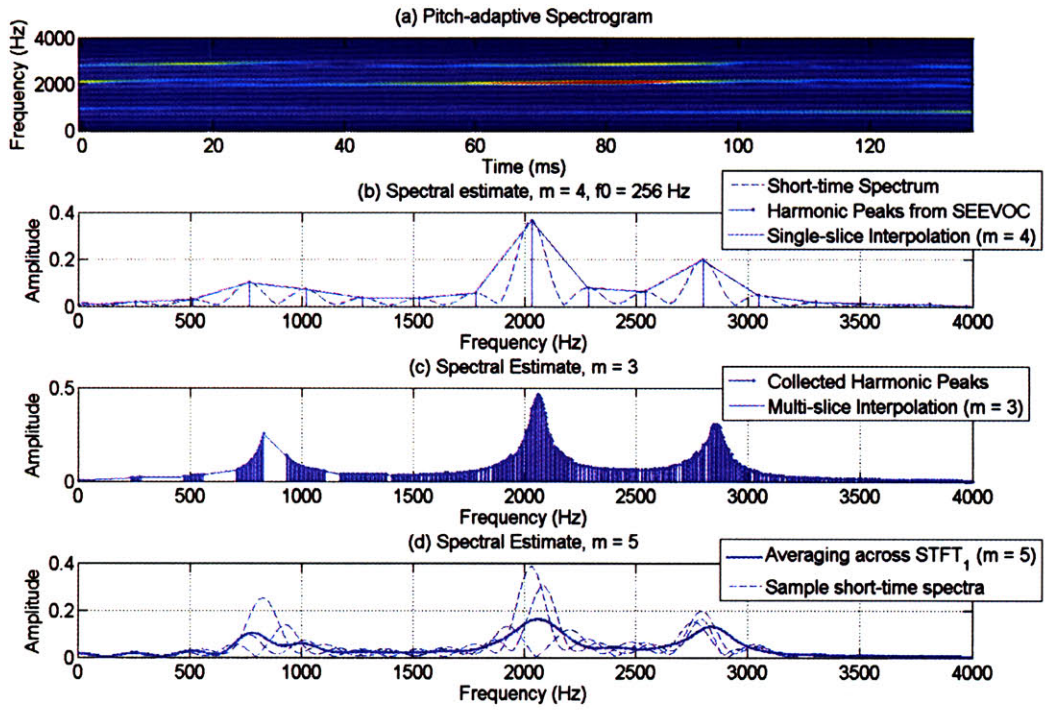

Figure 33 - Harmonic projection method; (a) Pitch-adaptive short-time spectrogram; (b) Single-slice peak-picking using SEEVOC algorithm at 256 Hz and single-slice interpolation (m = 4); (c) Projection of collected harmonics and linearly interpolated spectral envelope (m = 3); (d) Representative short-time spectra (dashed) used in the computing the resulting average spectrum (solid) (m = 5).

For comparison purposes, we show in Figure 34a the spectral estimates of m = 3, 4, and 5 along with the true formant envelope (denoted as $H(\omega)$). Linear interpolation of peaks from multiple slices (m = 3) appears to trace out $H(\omega)$ better than that of single-slice interpolation (m = 4). Specifically, the local maxima of the spectral estimate for m = 3 are closer to the true formant peaks of $H(\omega)$ than those of m = 4, thereby illustrating the benefits of obtaining additional spectral samples of the formant envelope by exploiting temporal change of pitch. Consequently, the formant peaks of the resulting linear prediction envelope for m = 3 appear closer to the true formant peaks than that of m = 4 (Figure 34b). Observe also for m = 3 a more jagged interpolation (e.g., near F1 = 860 Hz) in low-frequency regions than high-frequency regions. This is due to the relatively sparser sampling of harmonic peaks in low-frequency in comparison to the broader harmonic sampling in high-frequency regions (Section 3.1). In addition, the interpolation can appear "ragged" in certain regions (e.g., near 2500 Hz), presumably due to remaining main- and side-lobe interactions from short-time analysis affecting the heights of harmonic peaks. In comparison with m = 3, m = 5 generates a smoother spectral estimate and also appears to trace out $H(\omega)$. However, the spectral estimate of m = 5 contains harmonic-like structure and has lower peaks near the formants. This effect can be explained by observing the significant overlap of the main lobes corresponding to harmonic peaks in Figure 33b (e.g., near F2 = 2050 Hz) for two different spectral slices that correspond to distinct pitch values. While the pitch-adaptive short-time analysis aims to minimize interaction between harmonics for *single* spectral slices, it cannot account for the interactions of main- and side-lobes *across time*. Averaging across short-time spectra consequently invokes a smoothing effect on these

interactions. Nonetheless, the peaks of the resulting linear prediction envelope for m = 5 appear more closely matched in frequency to the true formant peaks overall than that resulting from m = 4, thereby providing additional evidence that exploiting temporal change of pitch can benefit formant estimation.

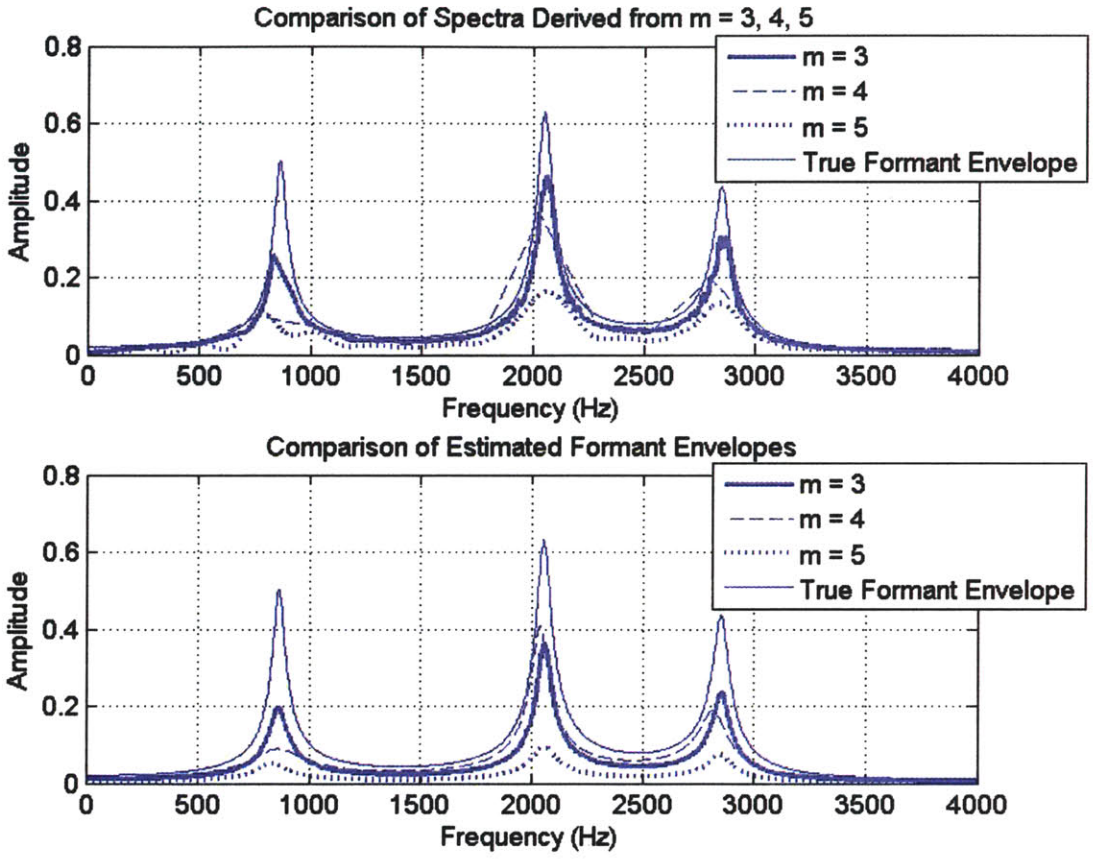

Figure 34 – Comparison of spectra (top) with resulting all-pole formant envelope estimates (bottom) for m = 3, 4, 5.

4.2.2 Grating Compression Transform Filtering

For m = 6, we implemented filtering of the GCT based on the model presented in Section 3.5. Localized regions were extracted from $STFT_I$ (denoted by $s[n,m]$) with a separable 2-D window $w[n,m]$. Specifically, a rectangular window in time with length corresponding to the full duration of the utterance was multiplied by a Hamming window in frequency with width set to 700 Hz; a 350-Hz frequency overlap was invoked.

Two GCTs were computed for each localized region. For GCT_I, a 2-D gradient operator [25] was applied to the entire STFT followed by windowing and removal of the DC component [17]; a 2-D DFT was then applied to generate the GCT. Specifically for a region centered at $n = n_0$, $m = m_0$:

$$D = \begin{bmatrix} 0 & 2 & 2 \\ -2 & 0 & 2 \\ -2 & -2 & 0 \end{bmatrix}$$

$$p_{win}[n,m] = w[n-n_0, m-m_0](s[n,m] *_{n,m} D). \tag{4.4}$$

$$p_1[n,m] = p_{win}[n,m] - \bar{p}_{win}$$

$$GCT_1(k,l) = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} p_1[n,m] e^{-j\frac{2\pi kn}{N}} e^{-j\frac{2\pi ln}{M}}$$

where $*_{n,m}$ denotes convolution across n and m and $\bar{p}_{win}$ denotes the average (DC) value of $p_{win}[n,m]$. The magnitude of GCT_1 was used to determine local estimates of $\hat{\omega}_0$ by peak-picking. For GCT_2, the DFT was computed directly from the windowed region:

$$p_2[n,m] = w[n-n_0, m-m_0]s[n,m]$$

$$GCT_2(k,l) = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} p_2[n,m] e^{-j\frac{2\pi kn}{N}} e^{-j\frac{2\pi ln}{M}}. \tag{4.5}$$

Figure 35 compares both GCT magnitudes and their respective patches in the time-frequency space. *For display purposes, the DC component of GCT_2 was removed.* The local maxima in $|GCT_2|$ (d) corresponding to harmonic line structure (dotted arrows) are accentuated in $|GCT_1|$ (e) due to the gradient operator. These peaks are located at $\hat{\omega} = \pm 0.025\pi$, $\hat{\Omega} = \mp 0.035\pi$ such that $\hat{\omega}_0 \approx 0.043$. The local maxima along the scale axis corresponds to a local portion of the formant envelope (solid arrows).

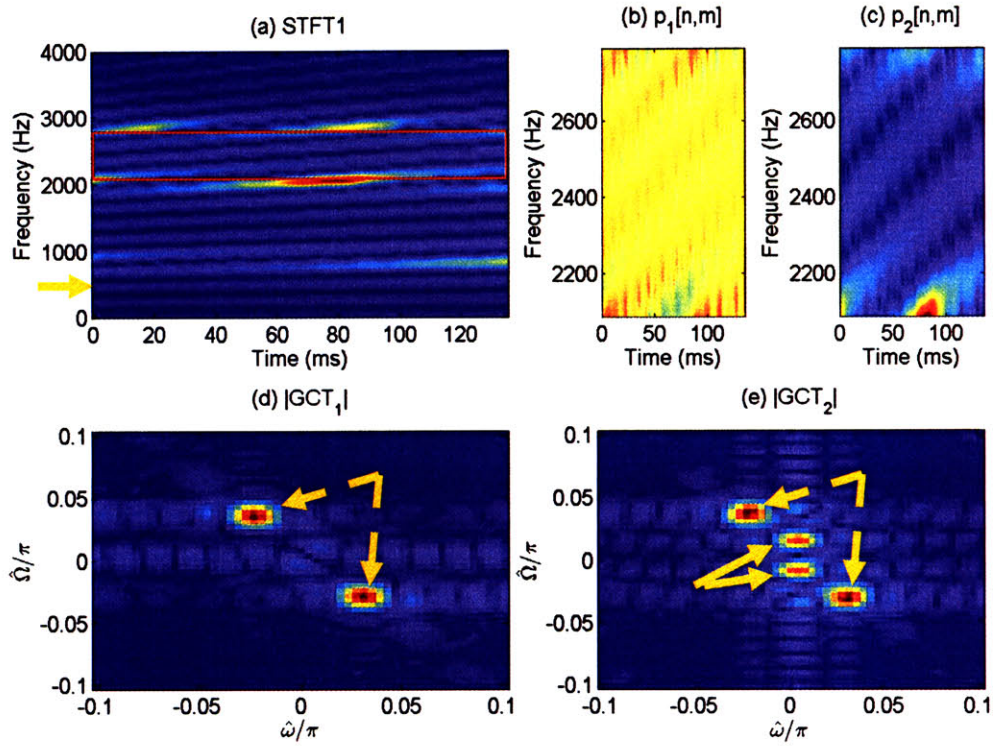

Figure 35 – (a) $STFT_1$ and localized region (rectangle); arrow – 350 Hz; (b) Localized region used for $\hat{\omega}_0$ estimate ($p_1[n,m]$) (c) Localized region used in filtering ($p_2[n,m]$) (d) $|GCT_1|$ computed from; dashed arrows – coherent mapping of harmonic structure (b); (e) $|GCT_2|$ computed from (c) with DC component removed for display purposes; solid arrows – mapping of local formant structure; dashed arrows – coherent mapping of harmonic structure.

GCT_2 was filtered with a 2-D elliptic filter designed by taking the product of two linear-phase low-pass filters in frequency in time. In time, the pass and stop bands were fixed to $0.25\omega_l$ and $0.5\omega_l$, respectively. ω_l corresponds to the $\hat{\omega}_0$ estimate derived from the lowest frequency-region of $STFT_1$ with center frequency of 350 Hz (arrow, Figure 35a). In frequency, we used a pass and stop band of $0.5\hat{\omega}_0$ and $\hat{\omega}_0$, with $\hat{\omega}_0$ corresponding to the local estimate for each region. The filter cutoffs were motivated from empirical observations showing that $\hat{\omega}_0$ tended to increase with frequency region (Figure 36). This effect is caused by the increased fanning of the harmonic line structure towards high-frequency regions; because the harmonic lines are no longer strictly parallel, $\hat{\omega}_0$ will be overestimated. Using the described filter cut-offs, we could therefore expect to obtain an adaptive low-pass elliptical filter that becomes more permissive along the scale direction for regions with increasing frequency. This is consistent with the improved source-filter separation in high-frequency regions previously discussed in Section 3.5.

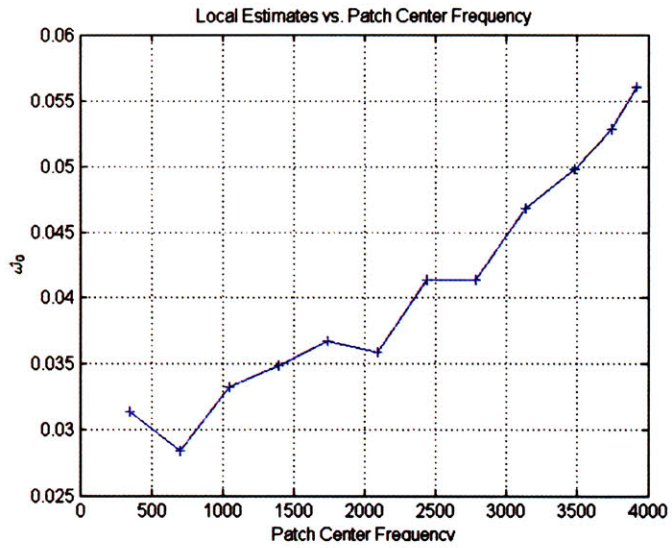

Figure 36 - $\hat{\omega}_0$ estimates as a function of the center frequency of each patch.

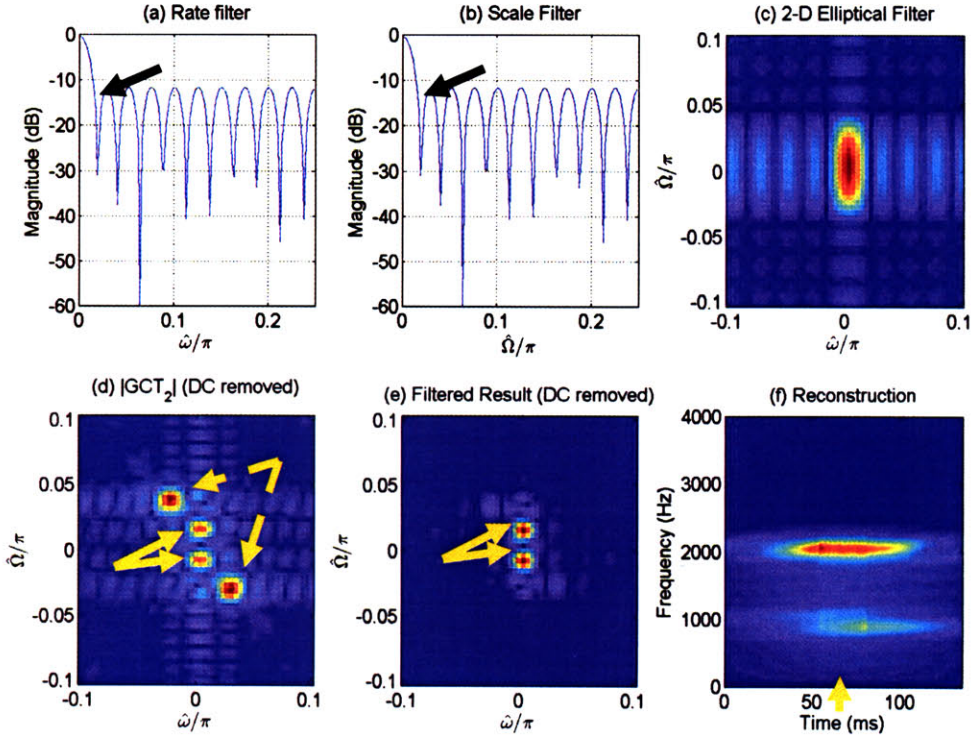

Figure 37 – GCT filtering process. (a) Magnitude of rate filter along $\hat{\omega}$; (b) Magnitude of scale filter along $\hat{\Omega}$; arrows – filter cutoffs; (c) Magnitude of product of (a) and (b); (d) GCT_2 magnitude (DC removed); dashed arrows – coherent mapping of harmonic line structure; solid arrows – mapping of local formant structure; (e) GCT_2 post-filtering, magnitude (DC removed); solid arrows - mapping of local formant structure remaining after filtering; (f) Time-frequency reconstruction via overlap-add; arrow - location in time of spectral slice extracted.

Figure 37 (a-c) shows the construction of the 2-D filter used on the localized region of Figure 35 with low-pass cut-off frequencies indicated by solid arrows. In Figure 37d we show GCT_2 with the harmonic (dotted arrows) and formant envelope (solid arrows) components indicated as before while Figure 37e shows the result of filtering GCT_2 with the elliptical filter of Figure 37c. As in Figure 35, the DC components have been removed for display purposes in these figures. Although we show only the magnitudes of the GCTs, filtering was done on the *complex* GCT_2. Observe that the harmonic components are removed from the GCT in Figure 37e as a result of the 2-D filtering, thereby leaving only the components corresponding to the local portion of the formant envelope. Finally, to generate the spectral estimate, a reconstructed time-frequency distribution was computed using overlap-add[8], and a spectral slice was extracted corresponding in time to the middle of the utterance (solid arrow). In Figure 38 (top) we show the spectral estimate generated from filtering in the GCT (m = 6) in comparison with the true formant envelope. Similar to the method of spectral slice averaging, the estimate appears smoother than that of m = 3; in addition, the estimate exhibits lower local maxima amplitudes and some harmonic-like structure indicative of the main- and side-lobe interactions of multiple spectral slices across time. The resulting formant envelope estimate from linear prediction is also shown in Figure 38 (bottom).

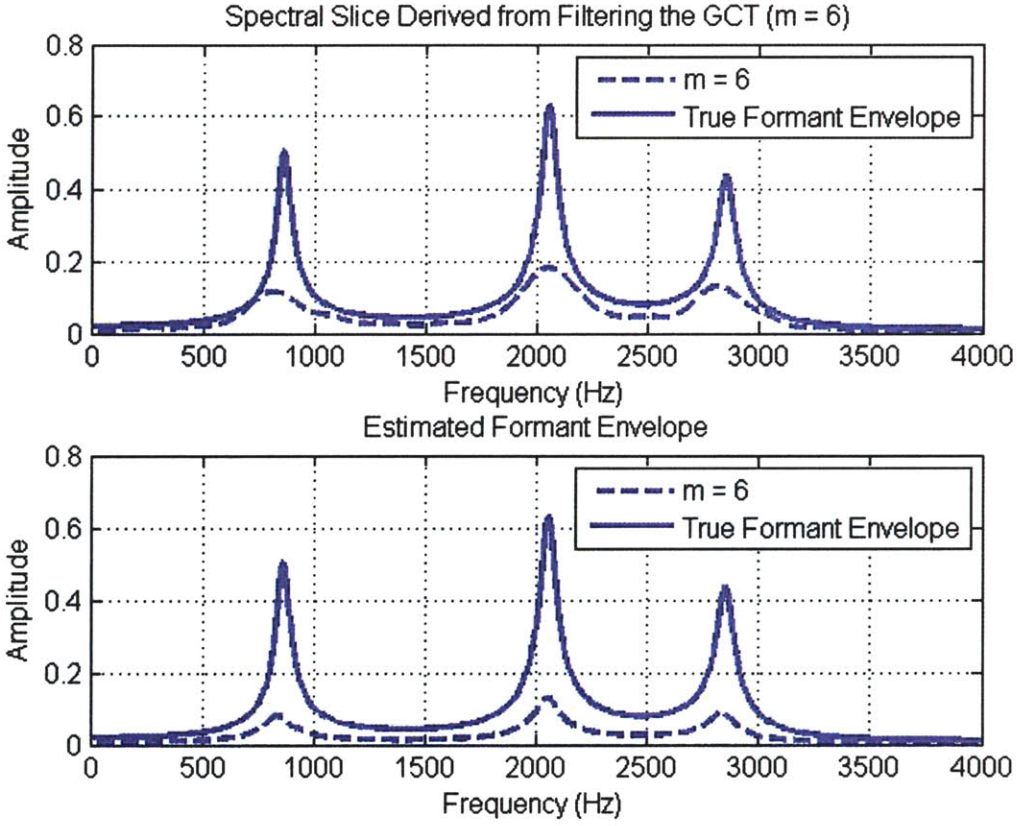

Figure 38 – Spectral slice from m = 6 in comparison with true formant envelope (top) with resulting all-pole formant envelope estimate (bottom).

[8] The Hamming window in the frequency direction was modified from its original form to sum to unity when overlapped by half of its length.

4.3 Baseline Formant Estimation

Two baseline formant estimation methods were implemented for comparison with those aiming to exploit temporal change of pitch. Specifically, a magnitude STFT (denoted as $STFT_0$) was computed for each utterance using a 20-ms Hamming window, 1-ms frame interval, and 2048-point DFT. A single spectral slice located in the middle of the utterance (denoted as $\left|X_{STFT_1}[k]\right|$) was extracted for use with linear prediction as was done for the spectra derived via m = 3-6; we refer to this as the traditional linear prediction baseline (m = 1). As previously discussed in Section 2.4, Rahman and Shimamura have suggested the use of homomorphic linear prediction to address high-pitch formant estimation [15]. To implement this method (m = 2), we used the same spectral slice of m = 1 to compute the real-cepstrum:

$$c_{STFT_1}[n] = \frac{1}{N}\sum_{k=0}^{N-1} \log\left|X_{STFT_1}[k]\right| e^{j\frac{2\pi kn}{N}} \tag{4.6}$$

where $N = 2048$. An ideal lifter with cut-off $\frac{0.6}{f_0}$ for $f_0 < 250$ Hz and $\frac{0.7}{f_0}$ for $f_0 > 250$ Hz was applied to the real cepstrum and an inverse DFT was computed to generate a magnitude spectrum to be used with linear prediction [15].

In Figure 39a, we show the spectral slice $\left|X_{STFT_1}[k]\right|$ of m = 1; Figure 39b shows a portion of the real cepstrum computed from $\left|X_{STFT_1}[k]\right|$ with the ideal lifter cutoff (arrow). We observe in the real cepstrum local peaks spaced at ~3.8 ms, consistent with a pitch value of ~260 Hz. The lifter is designed to isolate low-quefrency components corresponding to formant structure from these periodic components. In the resulting magnitude spectrum (Figure 39c), the harmonic peaks are no longer present, as can be expected from the spectral smoothing accomplished via liftering. This smoothing effect is distinct from that of spectral slice averaging and filtering in the GCT (m = 5 and 6). Whereas in m = 5 and 6 smoothing is performed across spectral slices in time, liftering operates on a single spectral slice. Figure 40 (top) shows the spectra used in the baseline methods along with the resulting formant envelope estimates (bottom) in comparison with the true formant envelope.

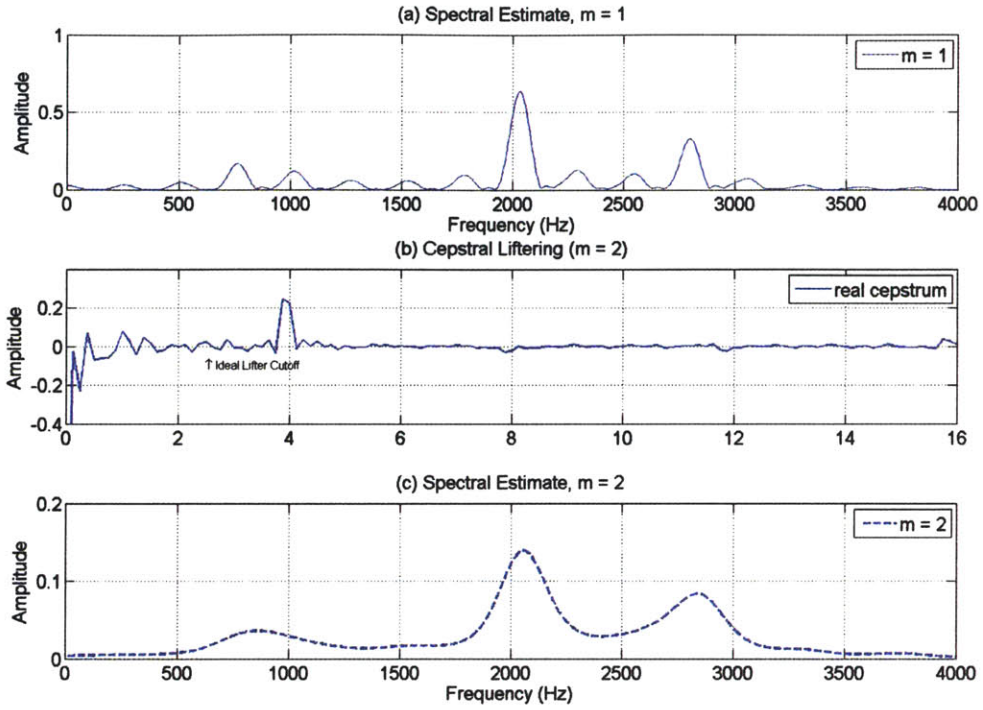

Figure 39 – (a) Spectral slice from $STFT_l$; (b) Cepstral analysis; arrow denotes ideal lifter cutoff; (c) Resulting spectrum from cepstral analysis. Analyses are performed on the synthesized vowel shown in Figure 31.

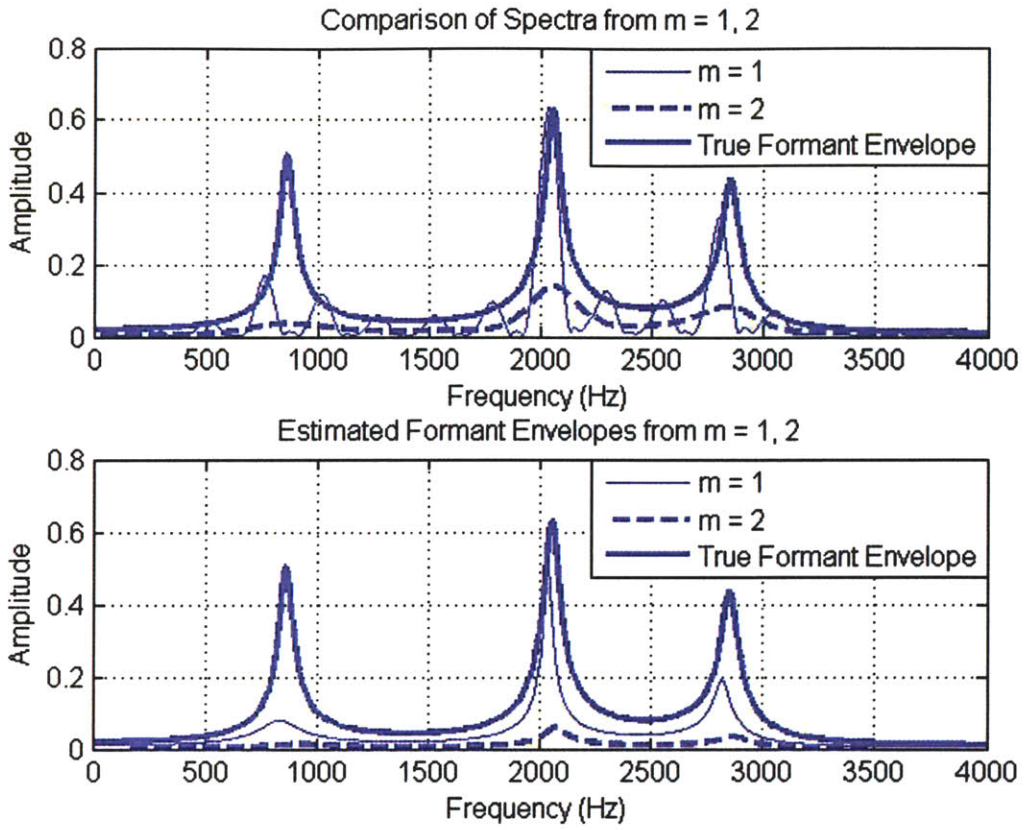

Figure 40 – Comparison of baseline spectral estimation methods (top) with resulting all-pole formant envelope estimates (bottom) for m = 1 (traditional linear prediction) and m = 2 (homomorphic linear prediction).

4.4 Conclusions

In this chapter, we have described our experimental framework for assessing the value of exploiting temporal change of pitch in formant estimation. We have discussed our vowel synthesis methodology and presented several methods of formant estimation aimed at addressing the spectral undersampling problem by exploiting pitch dynamics. Specifically, a number of techniques are proposed for generate a short-time spectrum for use with linear prediction in estimating formant locations. The harmonic projection method generates a spectrum by interpolating across a collection of harmonic peaks extracted from a pitch-adaptive spectrogram (m = 3); to control for the interpolation method itself, the harmonic peaks of a single spectral slice are also interpolated to generated a spectrum (m = 4). In addition, a simpler realization of the harmonic projection method is that of spectral slice averaging (m = 5); recall that this method has a simple interpretation in the context of the GCT. Finally, filtering in the GCT is done (m = 6) to exploit source-filter separability in a 2-D modulation space as a basis for generating a spectral slice of a vocal tract frequency response. We have chosen to compare the performance of these techniques with traditional and homomorphic linear prediction.

Chapter 5

Formant Estimation Results

In this chapter, we present the results of formant estimation using the previously discussed methods. Using the parameters in synthesis (e.g., vowel type v, starting pitch f_{0s}, etc.), several criteria of "goodness" can be extracted from the results. The primary metric from which other criteria are derived is the percent error between an estimated ($\hat{F}$) versus true formant (F) frequency value:

$$\%error = 100\frac{\left|\hat{F} - F\right|}{F} \tag{5.1}$$

Section 0 presents percent errors for distinct formants with fixed vowel type, pitch start, and pitch shift. In Section 5.2, we compute the average of percent errors across vowel type and pitch shifts to illustrate the effects of starting pitch on formant estimation accuracy. Similarly, Section 5.3 presents averages across vowels and pitch starts to illustrate the effect of the amount of pitch shift invoked. Finally, Section 5.4 presents averages across pitch starts and pitch shifts, thereby illustrating the relative performance across vowels. In the aforementioned sections, we display representative results for *females*, deferring discussion of performance across genders to Section 5.4, where we present a global goodness criterion based on averaging across all synthesis parameters. We conclude by summarizing our findings in Section 5.5. As a reference, we show in Table 3 a summary of the methods described in Chapter 4 for the formant estimation task to compared in this chapter.

Table 5 – Summary of methods used in formant estimation.

m=1	Traditional linear prediction (LP)
m=2	Homomorphic linear prediction (HLP)
m=3	Interpolation of collected harmonic peaks + LP
m=4	Interpolation using single slice + LP
m=5	Time-average of $STFT_l$ + LP
m=6	GCT-based filtering on $STFT_l$ + LP

5.1 Raw Percent Error

For the i^{th} formant of the v^{th} vowel with df_0 pitch shift, the raw percent error can be viewed as a function of the starting pitch f_{0s}:

$$E_{i,v,df_0}(f_{0s}) = 100\frac{\left|\hat{F}_{i,v}(f_{0s}) - F_{i,v}\right|}{F_{i,v}} \tag{5.2}$$

$F_{i,v}$ corresponds to the true i^{th} formant frequency of the v^{th} vowel. Some preliminary observations can be made for the traditional linear prediction baseline (m = 1). Figure 41 through Figure 43 show the raw results across methods for the female vowel */ae/* with $df_0 = 25$ Hz.

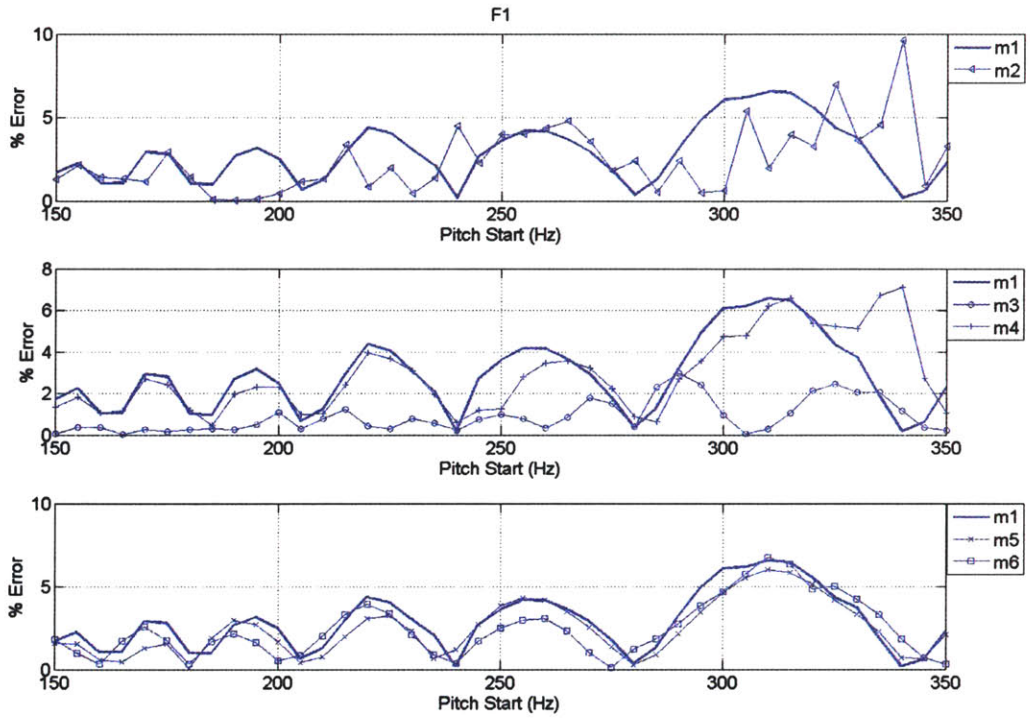

Figure 41 - Raw formant error $E_{i,v}(f_{0s}, df_0)$ for i = 1 (F1), v = 4 */ae/*, and $df_0 = 25$ Hz.

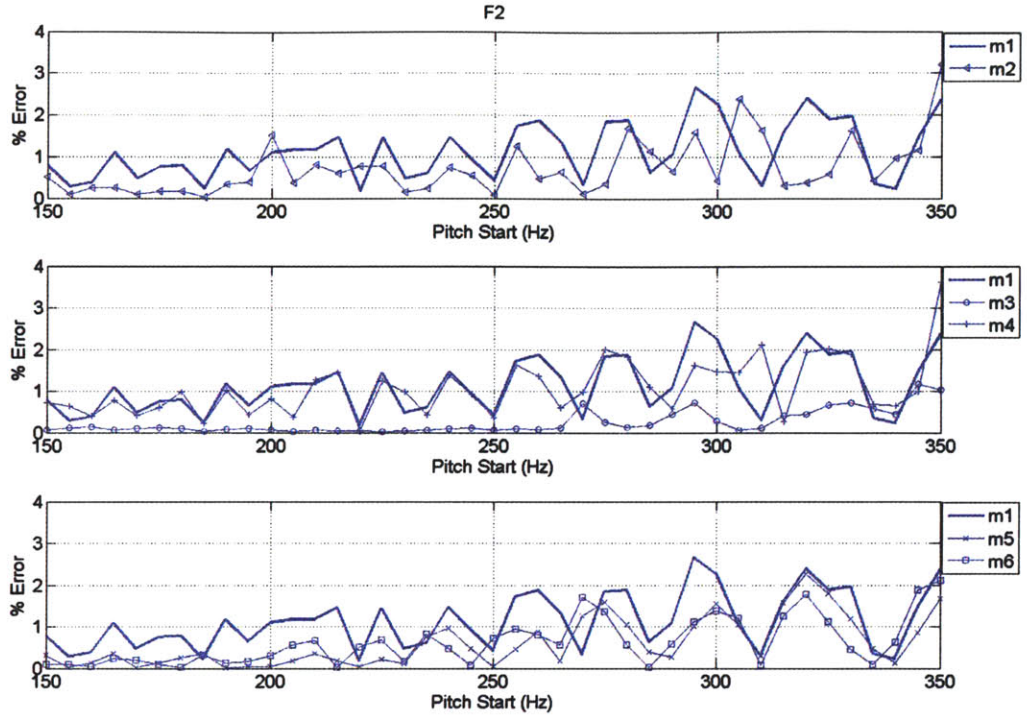

Figure 42 - Raw formant error $E_{i,v}(f_{0s}, df_0)$ for i = 2 (F2), v = *4 /ae/*, and $df_0 = 25$ Hz.

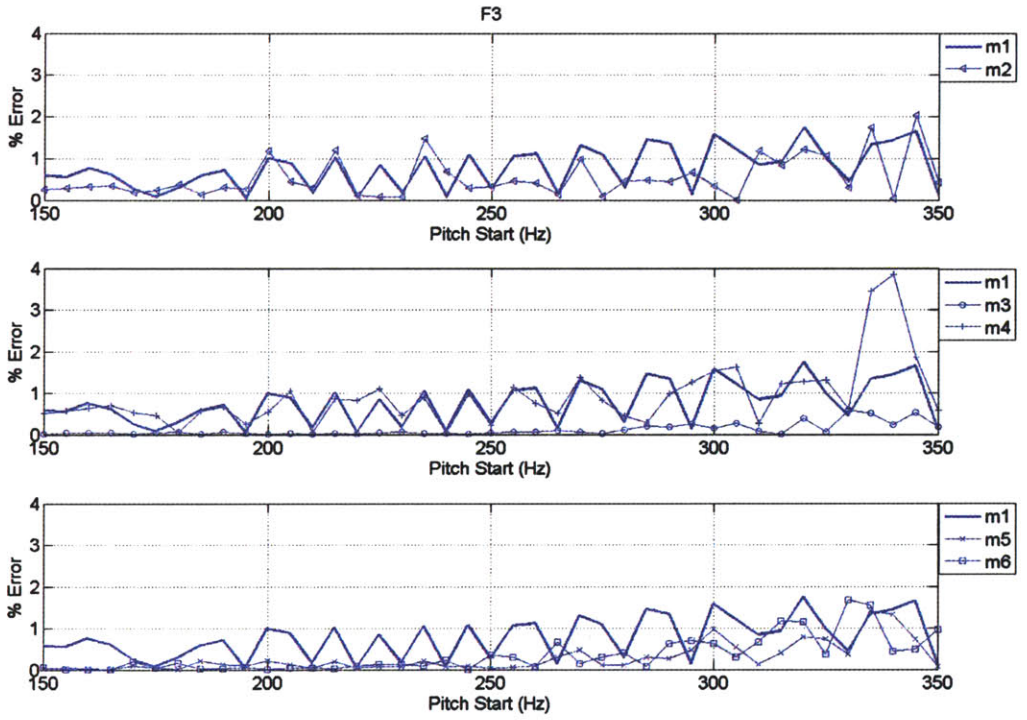

Figure 43 – Raw formant error $E_{i,v}(f_{0s}, df_0)$ for i = 3 (F3), v = *4 /ae/*, and $df_0 = 25$ Hz.

Some preliminary observations can be made for the traditional linear prediction baseline (m = 1). Recall that we use for analysis a 20-ms Hamming window and a 6th order linear prediction estimate. We observe that errors exhibit an "oscillatory" behavior across f_{0s}; in addition, the rate of oscillations tends to increase with formant number, with the fastest oscillations occurring for F3. These observations can at least partially be explained by the "fortuitous sampling" of the formant envelope by harmonic peaks proposed by Vallabha and Tuller [26] (Appendix C); specifically, as pitch changes, harmonics can move towards (away from) formant peaks, thereby leading to good (poor) formant estimates [26]. In accordance with this explanation, pitch changes will also invoke greater absolute changes in harmonic positions for higher frequency regions than lower regions (Section 3.1), such that F3 errors would be expected to oscillate more than F1 errors. Finally, we observe that the size of oscillations increasing with f_{0s}, consistent with the effect of spectral undersampling for higher-pitch formants.

Our results for homomorphic linear prediction (m = 2) are consistent with those reported in [15] in providing gains over traditional linear prediction under some conditions (e.g., F1 – f_{0s} = 295 Hz). Nonetheless, we observe that the errors via harmonic projection and interpolation (m = 3) can afford substantial error reductions; in addition, the similarity between the errors invoked with single-slice interpolation (m = 4) and m = 1 suggest that the error reduction via m = 3 is due to exploiting temporal change of pitch rather than the interpolation method itself. Similarly, m = 5 and m = 6 also appear to afford reductions in the error magnitude under certain conditions (e.g., F2 – m = 6). These results suggest that exploiting temporal change of pitch can improve formant estimation even under high-pitch conditions.

Observe that m = 2, 3, 5, and 6 exhibit some oscillatory behavior similar to m = 1; however, the local maxima of these oscillations can be lower than those of m = 1 (e.g., m = 3, all formants). For m = 2, 3, and 5, we interpret this effect in relation to the "fortuitous sampling" explanation of the oscillatory behavior as increasing the chances of harmonic peaks to align with formant peaks. Nonetheless, this is achieved differently between m = 2 and m = 3 and 5. Whereas cepstral liftering smoothes a spectrum across frequencies and can therefore distribute energy towards the true formant peaks, the projection and interpolation of harmonics does so across time. It appears that the latter method outperforms the former for this purpose (e.g., compare the peak errors for F2 between m = 2 and m = 3). For m = 6, we attribute the reduction in oscillation amplitude (e.g., for F2 and F3) to the improved source-filter separability invoked in 2-D modulation space.

While the raw error metric provides sufficient resolution to analyze individual simulation and estimation results, it remains difficult to compare the performance of different methods "overall". Specifically, across f_{0s}, no single method unanimously outperforms either baseline (m = 1, 2) method. To better interpret our results, several methods of averaging are proposed in the subsequent sections.

5.2 Averaging Across Vowels and Pitch Shifts

In this section, for the the i^{th} formant, we present the average of percent errors across vowels and df_0 as a function of f_{0s}, thereby assessing the overall effect of the starting pitch on formant estimation across methods:

$$E_i(f_{0s}) = \frac{100}{VD} \sum_{v=1}^{V} \sum_{d=1}^{D} \frac{\left| \hat{F}_{i,v,d}(f_{0s}) - F_{i,v} \right|}{F_{i,v}} \tag{5.3}$$

V and D correspond to the total number of vowels and pitch shifts, respectively; as in Section 0, we show the results of averaging for the female vowel *ae* in Figure 44-Figure 46.

Overall, all methods seeking to address the spectral undersampling problem of high-pitch formants (m = 2, 3, 5, 6) appear to provide gains over the traditional linear prediction baseline. Observe, however, that m = 3 unanimously outperforms the baseline for all formants and under all starting pitch values. As in Section 0, m = 4 notably performs worse than m = 3, implicating the role of temporal change in pitch. Finally, note that neither homomorphic linear prediction (m = 2) nor exploiting temporal change of pitch (m = 3, 5, 6) can correct for the trend in poorer formant estimates as pitch increases.

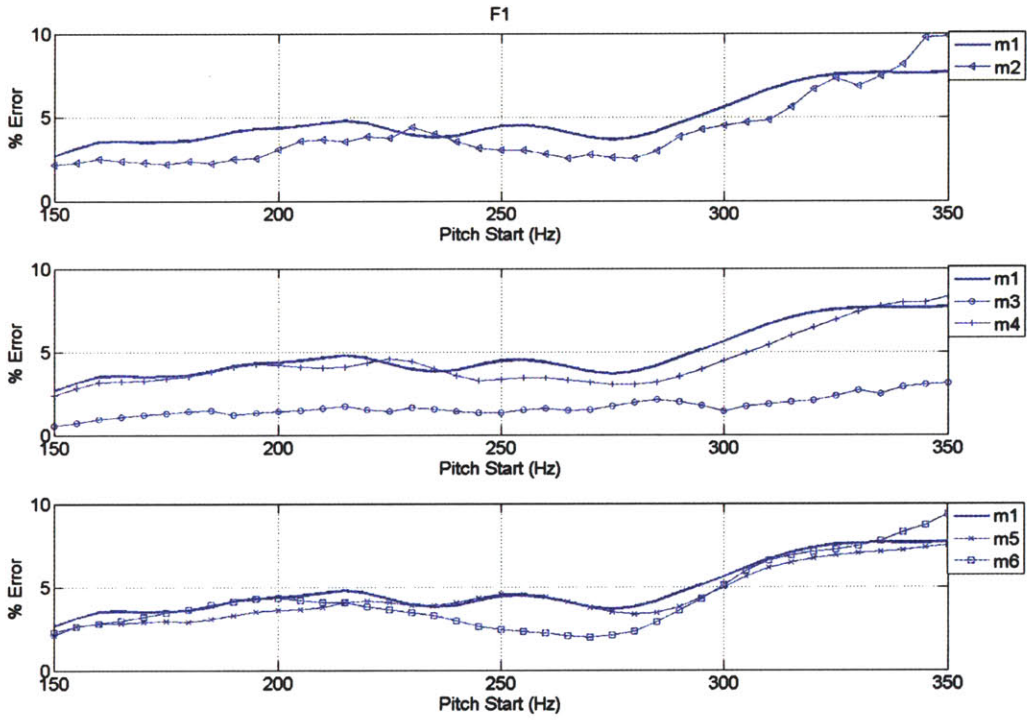

Figure 44 – Average across vowels and pitch starts, $E_i(f_{0s})$, for i = 1 (F1).

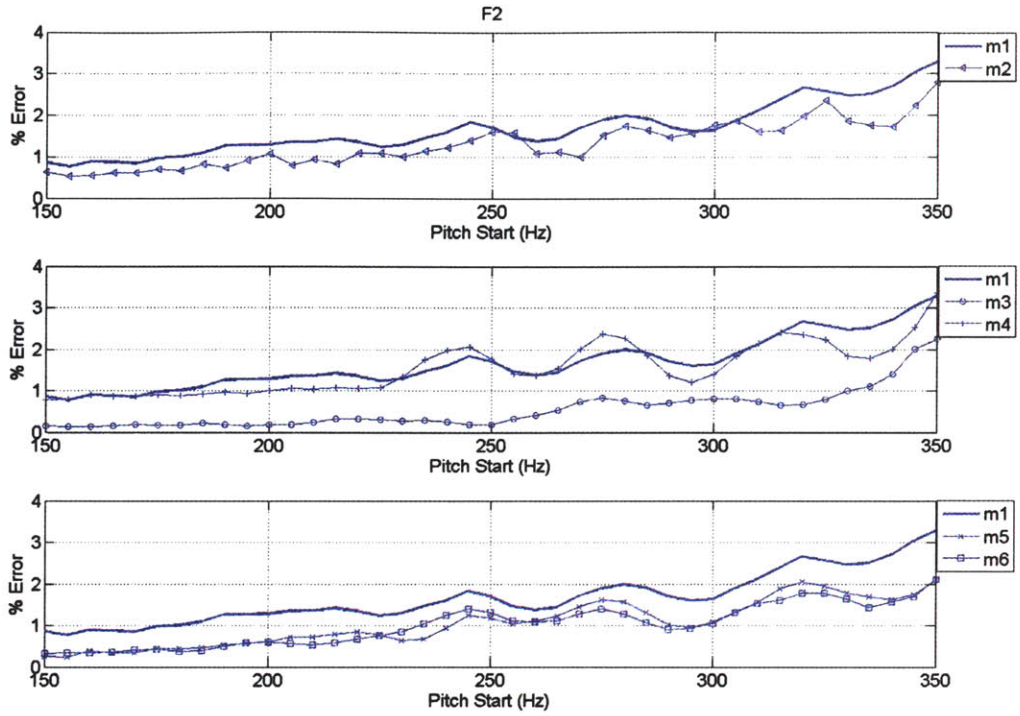

Figure 45 - Average across vowels and pitch starts, $E_i(f_{0s})$, for i = 2 (F2).

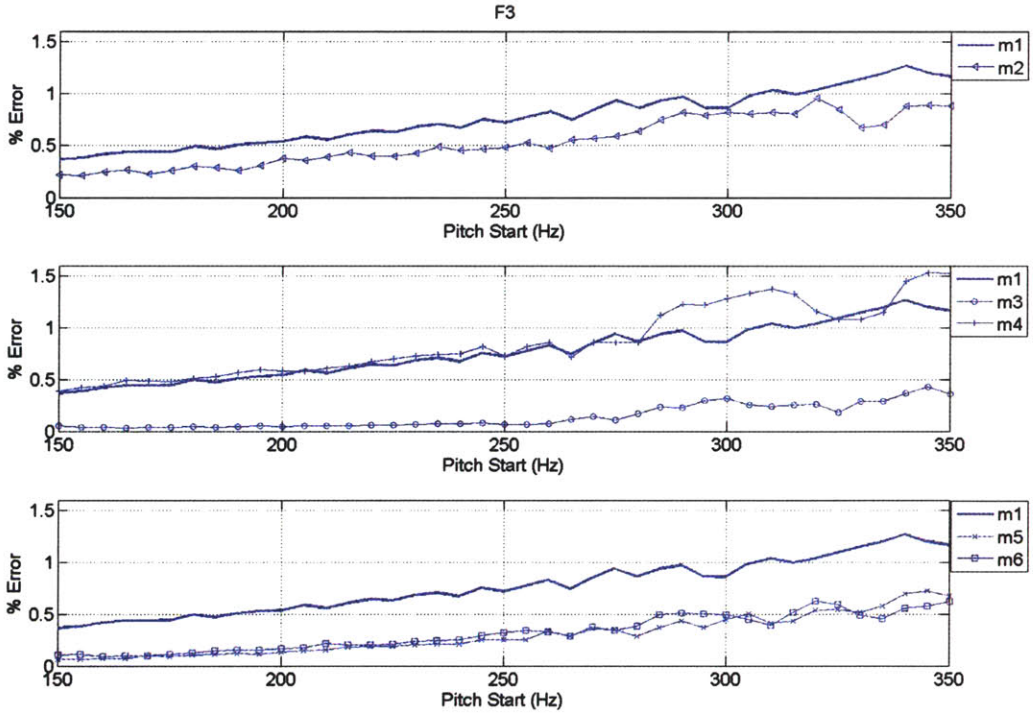

Figure 46 - Average across vowels and pitch starts, $E_i(f_{0s})$, for i = 3 (F3).

5.3 Averaging Across Vowels and Pitch Starts

In this section, for the the i^{th} formant, we present the average of percent errors across vowels and f_{0s} as a function of df_0, thereby gauging the overall effect of the pitch shift on formant estimation across methods:

$$E_i(df_0) = \frac{100}{VS} \sum_{v=1}^{V} \sum_{s=1}^{S} \frac{\left| \hat{F}_{i,v,s}(df_0) - F_{i,v} \right|}{F_{i,v}} \tag{5.4}$$

S corresponds to the number of pitch starts. As in Section 0, we show the results of averaging for the female vowel */ae/* in Figure 47 to Figure 49.

We observe in our results that those methods seeking to exploit temporal change of pitch (m = 3, 5, 6) exhibit greater error reductions with increasing df_0. The magnitude of the gains afforded by exploiting temporal change of pitch appears directly related to the amount of pitch change invoked; specifically, more change invokes larger gains. For m = 3 and 5, we interpret this as larger df_0 allowing for more spectral samples of the underlying formant envelope to be sampled and used for interpolation. Analogously, a larger pitch shift for m = 6 provides an increased $\hat{\theta}$ (Figure 7) such that the harmonic line structure will be transformed in the GCT to a coherent pair of components further off the scale axis. In contrast, we do not observe this same trend for those methods operating on single spectral slices (i.e., m = 1, 2, 4).

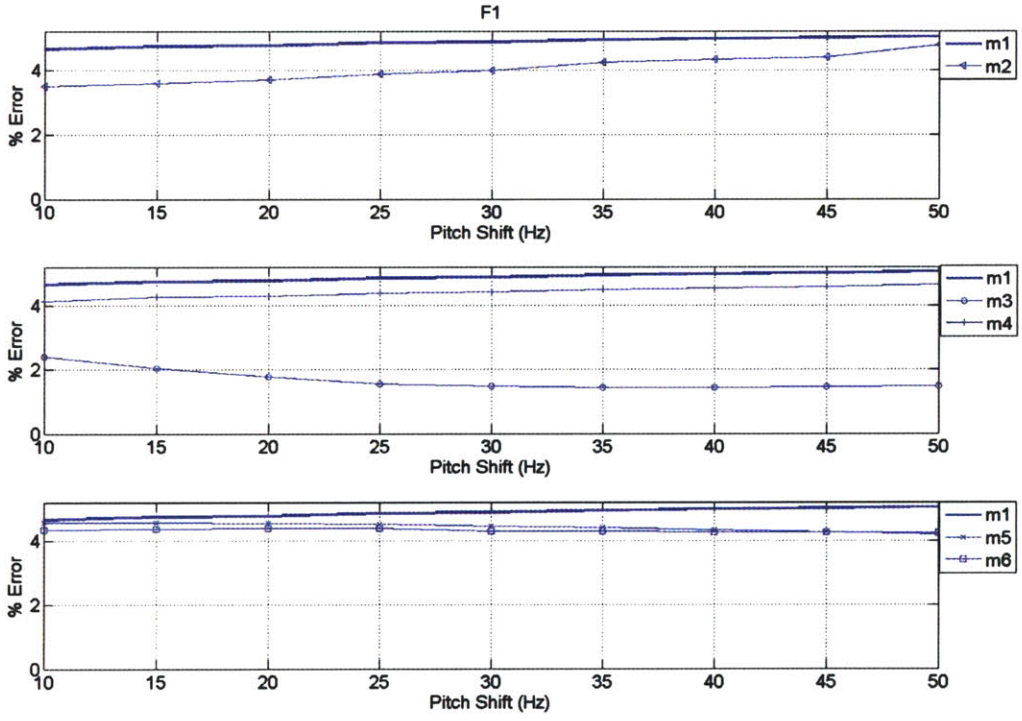

Figure 47 – Averages across vowels and pitch starts, $E_i(df_0)$, for i = 1 (F1).

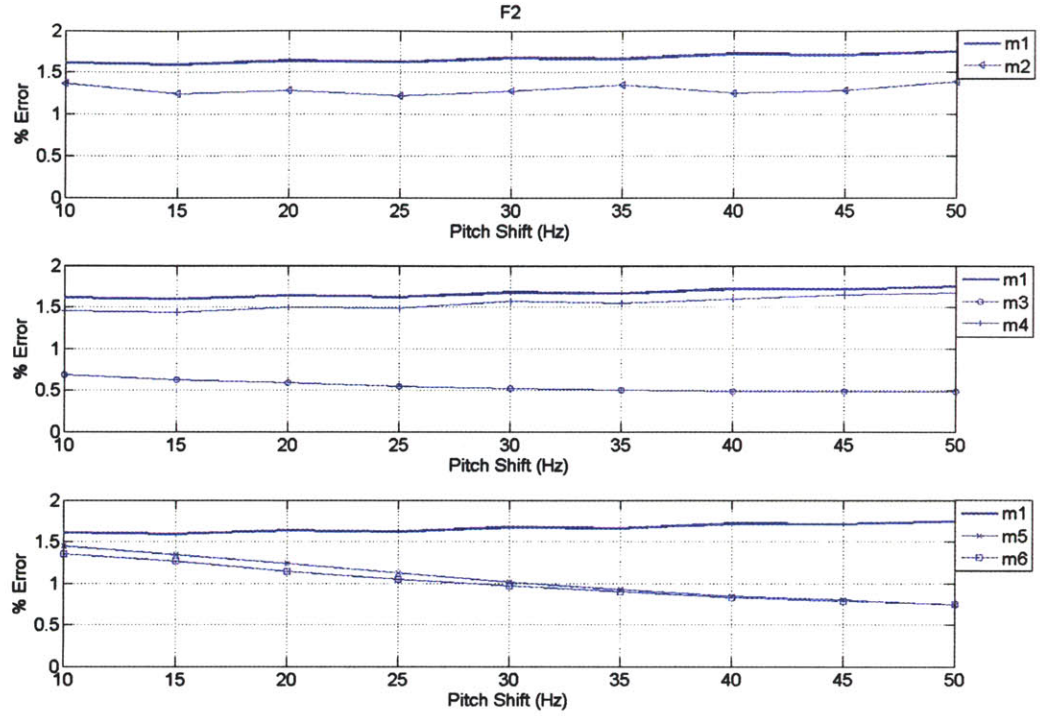

Figure 48 - Averages across vowels and pitch starts, $E_i(df_0)$, for $i = 2$ (F2).

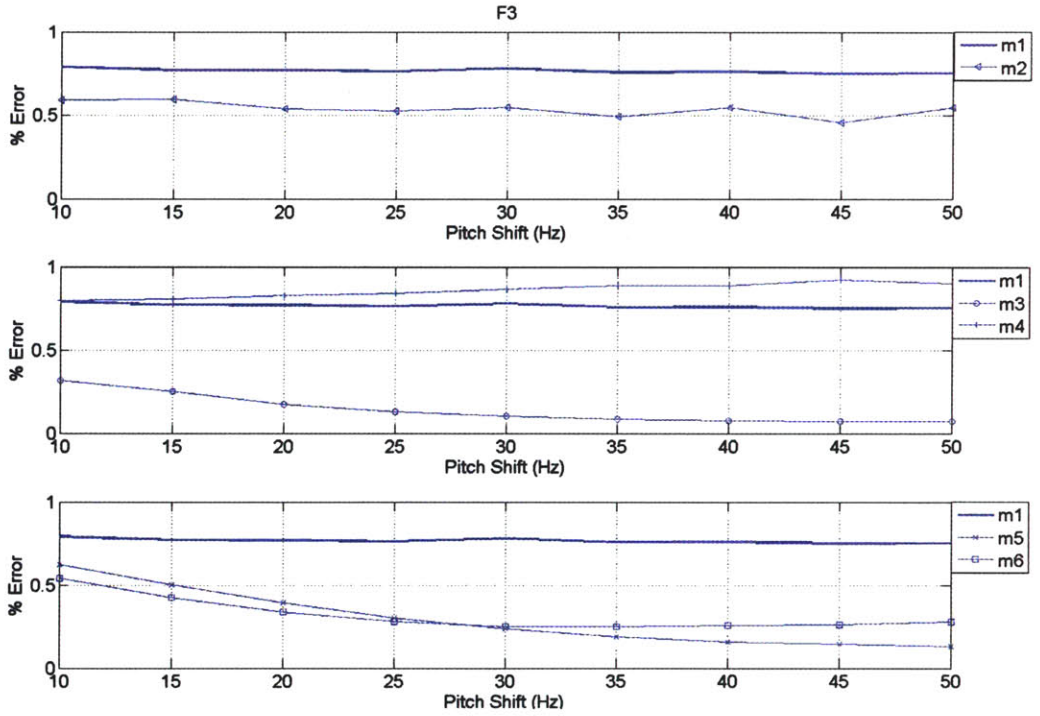

Figure 49 - Averages across vowels and pitch starts, $E_i(df_0)$, for $i = 3$ (F3).

5.4 Global Average

The previous sections focused on the results of females. To summarize our results with respect to gender, we propose a global goodness metric as the average percentage of formant-frequency errors across all df_0, f_{0s}, and v. Specifically, for the i^{th} formant and m^{th} method:

$$E_i(m) = \frac{100}{SDV} \sum_{s=1}^{S} \sum_{d=1}^{D} \sum_{v=1}^{V} \frac{\left| \hat{F}_{i,s,d,v}(m) - F_{i,v} \right|}{F_{i,v}} \tag{5.5}$$

with S, D, and V corresponding to the total number of f_{0s}, df_0, and vowels, respectively. Table 6 shows E_i for m = 1. Using this baseline, we compute the relative *gains* of E_i across all other methods for males, females, and children (Table 7 - Table 9):

$$R(m) = 100 \frac{E_i(1) - E_i(m)}{E_i(1)} \tag{5.6}$$

Table 6 - $E_i(m)$ for $m = 1$ (%).

	males	females	children
i = 1	3.13	4.88	5.68
i = 2	0.86	1.66	2.16
i = 3	0.38	0.77	0.91

Table 7 – Relative gains $R(m)$ for males (%)

	m=2	m=3	m=4	m=5	m=6
i = 1	21.09	66.45	9.58	20.45	9.58
i = 2	27.91	77.91	10.47	59.3	41.86
i = 3	23.68	71.05	-15.79	78.95	52.63

Table 8 - Relative gains $R(m)$ for females (%)

	m=2	m=3	m=4	m=5	m=6
i = 1	17.01	65.78	9.63	9.43	11.68
i = 2	21.69	67.47	7.23	36.14	39.16
i = 3	29.87	81.82	-11.69	61.04	58.44

Table 9 - Relative gains $R(m)$ for children (%)

	m=2	m=3	m=4	m=5	m=6
i = 1	4.93	62.50	12.32	6.16	6.16
i = 2	16.20	46.30	-14.35	26.85	30.56
i = 3	8.79	74.73	-36.26	49.45	51.65

For m = 1, average errors increase for each formant number from males to females to children; our global metric is therefore consistent with the trend observed in Section 5.2 regarding the

reduction in formant estimation accuracy with increasing pitch. For m = 2, our results provide further evidence that homomorphic linear prediction provides gains over traditional linear prediction (m = 1) in formant estimation [15]. Nonetheless, we observe that m = 3 outperforms all other methods with relative gains up ~81% over m = 1. Relative *losses* incurred by single-slice interpolation (m = 4) are again consistent with the role of changing pitch in improving formant estimation via m = 3. Spectral slice averaging (m = 5) and filtering in the GCT (m = 6) also provide gains over m = 1 and 2 for F2 and F3, though not F1. For m = 6, we believe the smaller relative gains in F1 stem from the reduced fanning of harmonic lines in lower frequency regions of the STFT; consequently, these lines are more likely to be mapped along the scale axis in the GCT, thereby reducing source-filter separability. Conversely, broader harmonic sampling for m = 3 and 5 and the increased source-filter separability in the GCT for m = 6 in high-frequency regions likely accounts for the larger gains in F3 than F1 and F2.

It is interesting here to compare the relative performance of m = 6 against m = 2 for F2 and F3. Analytically, recall that the cepstrum and the GCT are similar in transforming a multiplicative source-filter spectral model to an alternate domain where the source and filter are *additive* components; however, the GCT does so without invoking a homomorphic framework. The GCT appears to afford better separability than the cepstrum as assessed by estimating F2 and F3. Presumably, due to the rotational nature of transforming harmonic lines, filtering in the GCT avoids the over-smoothing done in liftering the cepstrum with an increasingly lower quefrency cutoff as pitch increases (Section 2.3.2).

One limitation in interpreting our global metric is that overall relative gains may be dominated by results from distinct vowels. For example, it is conceivable that the gains for females via m = 3, 5, and 6 are due to a large gain for the vowel */ae/* but small relative *losses* incurred for other vowels. To address this confound, we computed the average of percent errors across f_{0s} and df_0, thereby assessing the performance of each method across all vowels,

$$E_i(v) = \frac{100}{SD} \sum_{s=1}^{S} \sum_{d=1}^{D} \frac{\left| \hat{F}_{i,s,d}(v) - F_{i,v} \right|}{F_{i,v}} \tag{5.7}$$

We show in Figure 50 - Figure 52 representative results for F1, F2, and F3, for all vowels for females. The abscissa refers to vowels */ah/* (1), */iy/* (2), */ey/* (3), */ae/* (4), */oh/* (5), and */oo/* (6). The results of males and children are presented in Appendix D. Taken together, these results are generally consistent with the global goodness metric. For instance, observe that m = 5 and 6 provide gains over m = 1 and 2 for F2 and F3. In addition, m = 3 outperforms all other methods for all vowels and formants; concurrently, m = 4 performs consistently worse than m = 3. As previously discussed, the latter result emphasizes the benefits of exploiting temporal change of pitch. Nonetheless, it is interesting to observe that m = 4 can provide gains for all three formants over m = 1 under certain conditions (e.g., */ey/*). This is not entirely surprising, since the interpolation method itself removes harmonic structure and can generate a spectrum that resembles the true formant envelope. Note that this is similar to the spectral smoothing performed in the cepstrum for m = 2.

Several other notable observations can be made for individual vowels. With the exception of F1 for */ah/*, m = 2 for females generally provides gains over m = 1. From Table 3, recall that F1 and F2 are closely spaced (F1 = 850 Hz, F2 = 1220) with a frequency discrepancy of 370 Hz. This is the smallest frequency gap between any two formants for females. Relative to other vowels, the proximity of these formants likely invokes a higher quefrency component in the cepstrum that is

removed with liftering for the relatively high-pitch females (Section 2.3.2). As a result, it appears that while the resulting estimate of F2 maintains a gain over m = 1, removal of fine formant structure leads to a poorer estimate of F1.

For F1, */iy/* (2) exhibits the worst performance for all methods across vowels. In this case, F1 = 310 Hz; recall for females that the range of pitch values is 150 Hz to 400 Hz. At pitch values greater than 310 Hz, pitch harmonics can never sample the resonant peaks of these formants such that a poorer estimate will likely be obtained. While improving spectral sampling by exploiting temporal change in m = 3 can improve F1 estimates relative to both baselines, the method inherently cannot address this issue. The poor performance of F1 for */iy/* contrasts to the relatively (with respect to other vowels) good performance of the vowels */ah/* (1) and */ae/* (4), which have F1 = 850 and 860 Hz, respectively. Comparing those methods aiming to exploit temporal change of pitch (m = 3, 5, 6) between these two vowels, observe that their performance is similar. In contrast, the performance of homomorphic linear prediction (m = 2) is different by more than 1% between the two vowels. A possible cause of this discrepancy could be the broadened spectral sampling (or analogously, improved source-filter separability) at the relatively high-frequency region surrounding this formant. Presumably, while m = 3, 5, and 6 can exploit this effect of changing pitch, m = 2 cannot.

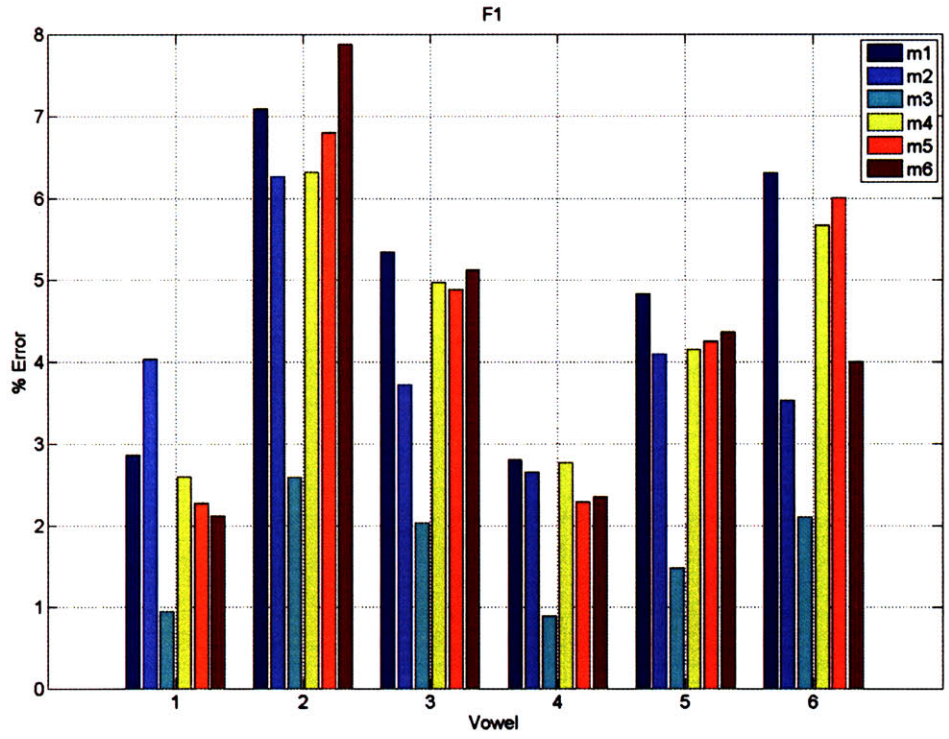

Figure 50 - Averages across pitch starts and pitch shifts, $E_i(v)$, for i = 1 (F1, females).

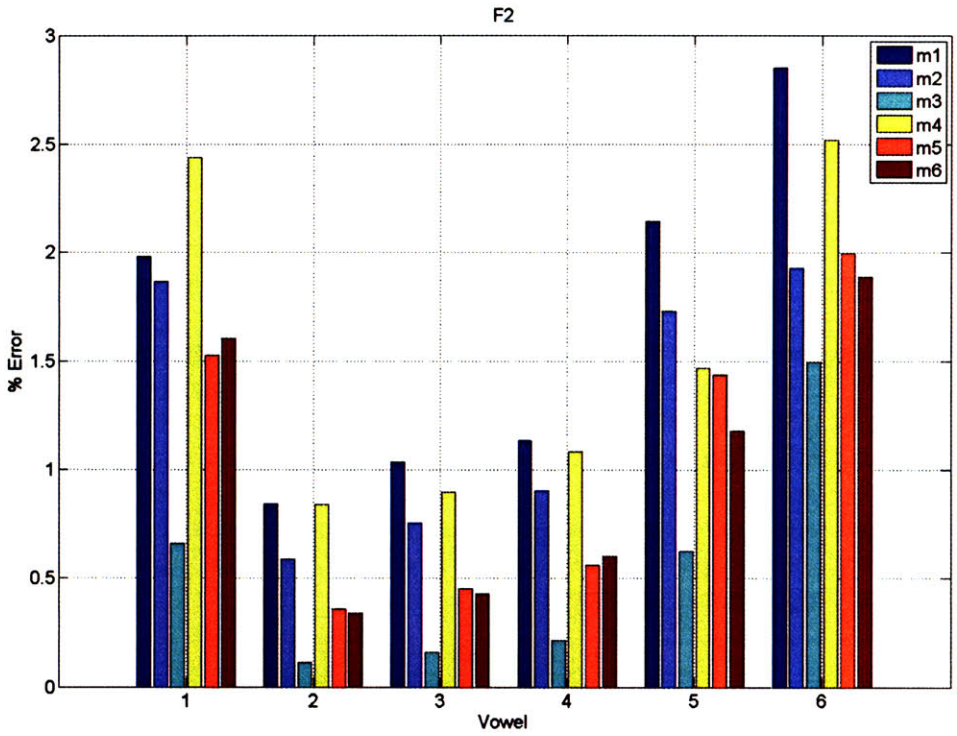

Figure 51 - Averages across pitch starts and pitch shifts, $E_i(v)$, for $i = 2$ (F2, females).

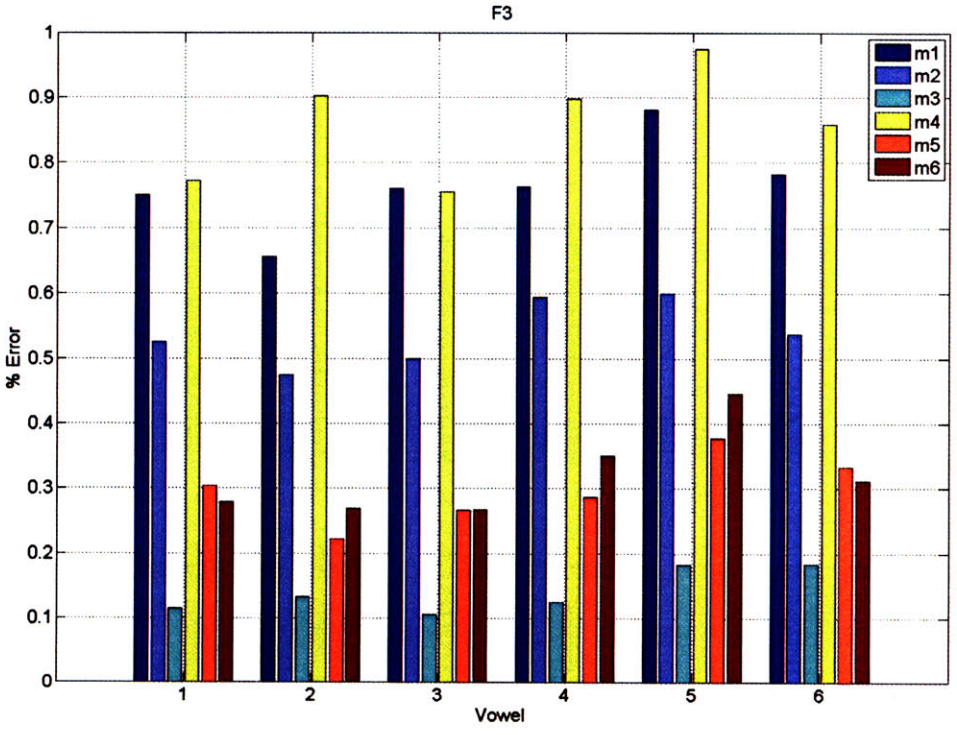

Figure 52 - Averages across pitch starts and pitch shifts, $E_i(v)$, for $i = 3$ (F3, females).

5.5 Conclusions

In this chapter, we have compared the results of traditional and homomorphic linear prediction with novel spectral estimation methods used in conjunction with linear prediction for formant frequency estimation. We have observed that those methods exploiting temporal change of pitch (m = 3, 5, 6) can provide gains over standard one-frame analysis used in our baseline methods (m = 1, 2) (Section 5.4). These gains are related to the amount of pitch change invoked, with greater change affording larger gains (Section 5.3). However, exploiting temporal change of pitch does not appear to reverse the *trend* of poorer formant estimates with increasing pitch (Section 5.2).

One caveat of our findings in relation to methodology is the discrepancy in short-time analysis methods used between our baseline methods (m = 1 and 2) and m = 3 through 6. Whereas for m = 3 through 6 short-time analysis was performed with a pitch-adaptive Blackman window, our baselines used a fixed-duration 20-ms Hamming window. To address the possibility that gains for m = 3, 5 and 6 were due to this difference in short-time analysis alone, we performed formant estimation using m = 1 and m = 2 but with a spectral slice extracted from the middle of $STFT_1$ instead of $STFT_0$ (i.e., traditional and homomorphic linear prediction using pitch-adaptive short-time analysis). A comparison of the global average metric resulting from this for m = 1 and 2 with m = 3 through 6 is presented in Appendix E and is consistent with our view that exploiting temporal change of pitch can improve high-pitch formant estimation.

Chapter 6

Speaker Recognition Experimentation

In this chapter, we assess the value of the proposed methods for deriving speech spectra in the particular application of speaker recognition. Specifically, system performance has been shown to exhibit a "gender gap", with better performance on male data sets versus females. We hypothesized that one contributing factor to this gap is the poorer spectral representation of formant structure due to the higher-pitch females. Motivated by our improved spectral estimates, *as characterized by improved formant estimation of high-pitch speech* (Chapter 5), we incorporated the proposed framework into a state-of-the-art speaker recognition system as a basis for feature extraction with the aim of addressing this gender gap.

This chapter is organized as follows. In Section 6.1, we present an overview of the speaker recognition system. Section 6.2 discusses the motivation and details of our experimental setup. We discuss the results of our experiments in Section 6.3 and summarize our findings in Section 6.4.

6.1 Speaker Recognition System Overview

In this thesis, we define the goal of a speaker recognition system as determining whether a speech waveform claiming to be a particular speaker corresponds to the claimant or an imposter. System development involves two primary components: feature extraction and pattern recognition. Pattern recognition can further be divided into *training* and *testing* steps. In this section, we discuss the details of the speaker recognition system used in assessing our proposed spectral estimation methods. Section 6.2 addresses feature extraction. Section 6.3 discusses the particular method of pattern recognition known as Gaussian mixture modeling in the *training* stage, and Section 6.4 discusses the *testing* stage and evaluation criteria.

6.1.1 Feature extraction

Feature extraction aims to obtain from a speech waveform a set of "useful" (as determined by the task) parameters via signal processing techniques. In this thesis, we employed mel-frequency cepstral coefficients (MFCC) [3, 27]. The baseline MFCCs (denoted as $MFCC(l)$ for l = 0, 1, ... L-1) are computed by performing short-time Fourier analysis on a speech waveform $x[n]$. For a single spectral slice $X[k]$ (for $k = 0,1...N-1$, where N is the DFT length), its squared-magnitude ($|X[k]|^2$) is weighted by a bank of 24 mel-frequency weightings (Figure 53); denoting each mel-frequency weighting as $H_{mel,l}[k]$, a set of log-mel-energies ($E_{mel}[l]$) are computed as:

$$E_{mel}[l] = \log\left(\sum_{k=0}^{N-1} |X[k]|^2 H_{mel,l}[k]\right) \tag{6.1}$$

Using the log and discrete-cosine transform operations, the MFCCs are then defined as:

$$MFCC(l) = \sum_{m=0}^{L-1} E_{mel}[m]\cos\left(\frac{\pi l}{L}(m+0.5)\right) \tag{6.2}$$

In this thesis, the feature vector is comprised of $MFCC(l)$ for $l = 1$ through 19.

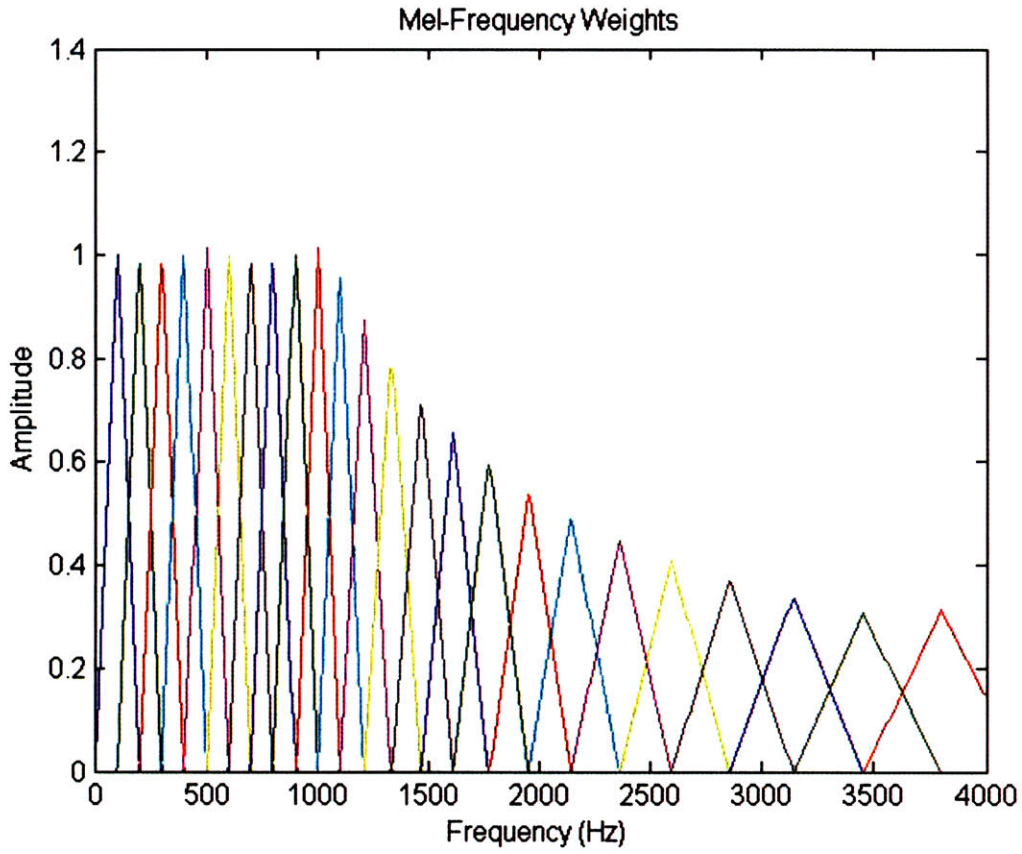

Figure 53 – Mel-frequency weightings.

6.1.2 Gaussian Mixture Modeling (Training)

In the training stage of pattern recognition, feature extraction is first performed on a speech corpus that is labeled by distinct speakers. A pool of features obtained from multiple speakers is initially used in estimating the parameters of a "universal" background speaker model (UBM) using Gaussian mixture modeling. Specifically, define $\vec{x}_t = [x_1, x_2, ..., x_D]$ as a vector of D features (e.g., MFCC) for $t = 1, 2, ...T$; here, T corresponds to the total number of feature vectors used in obtaining the UBM. The probability distribution of the set of T length-D vectors is modeled as a linear combination of M D-variate Gaussian distributions such that [28]

$$p(\vec{x}\,|\,UBM) = \sum_{m=1}^{M} w_m f_m(\vec{x})$$
$$f_m(\vec{x}) = \frac{1}{\sqrt{2\pi}^D |\Sigma_m|^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_m)'\Sigma_m^{-1}(\vec{x}-\vec{\mu}_m)} \tag{6.3}$$

where w_m corresponds to the m^{th} mixture weight such that $\sum_{m=1}^{M} w_m = 1$, $\vec{\mu}_m$ is the m^{th} mean vector, and Σ_m is the m^{th} covariance matrix. Assuming that the set of T length-D vectors are mutually independent, their likelihood for any given set of parameters $\vec{\mu}_m, w_m, \Sigma_m$ for m = 1, 2, ... M is:

$$p(\vec{x}_1, \vec{x}_2, ..., \vec{x}_T \,|\, \vec{\mu}_m, w_m, \Sigma_m; \forall m) = \prod_{t=1}^{T} p(\vec{x}_t \,|\, \vec{\mu}_m, w_m, \Sigma_m; \forall m) \tag{6.4}$$

Estimation of these parameters is done based on the criterion of maximum likelihood via the expectation maximization algorithm [28]. The UBM is therefore completely characterized by $\vec{\mu}_m, w_m, \Sigma_m$ for $m = 1, 2, ... M$.

The next step in training is developing *target* speaker models corresponding to distinct speakers. In contrast to the UBM, a set of $\vec{x}_t = [x_1, x_2, ..., x_D]$ are pooled for t = 1, 2, ..., T_s where T_s now corresponds to the number of feature vectors corresponding to a *single* speaker. We denote a single speaker as S_p for p = 1, 2, ... P where P corresponds to the total population of distinct speakers. With the same assumptions used in obtaining the UBM, we again wish to maximize the likelihood

$$p(\vec{x}_1, \vec{x}_2, ..., \vec{x}_{T_s} \,|\, \vec{\mu}_{m,p}, w_{m,p}, \Sigma_{m,p}; \forall m) = \prod_{t=1}^{T} p(\vec{x}_t \,|\, \vec{\mu}_{m,p}, w_{m,p}, \Sigma_{m,p}; \forall m) \tag{6.5}$$

with respect to a choice of parameters $\vec{\mu}_{m,p}, w_{m,p}, \Sigma_{m,p}$ for m = 1, 2, ... M that completely characterize S_p. Instead of estimating these parameters employing the expectation maximization algorithm, as was done for the UBM, the UBM parameters are used in an adaptation scheme discussed in [29] to obtain estimates of $\vec{\mu}_{m,p}, w_{m,p}, \Sigma_{m,p}$.

6.1.3 Gaussian Mixture Modeling (Testing)

Upon completion of training the UBM and the target speaker models, test utterances are used to evaluate the overall system. Given a test speech waveform $x_{test}[n]$[9] corresponding to one of the P target speakers (denoted here as S_{test}), a feature vector $\vec{x}_{test}$ is extracted as before. Next, P pair-wise trials are conducted across all targets speakers and the UBM by comparing $P(S_p \,|\, \vec{x}_{test})$

[9] The test waveform must *not* have been used in training the target speaker models or UBM.

with $P(UBM \mid \vec{x}_{test})$ (i.e., given the feature vector observation, the probability that it arose from the target speaker model S_p against the probability that it arose from the UBM):

$$P(S_p \mid \vec{x}_{test}) = \frac{p(\vec{x}_{test} \mid S_p)P(S_p)}{p(\vec{x}_{test})}$$
$$P(UBM \mid \vec{x}_{test}) = \frac{p(\vec{x}_{test} \mid UBM)P(UBM)}{p(\vec{x}_{test})} \tag{6.6}$$

If we consider the UBM as simply another target speaker and assume all speaker models are equally likely (i.e., $\frac{1}{P+1}$), then $P(UBM) = P(S_p)$. Under these conditions, note that $p(\vec{x}_{test}) = \sum_{p=1}^{P+1} p(\vec{x}_{test} \mid S_p)P(S_p)$ (where p now ranges from 1 to P+1 to account for the UBM), and is the same for both $P(S_p \mid \vec{x}_{test})$ and $P(UBM \mid \vec{x}_{test})$. To compare the two probabilities, we compute the log-likelihood ratio $L(\vec{x}_{test})$ defined as [29]:

$$L(\vec{x}_{test}) = \log\frac{P(S_p \mid \vec{x}_{test})}{P(UBM \mid \vec{x}_{test})} = \log\frac{p(\vec{x}_{test} \mid S_p)}{p(\vec{x}_{test} \mid UBM)} = \log p(\vec{x}_{test} \mid S_p) - \log p(\vec{x}_{test} \mid UBM) \tag{6.7}$$

Finally, a threshold c can be set such that a decision can be made regarding $\vec{x}_{test}$ as arising from either S_p or the UBM:

$$L(\vec{x}_{test}) \geq c \rightarrow \vec{x}_{test} \; from \; S_p$$
$$L(\vec{x}_{test}) < c \rightarrow \vec{x}_{test} \; from \; UBM \tag{6.8}$$

As previously mentioned, the task of speaker recognition in this thesis is to assess whether or not a speech waveform of a claimant corresponds to the claimed speaker or an imposter. We interpret the first condition of (6.8) as the case when $x_{test}[n]$ is "accepted" as the claimed speaker while the second corresponds to when $x_{test}[n]$ is "rejected" on grounds that it came from an imposter. Two types of errors can be made by this system: 1) if the claimant does *not* correspond to the claimed target speaker and $x_{test}[n]$ is accepted ("miss") and 2) if the claimant corresponds the claimed target speaker and $x_{test}[n]$ is rejected ("false alarm"). Depending on the choice of c, the system may be biased to invoke more false alarms than misses or vice versa. A detection error tradeoff (DET) curve plots, for all choices of c, the probabilities of these two errors. The equal-error rate (EER) is the operating point at which both errors are equal and is the metric used in this thesis for assessing recognition performance.

6.2 Experimental Setup

Our hypothesis in speaker recognition experimentation is that the short-time Fourier analysis performed in the standard mel-cepstral feature set (Section 6.1.1) is insufficient in representing

formant structure due to spectral undersampling, thereby contributing to the observed performance gap between males and females. We emphasize that spectral undersampling may not be the *only* cause of the gender gap. For instance, females may exhibit fewer formants over a fixed bandwidth relative to males, which cannot be addressed by simply improving spectral sampling. In addition, while we aim to improve the formant representation in the input spectrum to MFCCs, our methods may concurrently eliminate other characteristics of the spectrum (e.g., source fine structure) that contribute to speaker identity.

In this thesis, we present two sets of speaker recognition results: all-speech and vowel-only. In [30], we presented the results of the all-speech experiment suggesting a gender-gap closure on the TIMIT corpus using features sets motivated from our 2-D processing framework. These results motivated us to perform a more rigorous assessment of the potential benefits of our framework in addressing the gender gap. Herein we provide a discussion of our methodology in both sets of experiments. Section 6.2.1 discusses the motivation and choice of data set used and details of GMM training and testing. In Section 6.2.2, we give details of baseline and our proposed feature extraction methodologies.

6.2.1 Data Set and Gaussian Mixture Modeling

In this thesis, we perform speaker recognition experiments on male and female subsets of the TIMIT corpus [31]. TIMIT was recorded in the clear at a 16 kHz sampling rate, is phonetically balanced, and has a good representation of speaker types and dialects; in our experiments, all waveforms were first downsampled to 8 kHz. We chose TIMIT as our evaluation corpus in order to assess the value of our 2-D framework in feature extraction without the influence of uncontrolled background conditions (e.g., noise) and/or channel distortion.

In the all-speech experiment, the UBM was derived using 1680 sentences spoken by 1120 males and 560 females. In the vowel-only experiments, we sought to address the potential confound of the discrepancy in male and female speakers used in training the UBM and thereby used 1120 total sentences from 560 males and 560 females. In both sets of experiments, we used for the target speakers 137 females and 327 males; each target model was trained using 8 sentences. 128 Gaussian mixture components were used for both the UBM and target models. Evaluation was done using 273 and 653 sentences for females and males, respectively.

6.2.2 Feature Extraction

We describe here our feature extraction methodology for both sets of experiments. In all cases, the waveform was first pre-emphasized with a first-order high-pass filter $H(z) = 1 - 0.97z^{-1}$.

All-speech experiments: In the all-speech experiments, we used for a baseline feature set (denoted as $f_{all} = 0$) MFCCs derived as described in Section 6.1.1. Specifically, short-time spectra were derived using a 20-ms Hamming window and 10-ms frame interval followed by computation of the log-mel-energies and the DCT. The proposed features used instead a 10-ms Hamming window and a 2-ms frame interval for computing short-time spectra. At each time interval, the average of 10 spectral slices was computed and used as the input spectrum to the mel-cepstrum ($f_{all} = 1$). To control for the discrepancy in short-time analysis methods between the baseline and proposed features, MFCCs were also computed from spectra derived from the proposed method but *without* averaging ($f_{all} = 2$). All feature sets were allied with deltas features

across a 50-ms interval. Note that in this set of experiments, *all* of the speech in each sentence was used for extracting features.

Vowel-only experiments: In the vowel-only experiments, we aimed to isolate the potential effects on speaker recognition performance due specifically to spectral undersampling of vowels. Using the time-aligned transcriptions of the TIMIT corpus, we extracted speech corresponding *only* to vowels to be used in feature extraction. The vowels selected consisted of both monopthongs and dipthongs: /iy/, /ih/, /eh/, /ey/, /ae/, /aa/, /aw/, /ay/, /ah/, /ao/, /oy/, /ow/, /uh/, /uw/, /ux/, /er/, /ax/, /ix/, /axr/, /ax-h/. In addition, we did not employ delta features. As in the all-speech experiments, we used for a baseline feature set MFCCs computed from short-time analysis with a 20-ms Hamming window and 10-ms frame interval denoted as $f_{vo} = 0$ (though computed only for regions corresponding to vowels).

The proposed feature sets were motivated from the harmonic projection and spectral slice averaging methods described in Chapter 3 for exploiting temporal change of pitch across stationary vowels[10]. Motivated from our results in formant estimation, we used the same pitch-adaptive short-time analysis with a Blackman window of length four times the pitch period as in Section 4.2.1; we denote the resulting STFT as $STFT_{lf}$. To obtain estimates of pitch for use in determining window lengths, we employed an existing implementation[11] of the super-resolution pitch tracking algorithm with frame rate set to 1-ms [32].

For feature extraction using harmonic projection, we performed peak-picking using the SEEVOC algorithm [24] on spectral slices of $STFT_{lf}$. The peaks of several spectral slices were then merged and linearly interpolated across frequency as in formant estimation. The number of slices merged was based on a "window length" denoted as w. For example, for w = 20 ms, 20 slices of $STFT_{lf}$ were merged (consistent with the 1-ms frame interval used in computing $STFT_{lf}$). The interpolated spectrum was then used as the input spectrum (i.e., $|X[k]|$) for computing MFCCs as described in Section 6.1.1. These steps are illustrated in Figure 54 for w = 20 ms. Because our methods aim to exploit temporal changes in pitch, we experimented with four values of w: 20, 30, 40, and 50 ms to incorporate varying degrees of potential pitch dynamics in generating the MFCC input spectrum. Features were computed at a 10-ms frame interval to match the frame interval of the baseline feature set. We denote these feature sets as f_{vo} = 1-w (e.g., f_{vo} = 1-20 for w = 20 ms).

For the method of spectral slice averaging, features were generated in a similar fashion as in the harmonic projection method. Specifically, the input spectrum for computing MFCCs was derived by averaging several spectral slices of $STFT_{lf}$. As in f_{vo} = 1-w, the number of slices to average was determined by a choice of w. The values of w were the same as those in the harmonic projection method (20, 30, 40, and 50 ms). Features were also computed at a 10-ms frame interval to match the baseline feature set. We denote these features as f_{vo} = 2-w.

[10] Time was not permitted to employ filtering in the GCT for feature extraction.
[11] Software provided by Michael Brandstein.

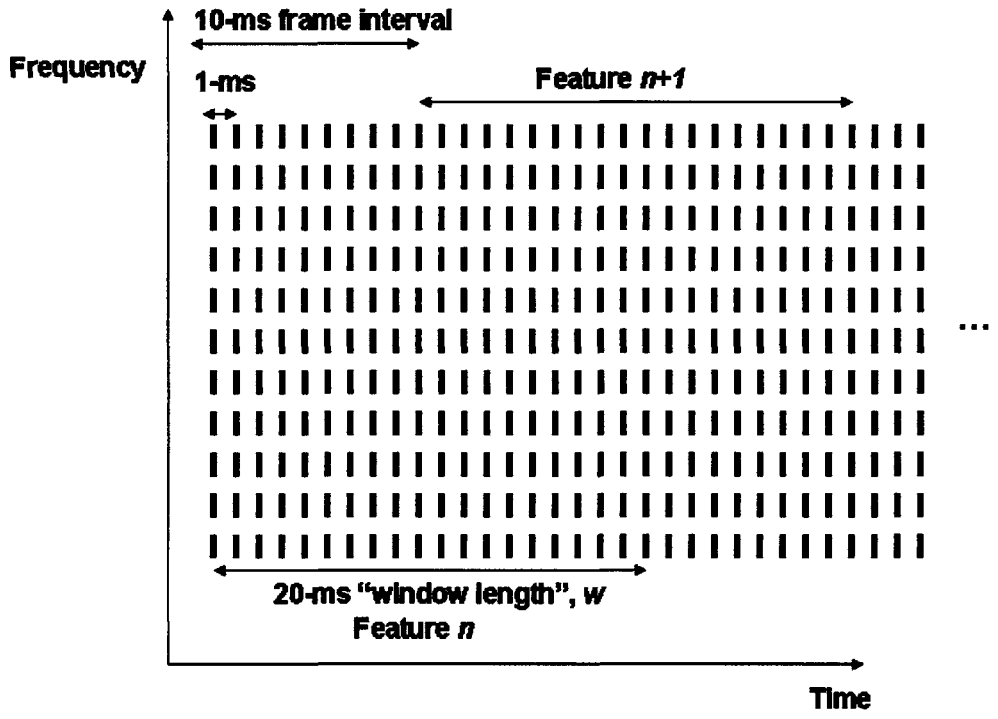

Figure 54 – Figure illustrating feature extraction method in vowel-only experiments for harmonic projection and spectral slice averaging with $w = 20$ ms. Dashed lines correspond to spectral slices of $STFT_{lf}$ computed at a 1-ms frame interval. For feature n, 20 slices are merged to generate the spectrum input for computing MFCCs; feature $n+1$ is generated from spectral slices located 10-ms (or 10 spectral slices) later in time to invoke a 10-ms frame interval for the feature stream.

6.3 Results

All-speech experiments: In Figure 55, we plot the detection-error tradeoff (DET) curves for the all-speech experiments while Table 10 lists the corresponding equal-error rates (EER)[12]. While the baseline features ($f_{all} = 0$) exhibit a gap between females and males, the proposed feature set ($f_{all} = 1$) affords an absolute reduction in EER of 2.26% for females while maintaining the performance of males. It appears then, that the proposed features close the gender gap. As previously noted, a caveat in interpreting these results is the discrepancy between the window lengths and frame rates used in the baseline and proposed methods. Observe that the performance of females using the 10-ms window and 2-ms frame interval ($f_{all} = 2$) *without* averaging resulted in an EER of 2.85%. While this result shows that the proposed method may afford the gain in females due to the choice of window and frame interval in short-time analysis alone, the larger absolute gain of the proposed method appears promising.

[12] We refer the reader to [3] for a detailed discussion of these performance measures.

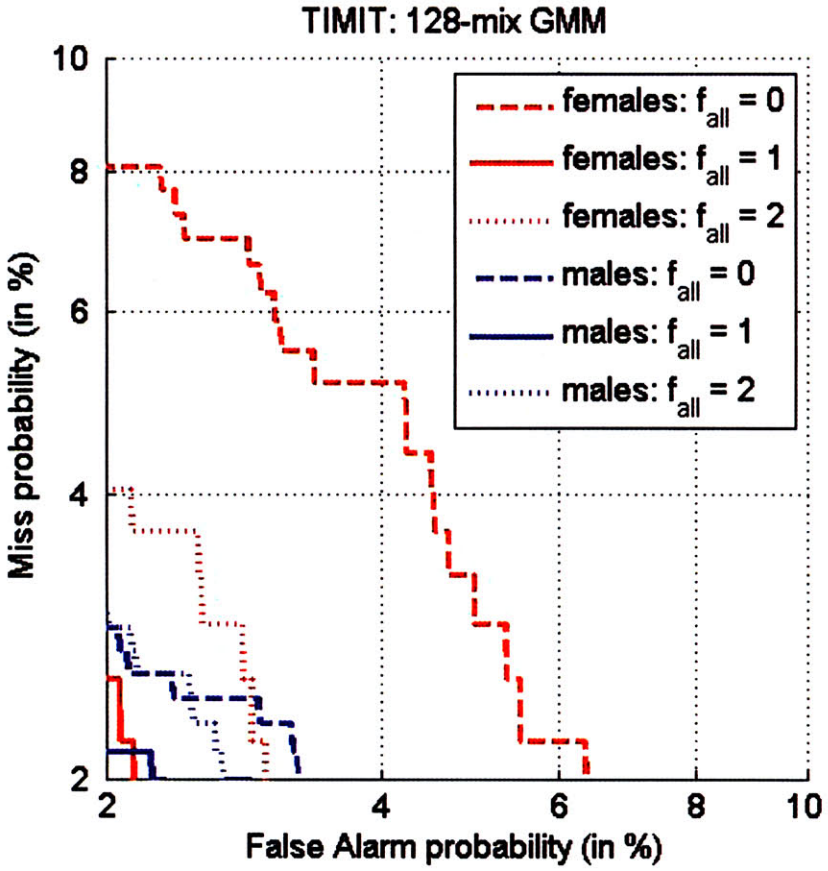

Figure 55 – DET plot for all-speech experiments.

Table 10 – EER for all-speech experiments (%). Confidence intervals are at the 95% level.

	f_{all} = 0 (baseline)	f_{all} = 1 (proposed)	f_{all} = 2 (control)
males	1.86 < 2.45 < 3.39	1.53 < 2.15 < 2.80	1.88 < 2.49 < 3.01
females	3.12 < 4.41 < 5.64	1.55 < 2.15 < 3.30	2.01 < 2.85 < 3.72

Vowel-only experiments: In Table 10 and Table 11, we present the EER values for males and females using the method of harmonic projection (f_{all} = 1-*w*) and spectral slice average (f_{vo} = 2-*w),* respectively. Baseline MFCC (f_{all} = 0) results are also presented as a reference; observe for the baseline that there exists a performance gap between males and females of 2.24% in absolute EER. Similar observations can be made for both methods aiming to exploit temporal dynamics of pitch. In particular, neither method alone outperforms the baseline MFCCs. As previously discussed, one cause of this may be that in the process of improving the spectral representation of formant structure, we have discarded other distinctive spectral content, particularly glottal source fine structure. The performance of both methods also degrades as *w* increases for both genders. This effect may be due to the presence of dipthongs used in feature extraction. As *w* increases, more slices are merged over a longer duration such that the vowel cannot be assumed to be stationary (as in a monopthong). The resulting interpolation of harmonic peaks is likely distorted in the sense that it does not represent spectral samples of a single formant envelope. Finally, it interesting to observe that under certain conditions, the proposed methods appear to "close" the gap though at the cost of degrading male performance (e.g., f_{vo} = 1-20, f_{vo} = 1-40); further experimentation is necessary to interpret this finding.

Table 11 – Equal error rates for vowel-only experiment (%) for the method of harmonic projection (f_{vo} = 1-w) for all values of w and the baseline MFCC features (f_{vo} = 0). Confidence intervals are at the 95% level.

	males	females
f_{vo} = 0	3.64 < 4.01 < 6.89	4.78 < 6.25 < 8.49
f_{vo} = 1-20	6.55 < 7.52 < 8.41	6.34 < 7.72 < 10.26
f_{vo} = 1-30	6.54 < 7.71 < 8.75	7.57 < 8.91 < 10.71
f_{vo} = 1-40	7.20 < 8.44 < 10.14	6.92 < 8.28 < 10.75
f_{vo} = 1-50	7.51 < 8.79 < 10.16	10.38 < 12.13 < 15.43

Table 12 - Equal error rates for vowel-only experiment (%) for the method of spectral slice averaging (f_{vo} = 2-w) for all values of w and the baseline MFCC features (f_{vo} = 0). Confidence intervals are at the 95% level.

	males	females
f_{vo} = 0	3.64 < 4.01 < 6.89	4.78 < 6.25 < 8.49
f_{vo} = 2-20	7.23 < 8.13 < 9.47	5.72 < 6.64 < 9.33
f_{vo} = 2-30	7.35 < 8.53 < 9.64	6.64 < 8.09 < 10.26
f_{vo} = 2-40	7.51 < 8.51 < 9.64	7.26 < 9.19 < 11.13
f_{vo} = 2-50	8.31 < 9.20 < 10.29	8.86 < 10.74 < 13.27

To address the possibility that the degraded performance of the proposed methods was due to the discarding of source fine structure, we also obtained results from fusion of log-likelihood ratios between the MFCC baseline and the harmonic projection method. A linear fusion with equal weighting between f_{vo} = 0 and f_{vo} = 1-20 was performed for both males and females. Figure 56 shows a DET plot comparing f_{vo} = 0, f_{vo} = 1-20, and the fused results denoted as f_{vo} = 0+1-20. The fused result appears to provide a gain over the baseline MFCC feature set alone for females while maintaining the performance of males. Table 13 lists the EER of the fused result showing an absolute reduction in EER of 1.47% for females relative to the MFCC baseline; in addition, male performance is moderately degraded with an increase in absolute EER of 0.1%. Recall that for the MFCC baseline, a gap of 2.24% exists between males and females. The proposed features appear then to provide complementary information to this baseline with an effect of reducing this gap by 1.57% from the results of f_{vo} = 0+1-20. Nonetheless, this gap closure is obtained at the cost of a moderate reduction in the performance of males of 0.1%.

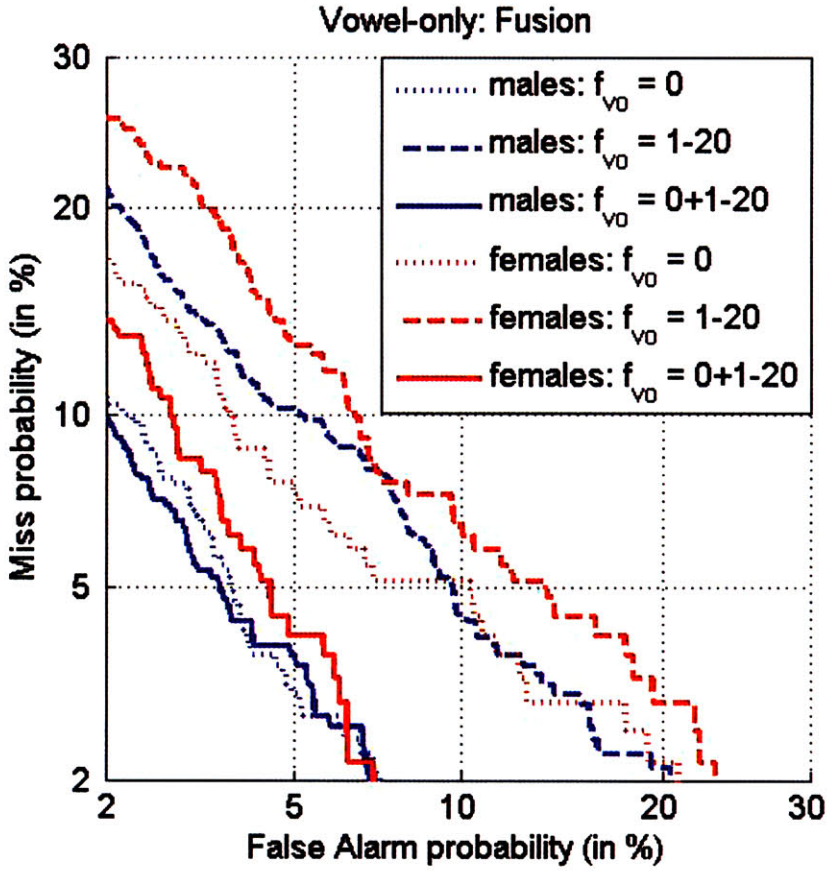

Figure 56 - DET plot comparing f_{vo} = 0, f_{vo} = 1-20, and the fused result denoted as f_{vo} = 0+1-20.

Table 13 - Equal error rates for vowel-only experiment (%) for the baseline MFCC features (f_{vo} = 0), the method of harmonic projection with w = 20 ms (f_{vo} = 1-20) for all values of w and f_{vo} = 0+1-20. Confidence intervals are at the 95% level.

	males	females
f_{vo} = 0	3.64 < 4.01 < 6.89	4.78 < 6.25 < 8.49
f_{vo} = 1-20	6.55 < 7.52 < 8.41	6.34 < 7.72 < 10.26
f_{vo} = 0+1-20	3.48 < 4.11 < 5.25	3.68 < 4.48 < 6.19

One caveat in interpreting these results is whether the observed improvement for females was due to improved spectral sampling or the interpolation method itself. To address this confound, we extracted features as in f_{vo} = 2-20 but performed interpolation on a set of harmonic peaks derived from a *single* spectral slice (i.e., without merging); this feature set is denoted as f_{vo} = 1-20-*single*. The results of this feature set alone are listed in

Table 14 along with the baseline MFCC results. In addition, we show the results of fusing the baseline with f_{vo} = 1-20-*single* denoted as f_{vo} = 0+1-20-*single*. Observe for the fused result that the performance of males are comparable to those of f_{vo} = 0+1-20 while that of females is 0.30% worse than that of f_{vo} = 0+1-20. While this shows that the gains obtained using f_{vo} = 0+1-20 may partially be due to the interpolation method itself, the larger absolute gain afforded by f_{vo} = 0+1-20 appears promising.

Table 14 - Equal error rates for vowel-only experiment (%) for the baseline MFCC features ($f_{vo} = 0$), and the control method of harmonic projection with $w = 20$ ms (f_{vo} = 1-20-*single*) and their fusion (f_{vo} = 0+1-20-*single*). Confidence intervals are at the 95% level.

	males	females
$f_{vo} = 0$	3.64 < 4.01 < 6.89	4.78 < 6.25 < 8.49
f_{vo} = 1-20-*single*	5.73 < 6.75 < 8.11	6.30 < 7.80 < 9.38
f_{vo} = 0+1-20-*single*	3.51 < 4.10 < 5.06	3.70 < 4.78 < 6.64

6.4 Conclusions

In this chapter, we have employed a 2-D speech processing framework in feature extraction for speaker recognition experiments performed on the TIMIT corpus. Preliminary results appear to implicate the benefits of exploiting temporal change of pitch to derive an improved spectral input to the standard MFCC feature set. Specifically, a closure in an observed gender performance gap was obtained in two separate experiments by employing spectral slice averaging (all-speech) and the harmonic projection method (vowel-only). It remains unclear why spectral slice averaging alone in the all-speech experiments provided a gain over the baseline MFCC feature set whereas harmonic projection in the vowel-only experiments afforded gains only when fused with the MFCC baseline. Several discrepancies between the all-speech and vowel-only experiments including short-time analysis window length (fixed 10 ms vs. pitch-adaptive), feature stream frame rate (2-ms vs. 10-ms), and training data (gender imbalanced vs. gender balanced) may have caused this difference in performance. In addition, recall that while the UBM was derived from a gender balanced data set, the number of male and female target speakers differed between the all-speech and vowel-only experiments. More rigorous testing is therefore needed in future work to fully assess the value of these proposed methods.

Chapter 7

Conclusions and Future Work

7.1 Summary

In this thesis, we have proposed the use of a 2-D speech processing framework for addressing high-pitch formant estimation. This framework is inspired by both psychophysical and physiological evidence implicating the auditory system's use of *temporal changes* in speech. We began in Chapter 2 by discussing a short-time model of speech production for vowel sounds and the effects of high-pitch on standard formant estimation methods. We showed that these effects had several analogous interpretations illustrating the lack of robustness of these standard techniques. One interpretation in particular is that of spectral undersampling, which we adopted for the entirety of this thesis.

In Chapter 3, we provided a thorough phenomenological analysis of three realizations of a 2-D speech processing framework: (pitch) harmonic projection, the Grating Compression Transform (GCT) [9], and a comprehensive model of auditory signal processing ending in the auditory cortex [8]. Our aim in this analysis was to motivate an improved method of high-pitch formant estimation when pitch is changing the formant envelope is assumed to be stationary. Under these conditions, we observed that harmonic projection improves the spectral sampling of the formant envelope while the GCT invokes improved source-filter separability in a 2-D modulation space. These observations were also justified analytically. While the auditory model invokes a similar 2-D modulation space and highlights formant structure and pitch in several ways, we were unable to argue for the same type of source-filter separability as in the GCT. We concluded in Chapter 3 by motivating several methods of deriving speech spectra for improving high-pitch formant estimation.

Chapter 4 described a set of formant estimation methods motivated from Chapter 3 an experimental framework for evaluating them on synthesized speech in relation to standard techniques. In Chapter 5, we presented the results of this evaluation; these results suggested that exploiting temporal change of pitch can provide improved estimates of formant frequencies, even under conditions of high pitch. Finally, Chapter 6 presented preliminary results of adopting our 2-D processing framework for the speaker recognition application. Our results appeared promising in addressing a performance "gender gap" in recognition performance though further testing is required.

7.2 Future Work

Motivated from our preliminary results in speaker recognition, future work will aim to further assess the value of applying the 2-D processing framework in addressing the performance gender gap. In addition to the harmonic projection and spectral slice averaging methods, the GCT may also be used as a basis for feature extraction. For an all-speech experiment, one modification to the filtering implementation used in formant estimation may be to adapt *both* the rate and scale

filter cut-offs as opposed to just the scale direction (Section 4.2.2.). This could account for formant and syllabic transitions occurring in dipthongs and running speech. In addition, we aim to revisit the all-speech and vowel-only experiments to better understand the previously noted performance discrepancy. Finally, the improved estimates of high-pitch formants presented in Chapter 5 motivate feature sets derived from improved inverse filtering to better exploit glottal source characteristics in the speaker recognition application.

As previously discussed, the generalized 2-D framework employs *any* time-frequency distribution followed by *any* 2-D transform. An area of future work could be to explore time-frequency distributions other than the STFT such as the Fan-chirp transform [33] as a basis for the GCT as it has been shown to enhance harmonic structure. In addition, recall from Section 3.2 that the GCT approximates harmonic line structure as being parallel under conditions of changing pitch. Consequently, we observed in Figure 36 that estimates of the spatial frequency of harmonic lines ($\hat{\omega}_0$) increased for higher frequency regions due to the increased fanning of harmonic lines. Although we exploited this property in formant estimation using the GCT[13], this observation highlights the limits of the approximation made in the GCT. In Appendix F, we provide an analytical investigation of this effect to motivate a novel 2-D transform that coherently maps a *fanned* line structure to an alternate 2-D space. Finally, further investigation of the potential benefits of the auditory representation proposed by Chi, et al. [8] is also necessary.

The proposed model for source-filter separability in Section 3.5 may also have implications for the two-speaker separation problem. In [10] and [9], Quatieri and Ezzat, et al. observed that the STFT magnitude of two simultaneous speakers can sometimes resemble the sum of their individual STFT magnitudes. Denoting one speaker's contribution as $s_1[n,m]$ and the other as $s_2[n,m]$ and invoking the model proposed in Section 3.5, their sum $s[n,m]$ can be expressed analytically as

$$\begin{aligned} s_1[n,m] &= a_1[n,m] + a_1[n,m]\cos(\hat{\omega}_1\Phi_1(n,m;\hat{\theta}_1)) \\ s_2[n,m] &= a_2[n,m] + a_2[n,m]\cos(\hat{\omega}_2\Phi_2(n,m;\hat{\theta}_2)) \\ s[n,m] &= s_1[n,m] + s_2[n,m] \\ &= a_1[n,m] + a_1[n,m]\cos(\hat{\omega}_1\Phi_1(n,m;\hat{\theta}_1)) + a_2[n,m] + a_2[n,m]\cos(\hat{\omega}_2\Phi_2(n,m;\hat{\theta}_2)) \end{aligned} \tag{7.1}$$

Observe that while the slowly-varying components $a_1[n,m]$ and $a_2[n,m]$ will likely overlap near the rate-scale origin, their modulated versions can be separated using 2-D filtering based on their local spatial frequencies $\hat{\omega}_1$ and $\hat{\omega}_2$ and/or their orientations described by $\hat{\theta}_1$ and $\hat{\theta}_2$. These parameters may be estimated as was done in our formant estimation method employing the GCT and used to separately *demodulate* $a_1[n,m]\cos(\hat{\omega}_1\Phi_1(n,m;\hat{\theta}_1))$ and $a_2[n,m]\cos(\hat{\omega}_2\Phi_2(n,m;\hat{\theta}_2))$ such that $a_1[n,m]$ and $a_2[n,m]$ can be recovered, thereby motivating a simple method of speaker separation.

One caveat of this approach is that the STFT magnitude is not generally linear. An alternative short-time analysis method such as a short-time discrete-cosine transform (which is linear) may

[13] Recall that an increasingly permissive filter along the scale axis was generated as the frequency region of analysis increased, presumably preventing oversmoothing.

be employed instead; further work is needed to assess whether the proposed model of Section 3.5, is well suited to this time-frequency distribution such that the described approach to speaker separation is feasible.

Appendix A

Autocorrelation method of linear prediction

In this thesis, the autocorrelation method of linear prediction has been used for formant estimation when applied to spectral estimates derived from a number of methods. This section provides the derivation of this method for the source-filter model of speech production as well as its interpretations in the time and spectral domains.

A.1 Normal equations

Define the linear prediction error of a short-time segment of speech $x_n[m]$ (via the all-pole model) as

$$e_n[m] = x_n[m] - \sum_{k=1}^{p} \alpha_k x_n[m-k] \qquad 0 \le m \le N+p-1 \tag{0.1}$$

where N corresponds to the length of the window. To determine α_k, the sum of squared errors $E_n = \sum_{m=0}^{N+p-1} e_n^2[m]$ is minimized:

$$\begin{aligned} E_n &= \sum_{m=0}^{N+p-1} (x_n[m] - \sum_{k=1}^{p} \alpha_k x_n[m-k])^2 \; so\; that \\ \frac{\partial E_n}{\partial \alpha_i} &= 2 \sum_{m=0}^{N+p-1} \left[(x_n[m] - \sum_{k=1}^{p} \alpha_k x_n[m-k]) \frac{\partial}{\partial \alpha_i} (x_n[m] - \sum_{k=1}^{p} \alpha_k x_n[m-k]) \right]. \end{aligned} \tag{0.2}$$

Rearranging (0.2) and setting $\frac{\partial E_n}{\partial \alpha_i}$ equal to zero, we obtain

$$\begin{aligned} 0 &= 2 \sum_{m=0}^{N+p-1} \left[(x_n[m] - \sum_{k=1}^{p} \alpha_k x_n[m-k])(-x_n[m-i]) \right] \\ \sum_{m=0}^{N+p-1} x_n[m] x_n[m-i] &= \sum_{k=1}^{p} \alpha_k \sum_{m=0}^{N+p-1} x_n[m-k] x_n[m-i] \end{aligned} \tag{0.3}$$

which can be rewritten as

$$r_n[i] = \sum_{k=1}^{p} \alpha_k r_n[k-i] \quad 1 \le i \le p$$
$$r_n[i] = x_n[n] * x_n[-n] \tag{0.4}$$

where $r_n[i]$ denotes the autocorrelation of the short-time segment analyzed.

A.2 Time-domain interpretation

Consider the impulse response of the all-pole model of Section 2.3 with $A = 1$ for simplicity

$$h[n] = \sum_{k=1}^{p} \alpha_k h[n-k] + \delta[n]. \tag{0.5}$$

For $h[n]$ causal, multiplying both sides of (0.5) by $h[n-i]$ for $1 \le i \le p$ and summating across all n yields [3]:

$$\sum_{n=-\infty}^{\infty} h[n]h[n-i] = \sum_{n=-\infty}^{\infty} \left(\sum_{k=1}^{p} \alpha_k h[n-k]h[n-i] + \delta[n]h[n-i] \right). \tag{0.6}$$

Since $h[n]$ is causal and $i \ge 0$, $\delta[n]h[n-i] = h[-i] = 0$. The left-hand side of (0.6) is the autocorrelation of $h[n]$ denoted by $r_h[i]$. Similarly, for the term $\sum_{n=-\infty}^{\infty} \sum_{k=1}^{p} \alpha_k h[n-k]h[n-i]$, letting $m = n - i$ yields $\sum_{k=1}^{p} \alpha_k \sum_{n=-\infty}^{\infty} h[m]h[m-(k-i)] = \sum_{k=1}^{p} \alpha_k r_h[k-i]$. In this case, $r_h[n]$ corresponds to the true autocorrelation of the all-pole model such that the normal equations

$$r_h[i] = \sum_{k=1}^{p} \alpha_k r_h[k-i] \quad 1 \le i \le p \tag{0.7}$$

will solve for the exact values of α_k [3].

In contrast to the single-impulse input case, voiced speech is typically modeled as the result of a *periodic* impulse train input to the all-pole system (here, a pure impulse train) such that

$$s[n] = \sum_{k=-\infty}^{\infty} h[n-kP] \tag{0.8}$$

as was observed in Figure 1. The autocorrelation estimate required for the normal equations is

$$
\begin{aligned}
r_n[i] &= \sum_{k=-\infty}^{\infty} s[k]s_1[-(i-k)] = \sum_{k=-\infty}^{\infty} s[k]s_1[k-i] \\
&\sum_{k=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} h[k-mP]\sum_{l=-\infty}^{\infty} h[k-i-lP] \\
&\sum_{m=-\infty}^{\infty}\sum_{l=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} h[k-mP]h[k-i-lP] \\
&\sum_{m=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} r_h[(l-m)P+i] \\
&\sum_{u=-\infty}^{\infty} r_h[i-uP]
\end{aligned}
\tag{0.9}
$$

where we have made the substitution $u = m - l$ in the final step. In practice, m and l will range over finite values depending on the length of the window such that the resultant autocorrelation estimate will have decreasing amplitude for large values of i [3].

A.3 Frequency-domain interpretation

From Section 2.3, the all-pole model frequency response of the vocal tract $H(\omega)$ is

$$
\begin{aligned}
H(\omega) &= \frac{A}{A(\omega)} \\
A(\omega) &= 1 - \sum_{k=1}^{p} \alpha_k e^{-j\omega k}
\end{aligned}
\tag{0.10}
$$

Define the spectral-domain representation of the prediction error as

$$
E(\omega) = X(\omega)A(\omega) \tag{0.11}
$$

$Q(\omega)$ as defined in Section 2.3.1 can then be rewritten as

$$
\begin{aligned}
Q(\omega) &= \log|X(\omega)|^2 - \log|H(\omega)|^2 = \log\left|\frac{X(\omega)}{H(\omega)}\right|^2 \\
&= \log\left|\frac{E(\omega)}{A}\right|^2
\end{aligned}
\tag{0.12}
$$

Substituting (0.12) into the error criterion I of Equation (2.12),

$$I = \int_{-\pi}^{\pi} \left[\left| \frac{X(\omega)A(\omega)}{A} \right|^2 - \log \left| \frac{X(\omega)A(\omega)}{A} \right|^2 - 1 \right] \frac{d\omega}{2\pi}$$
$$= \frac{1}{A^2} \int_{-\pi}^{\pi} |X(\omega)A(\omega)|^2 \frac{d\omega}{2\pi} + \log A^2 + \int_{-\pi}^{\pi} \log |A(\omega)|^2 \frac{d\omega}{2\pi} + \int_{-\pi}^{\pi} |X(\omega)|^2 \frac{d\omega}{2\pi} - 1 \quad . \quad (0.13)$$

Under the assumption that the roots of $A(\omega)$ have magnitude less than unity, the term $\int_{-\pi}^{\pi} \log |A(\omega)|^2 \frac{d\omega}{2\pi}$ can be shown to equal zero:

$$|A(\omega)|^2 = A(\omega)A^*(\omega)$$
$$\int_{-\pi}^{\pi} \log |A(\omega)|^2 \frac{d\omega}{2\pi} = \int_{-\pi}^{\pi} \log A(\omega) \frac{d\omega}{2\pi} + \int_{-\pi}^{\pi} \log A^*(\omega) \frac{d\omega}{2\pi} \quad (0.14)$$
$$= c_{A(\omega)}[n] + c_{A^*(\omega)}[n]\Big|_{n=0}$$

where $c_{A(\omega)}[n]$ and $c_{A^*(\omega)}[n]$ correspond to the complex cepstra of $A(\omega)$ and $A^*(\omega)$, respectively. Since $A(\omega)$ has roots with magnitude less than unity, $A^*(\omega)$ will also have roots with magnitude less than unity. At $n = 0$, $c_{A(\omega)}[n]$ and $c_{A^*(\omega)}[n]$ will then both be equal to $\log(1) = 0$ from properties of the complex cepstrum [4]. Equation (0.13) then becomes

$$I = \frac{1}{A^2} \int_{-\pi}^{\pi} |E(\omega)|^2 \frac{d\omega}{2\pi} + \log A^2 + \int_{-\pi}^{\pi} |X(\omega)|^2 \frac{d\omega}{2\pi} - 1 . \quad (0.15)$$

Note that the dependence on the set of α_k in (0.15) is exclusively in $\frac{1}{A^2} \int_{-\pi}^{\pi} |E(\omega)|^2 \frac{d\omega}{2\pi}$ such that minimization of I via differentiation with respect to the α_k's is equivalent to minimizing $\int_{-\pi}^{\pi} |E(\omega)|^2 \frac{d\omega}{2\pi} = \sum_{n=0}^{N-1} |e[n]|^2$ as in the time domain (Section A.1).

Appendix B

Cepstrum of a windowed impulse train

From Section 2.3.2 we define $g_w[n] = w[n]g[n]$ with $g[n] = \sum_{k=-\infty}^{\infty} \delta[n-kP]$ such that

$$\begin{aligned} g_w[n] &= w[n]\sum_{k=-\infty}^{\infty}\delta[n-kP] \\ &= \sum_{k=0}^{N-1} w[kP]\delta[n-kP] \end{aligned} \tag{0.16}$$

Defining an alternate sequence [4]

$$\begin{aligned} w_P[k] &= w[kP] \quad k = 0,1...L-1 \\ &= 0 \qquad otherwise \end{aligned} \tag{0.17}$$

$g_w[n]$ is equivalent to $w_P[k]$ *upsampled* by a factor of P

$$\begin{aligned} g_w[n] &= w_P[n/P] \qquad n = 0,\pm P,\pm 2P... \\ &= 0 \qquad\qquad otherwise \end{aligned} \tag{0.18}$$

The Fourier transform of $g_w[n]$ is then $G_w(\omega) = W_p(\omega P)$. Let $c_w[n]$ denote the cepstrum of $w_P[k]$. The cepstrum of $g_w[n]$, denoted by $c_g[n]$ is then $w_P[k]$ upsampled by a factor P:

$$\begin{aligned} &W_p(\omega) = \sum_{k=0}^{L-1} w[kP]e^{-j\omega k} \quad so\ that \\ &c_w[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log[W_p(\omega)]e^{j\omega n}d\omega \\ &c_g[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log[W_p(\omega P)]e^{j\omega n}d\omega \end{aligned} \tag{0.19}$$

Therefore,

$$\begin{aligned} c_g[n] &= c_w[n/P] \qquad n = 0,\pm P,\pm 2P,... \\ &= 0 \qquad\qquad otherwise \end{aligned} \tag{0.20}$$

Appendix C

Pitch effects in formant frequency estimation using linear prediction

As previously discussed (Section 2.2), the narrow-band speech spectrum for vowels has energy concentrated at the pitch harmonics. Consequently, certain pitch values can "fortuitously" sample the formant envelope (denoted as $|H(\omega)|$) near its peaks (even under high-pitch conditions) such that an accurate estimate can be expected via the spectral matching condition in linear prediction [26]. Consider again the female vowel */ae/* with F1, F2, and F3 set to 860, 2050, and 2860 Hz, respectively, and B1, B2, and B3 set to 54, 65, and 70 Hz, respectively. Figure 61 shows the raw formant errors for the traditional linear prediction baseline (m = 1) with a 25-Hz pitch shift as in Section 0. Recall that in this method, a spectral slice extracted from the middle of the utterance is used for linear prediction. For a starting pitch of $f_{0s} = 205$ Hz and a pitch shift of 25 Hz, the synthesized vowel will exhibit a pitch of 217.5 Hz in the middle of the utterance. We denote this center pitch value in general as f_{center} which can be obtained from the downsampled pitch contour described in Section 4.2.1. The fourth harmonic of this pitch value is near F1 = 860 Hz: $\frac{860}{217.5} \approx 4$. We may then expect to obtain an accurate estimate of F1 due to this "fortuitous sampling".

Observe in Figure 57 (bottom) that the raw error indeed exhibits a minimum at a pitch start of 205 Hz (with $f_{center} = 217.5$ Hz). To further assess the validity of this "fortuitous sampling" explanation of local minima, we plot in Figure 57 (top) the product $\breve{F}_i(f_{0s})$:

$$\begin{aligned} \breve{F}_i(f_{0s}) &= nf_{center} \\ n &= round(\frac{F_i}{f_{center}}) \end{aligned} \tag{0.21}$$

with $df_0 = 25$ Hz for i = 1. In Figure 57 (bottom) we highlight points on the raw formant error curve that invoke a f_{center} such that the n^{th} harmonic is "near" the i^{th} formant. For example, a pitch start of 160 Hz invokes an f_{center} of 172.5 Hz. The 5th harmonic (n = 5) of this pitch value is near the first formant (i.e., $\frac{860}{172.5} \approx 5$, $\breve{F}_i(f_{0s}) = 5 \cdot 172.5 = 862.5$ Hz). An additional constraint imposed on these points is that $|H(\omega = \breve{F}_i(f_{0s}))|$ is within 1 dB of $|H(\omega = F_i)|$. We show in Figure 60 the region near the formant peaks where such pitch harmonics may lie (*). Observe with this criterion that two local minima have been highlighted at $f_{0s} = 160, 205$ Hz,

consistent with the "fortuitous sampling" interpretation. Nonetheless, $f_{0s} = 275$ Hz is also highlighted and is not strictly a local minimum. Figure 58 and Figure 59 further illustrate this loss of generality for F2 and F3, with highlighted points sometimes corresponding to local *maxima*.

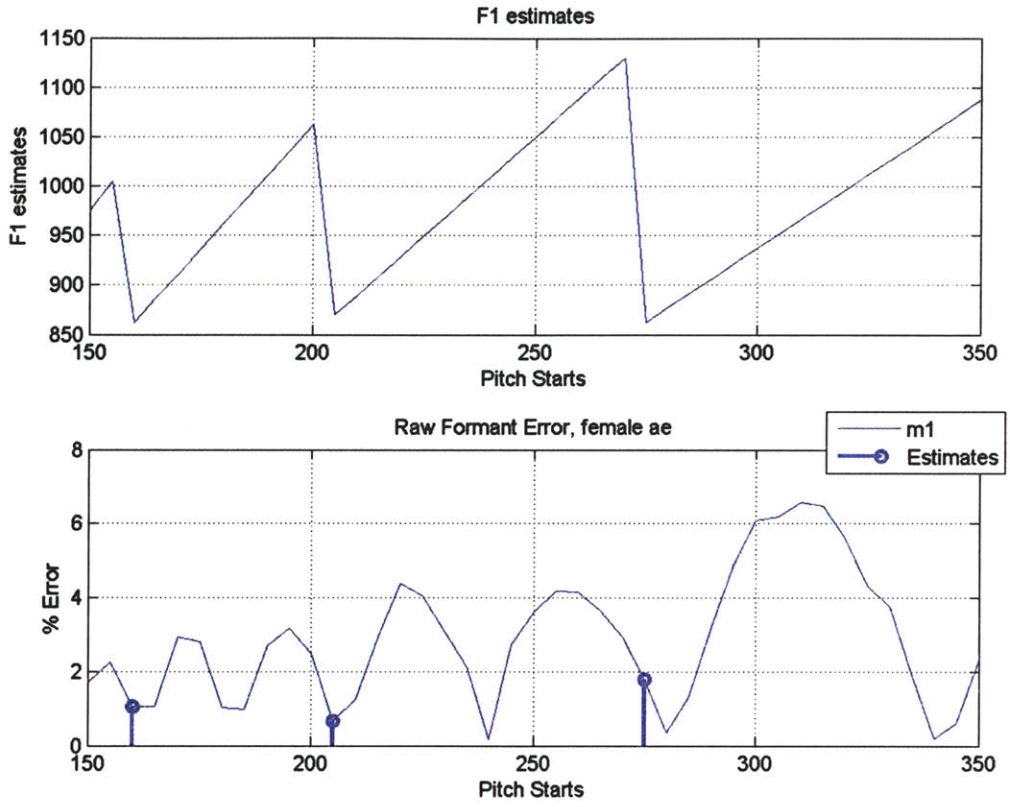

Figure 57 – F1 estimates based on nearest harmonic number (top) as a function of f_{0s}; Raw formant errors for m = 1 (solid curve) and the set of corresponding f_{0s} values (stem plot) invoking f_{center} pitch harmonics that are considered "near" the F1 peak (bottom).

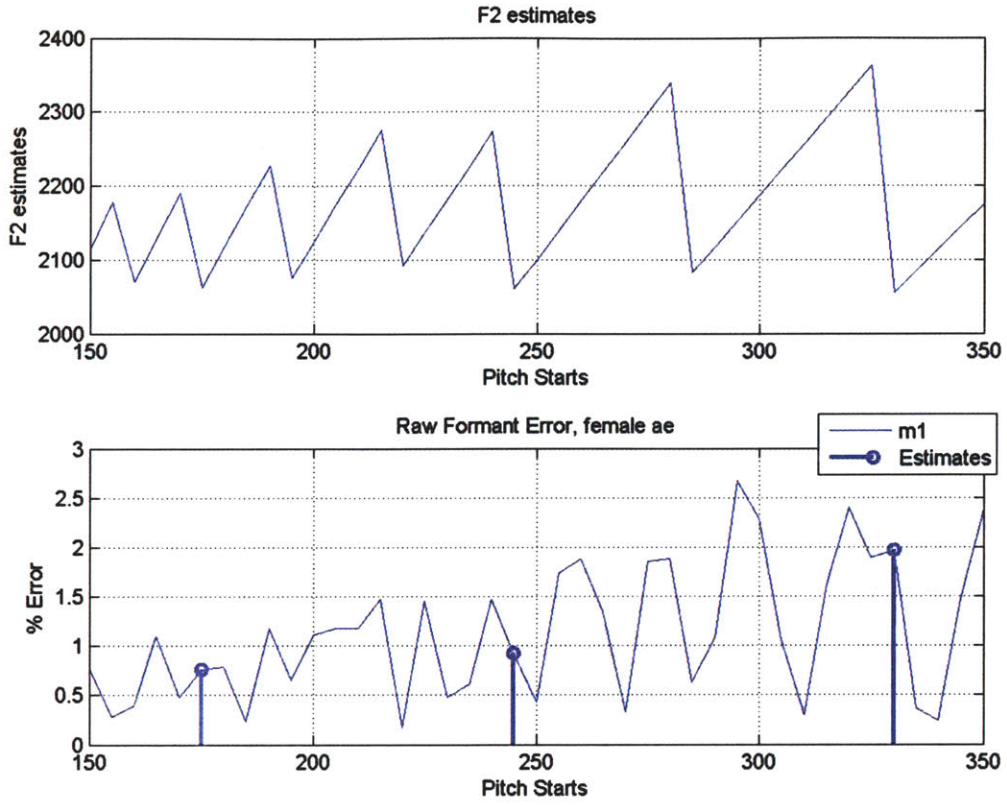

Figure 58 – F2 estimates based on nearest harmonic number (top) as a function of f_{0s}; Raw formant errors for m = 1 (solid curve) and the set of corresponding f_{0s} values (stem plot) invoking f_{center} pitch harmonics that are considered "near" the F2 peak.

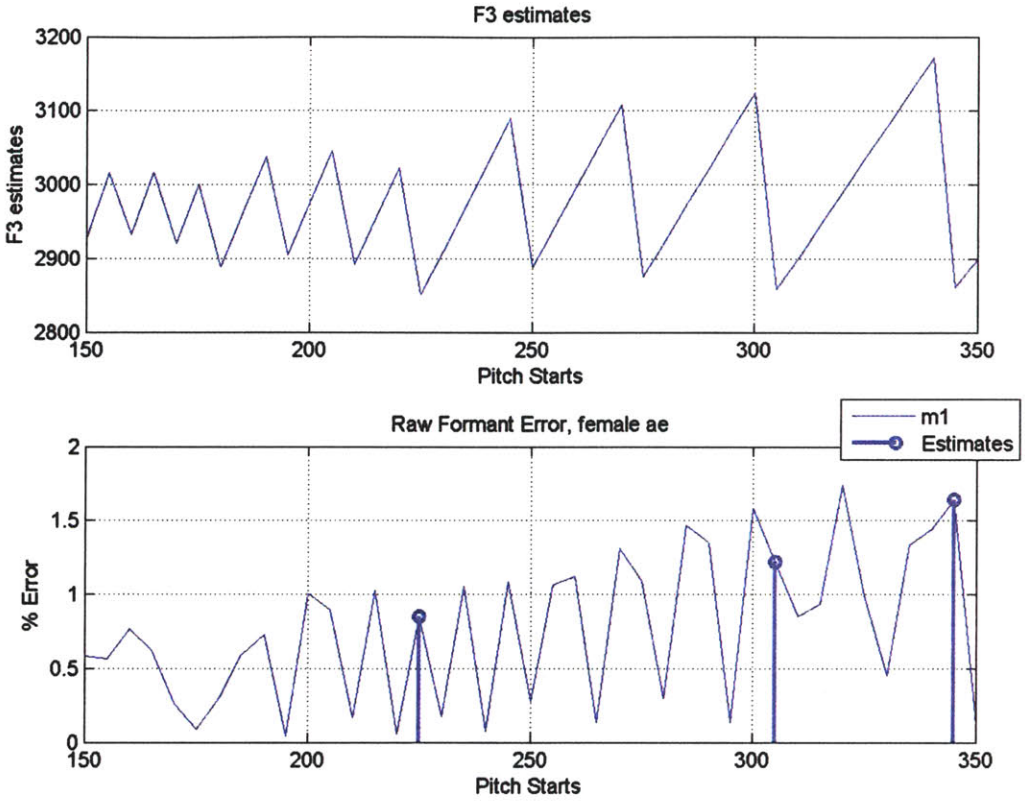

Figure 59 – F3 estimates based on nearest harmonic number (top) as a function of f_{0s}; Raw formant errors for m = 1 (solid curve) and the set of corresponding f_{0s} values (stem plot) invoking f_{center} pitch harmonics that are considered "near" the F3 peak.

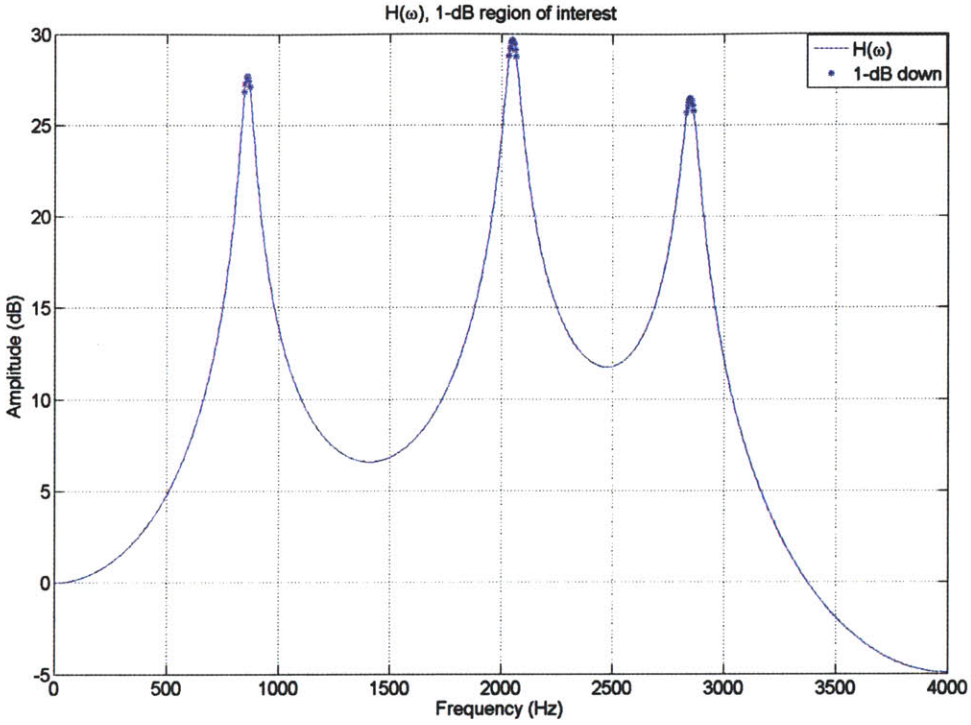

Figure 60 – True formant envelope for female /*ae*/ with points 1 dB down from the formant peak denoted by asterisks. Points are chosen from the set of $\breve{F}_i(f_{0s}) = nf_{center}$ derived from Equation (0.21) that fall in the regions highlighted by this amplitude criterion.

The observed loss of generality of the "fortuitous sampling" explanation of local minima in the raw formant error curves can be explained by noting that via linear prediction, formant frequency estimates are not obtained independently as is assumed by this explanation. We plot the error curves for all three formants in Figure 61 to illustrate the interactions between formant estimation accuracy. Observe that at $f_{0s} = 225$ Hz (arrow), all three formant errors exhibit local maxima. This may be expected for F1 (860 Hz) and F2 (2050 Hz) since the resulting $f_{center} = 237.5$ Hz would not invoke a harmonic near these formants ($\frac{860}{237.5} = 3.62$, and $\frac{2050}{237.5} = 8.63$). However, this is unexpected for F3 since 272.5 Hz invokes a 12th harmonic that corresponds to the formant frequency (F3 = 2850 Hz, $\frac{2850}{272.5} = 12$). In conclusion, while the "fortuitous sampling" explanation of oscillations in formant estimation errors can account for some of the observed effects, oscillations are also due in part to the interactions *between* individual formant estimates (e.g., F1 vs. F2) in linear prediction.

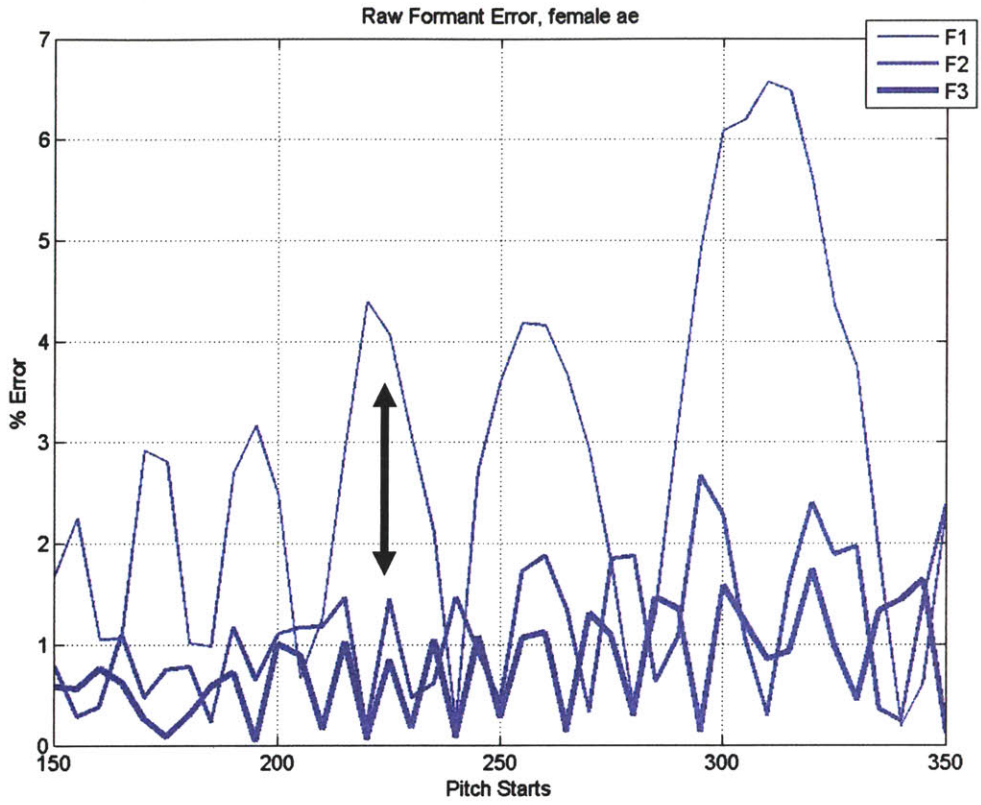

Figure 61 – All formant errors for m = 1 as a function of f_{0s}; Arrow indicates a pitch start of 225 Hz referred to in the discussion.

Appendix D

Male and children results for averaging across pitch starts and pitch shifts

As in Section 5.4, we present in this section the average of the raw formant errors across pitch starts and pitch shifts for males and children, thereby showing average errors across vowels. For reference, we show in Table 15 the formant estimation methods used. The abscissa refers to vowels */ah/* (1), */iy/* (2), */ey/* (3), */ae/* (4), */oh/* (5) for males, */oo/* (6), and */uh/* (5) for children.

Table 15 – Summary of formant estimation methods

m=1	Traditional linear prediction (LP)
m=2	Homomorphic linear prediction (HLP)
m=3	Interpolation of collected harmonic peaks + LP
m=4	Interpolation using single slice + LP
m=5	Time-average of $STFT_1$ + LP
m=6	GCT-based filtering on $STFT_1$ + LP

Taken together, these results are generally consistent with the global average metrics presented in Section 5.4. For instance, observe that for males, m = 3 outperforms all other methods for all conditions. In addition, m = 4 consistently performs worse than m = 3, thereby highlighting the benefits of exploiting temporal change of pitch. Observe also that m = 5 and 6 consistently outperform m = 1 and 2 for F2 and F3 but not F1. For children, similar results were obtained for m = 3, 5, and 6 across vowels and formants; nonetheless, two exceptions were observed. Specifically, the F2 estimate of */oo/* using m = 3 has worse performance than either baseline. Likewise, the F2 estimate of */ah/* using m = 5 is worse than the m = 2 baseline. Finally, observe for males that m = 5 tends to outperform m = 6 as suggested by the global average metric presented in Section 5.4. It is unclear why these effects occur, and further experimentation with the GCT and harmonic projection methods are necessary.

It is of interest here to recall that in Section 5.4, we interpreted the poorer performance of F1 using m = 5 and 6 (in females) as reflecting the reduced source-filter separability of the GCT in low-frequency regions of the STFT (from reduced fanning of the harmonic line structure when pitch is changing). Recall from Figure 50 that m = 5 and m = 6 were able to provide gains over m = 1 and 2 for all vowels except */iy/* and */oo/*. Here, we observe that this is also the case for both males and children. From Table 3 of Section 4.1, note that the F1 frequencies for these two vowels are the *lowest* relative to others for the same gender, which is consistent with this interpretation.

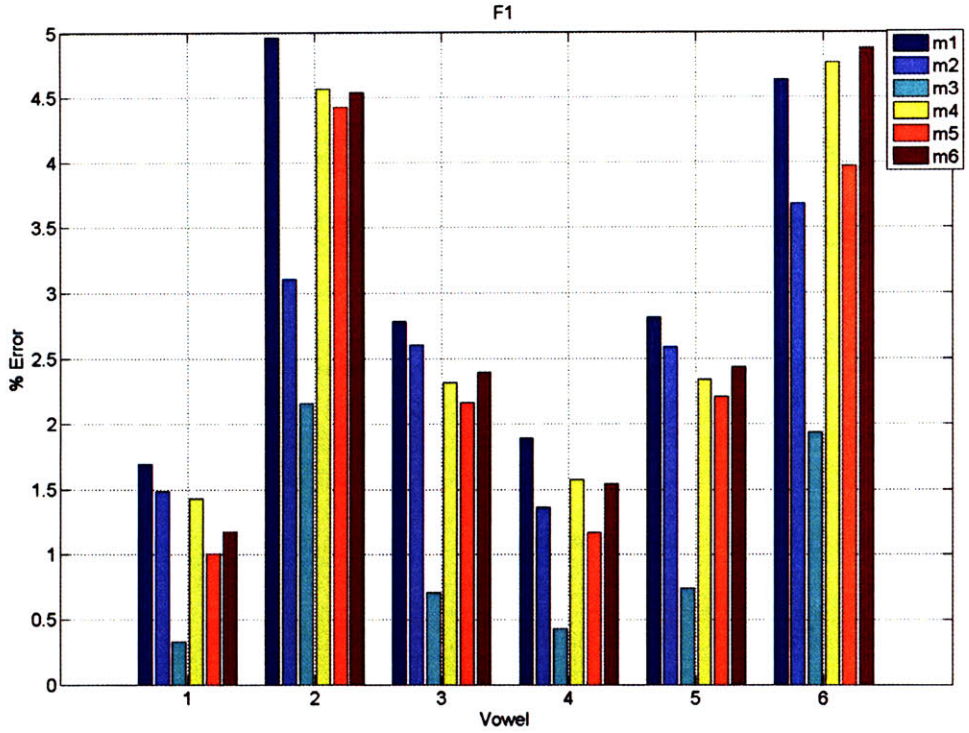

Figure 62 – Averages across pitch starts and pitch shifts, $E_i(v)$, for $i = 1$ (F1, males).

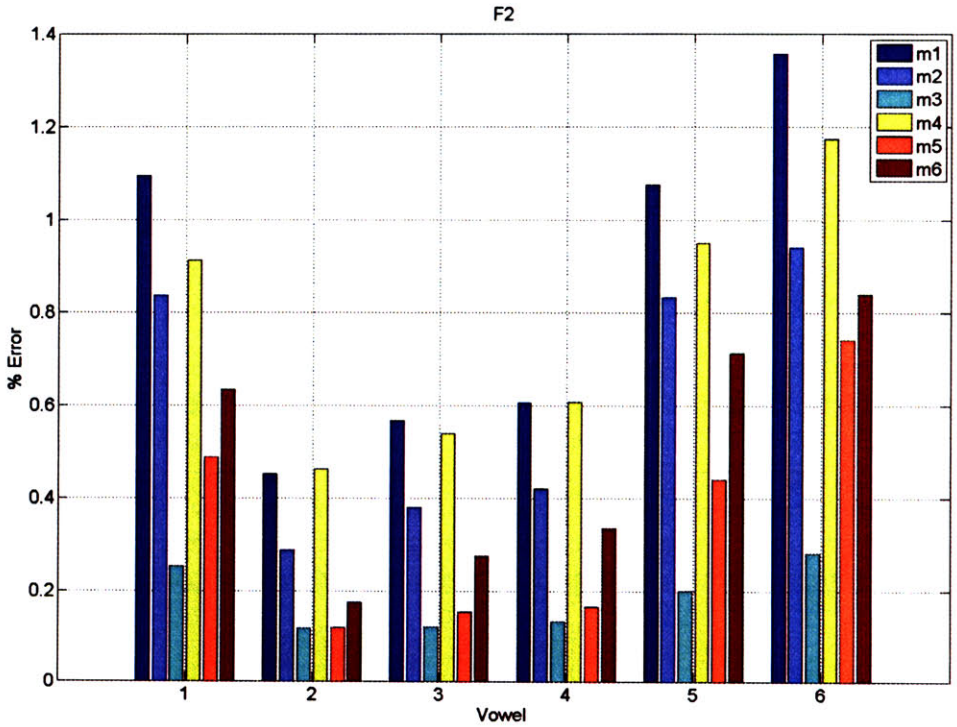

Figure 63 – Averages across pitch starts and pitch shifts, $E_i(v)$, for $i = 2$ (F2, males).

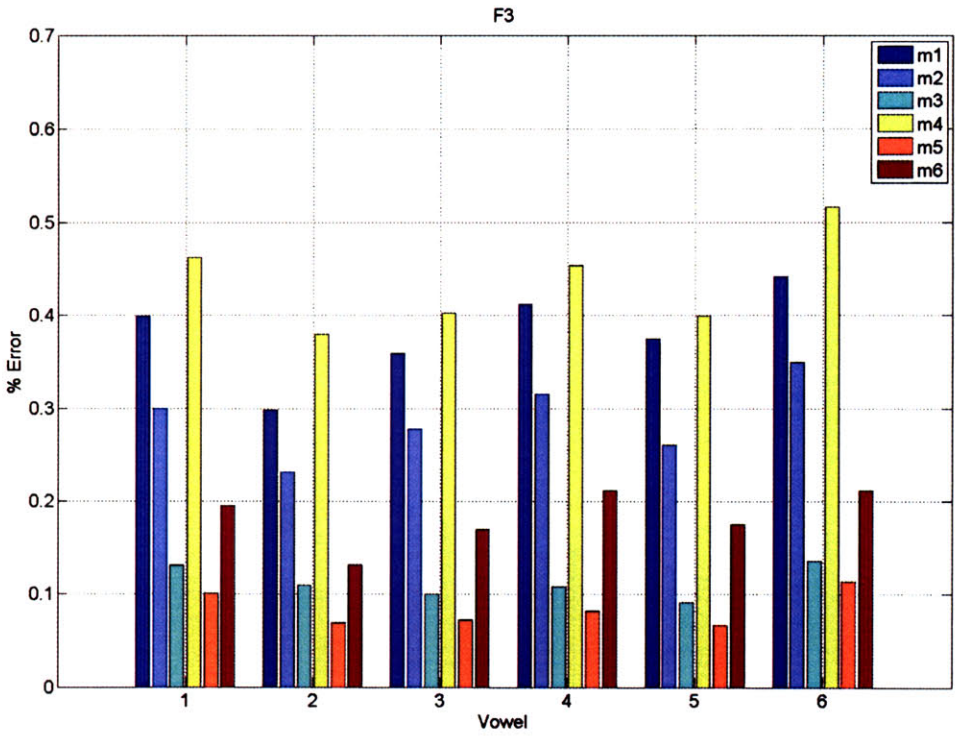

Figure 64 – Averages across pitch starts and pitch shifts, $E_i(v)$, for $i = 3$ (F3, males).

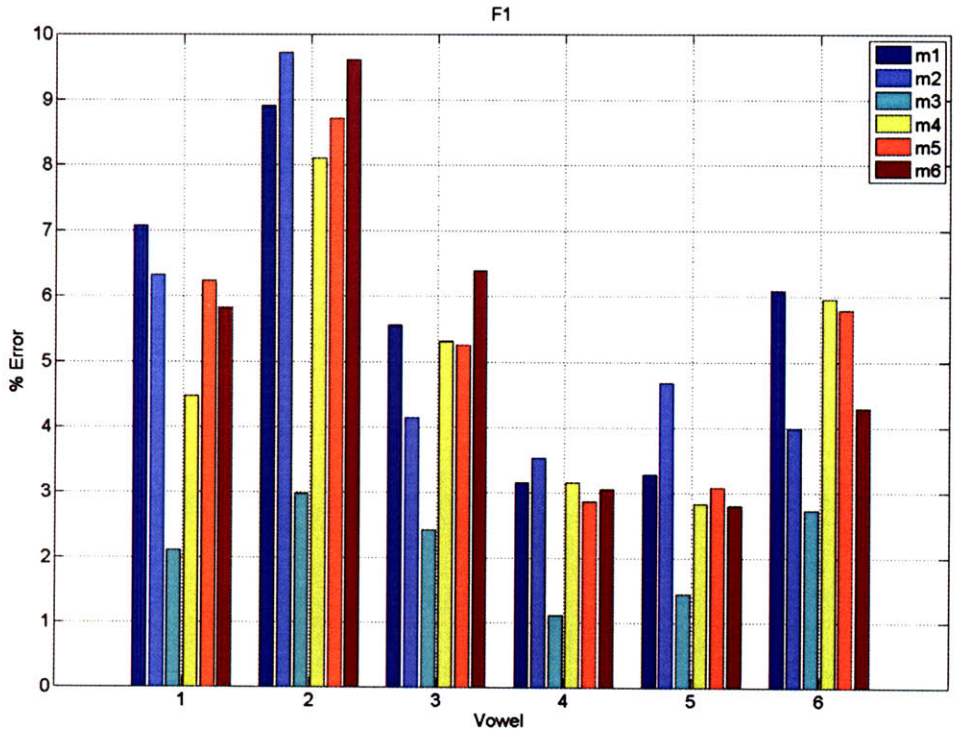

Figure 65 - Averages across pitch starts and pitch shifts, $E_i(v)$, for $i = 1$ (F1, children).

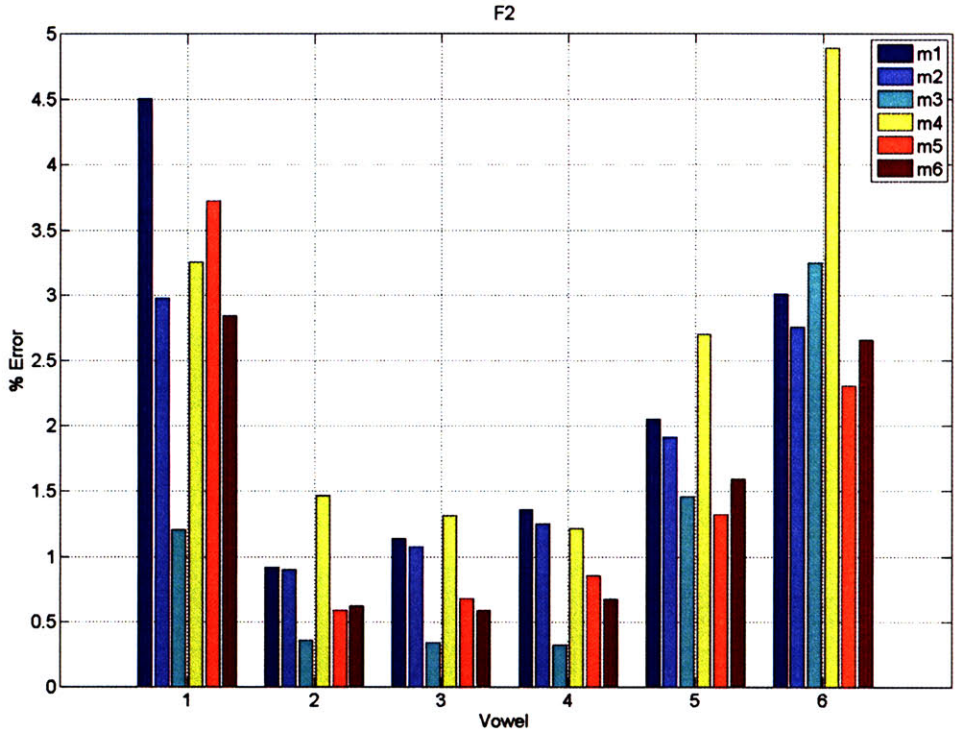

Figure 66 – Averages across pitch starts and pitch shifts, $E_i(v)$, for $i = 2$ (F2, children).

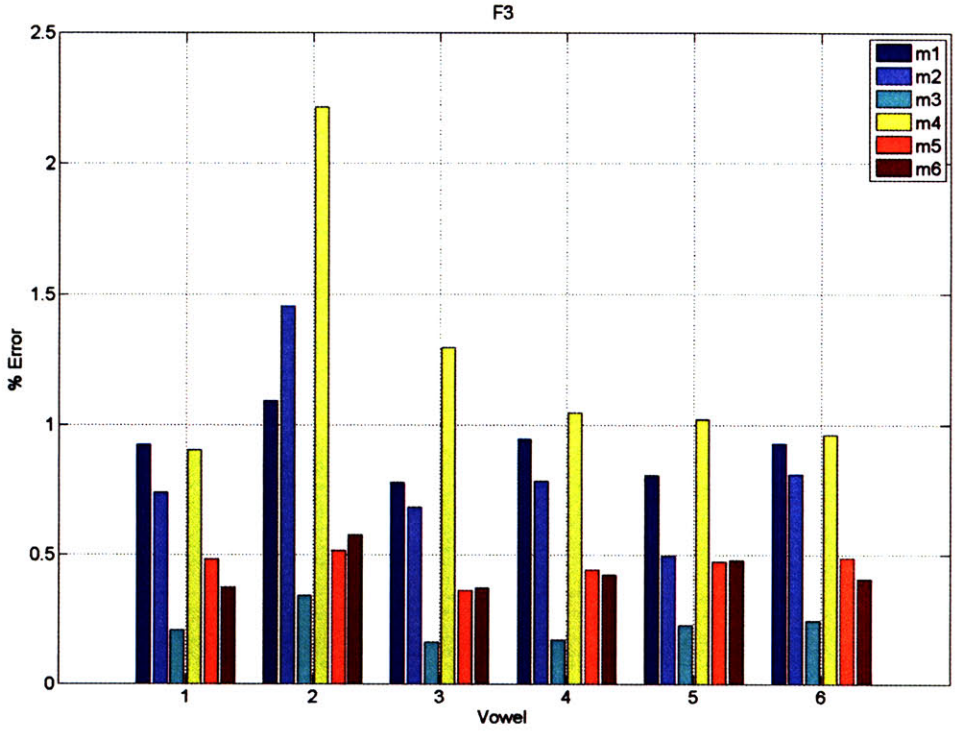

Figure 67 – Averages across pitch starts and pitch shifts, $E_i(v)$, for $i = 2$ (F2, children).

Appendix E

Results of baseline formant estimation methods using pitch-adaptive short-time analysis

In this section, we present results of using traditional and homomorphic linear prediction with pitch-adaptive short-time analysis. For both methods, a spectral slice used was extracted from the middle of the pitch-adaptive $STFT_1$ (Section 4.2.1) instead of $STFT_0$ (not pitch-adaptive). We denote these methods as m = 1p and m = 2p. Our aim is to control for the possibility that pitch-adaptive short-time analysis alone invoked the gains we observed using the methods of harmonic projection and filtering in the GCT. Table 16 and Table 17 show the global average metric (across vowels, pitch starts, and pitch shifts, Section 5.4) m = 1p and m = 1 (as in Section 5.4), respectively. Table 18 and Table 19 show the results of the global average metric for m = 2p and m = 2 (also from Section 5.4), respectively. In Table 20 and Table 21, we show the relative gains of those methods employing the pitch-adaptive spectrogram with respect to those employing the fixed-window spectrogram. We observe that the performance is similar for traditional linear prediction using either a pitch-adaptive or fixed-length window for short-time analysis. Specifically, the maximum relative performance of m = 1p to m = 1 is a 5.07% gain. In contrast, employing the pitch-adaptive window for homomorphic linear prediction appears to *degrade* performance overall relative to a fixed-length window. Relative losses up to 21% (i.e., females, F3) are observed; however, for two conditions, F1 of females and F1 of children, the m = 2p exhibits relatives of 8.85% and 20.44%. Further investigation is necessary to better understand these effects. Nonetheless, our results overall suggest that the gains obtained for the harmonic projection and GCT methods were not due to the pitch-adaptive short-time analysis itself, and are most likely due to the effects of exploiting temporal change of pitch.

Table 16 – Global average metric[14] across vowels, pitch starts and pitch shifts for m = 1p.

	males	females	children
i = 1	3.15	4.85	5.66
i = 2	0.83	1.61	2.05
i = 3	0.37	0.79	0.94

Table 17 – Global average metric across vowels, pitch starts and pitch shifts for m = 1.

	males	females	children
i = 1	3.13	4.88	5.68
i = 2	0.86	1.66	2.16
i = 3	0.38	0.77	0.91

[14] Results have been rounded to two significant digits for display but were not rounded in computing relative gains.

Table 18 – Global average metric across vowels, pitch starts and pitch shifts for m = 2p.

	males	females	children
i = 1	2.63	3.69	4.30
i = 2	0.65	1.46	2.10
i = 3	0.29	0.65	0.86

Table 19 – Global average metric across vowels, pitch starts and pitch shifts for m = 2.

	males	females	children
i = 1	2.47	4.05	5.40
i = 2	0.62	1.30	1.81
i = 3	0.29	0.54	0.83

Table 20 – Relative gains (%) of m = 1p with respect to m = 1.

	males	females	children
i = 1	-0.57	0.60	0.23
i = 2	2.78	3.20	5.07
i = 3	2.74	-2.93	-3.22

Table 21 – Relative gains (%) of m = 2p with respect to m = 2.

	males	females	children
i = 1	-6.43	8.85	20.44
i = 2	-5.25	-13.00	-15.67
i = 3	-0.80	-21.24	-3.27

Appendix F

Coherent mapping of a fanned-line structure

We observed in Section 4.2.2 that under conditions of changing pitch, the STFT is characterized by a fanned line structure. This characteristic leads to an increase in local estimates of the spatial frequency ($\hat{\omega}_0$) corresponding to an assumed *parallel* harmonic line structure in the GCT (Figure 36). Herein we provide an analytical investigation of the effects of a linearly changing pitch on the STFT, thereby motivating a novel 2-D transform that is able to coherently transform a fanned line structure.

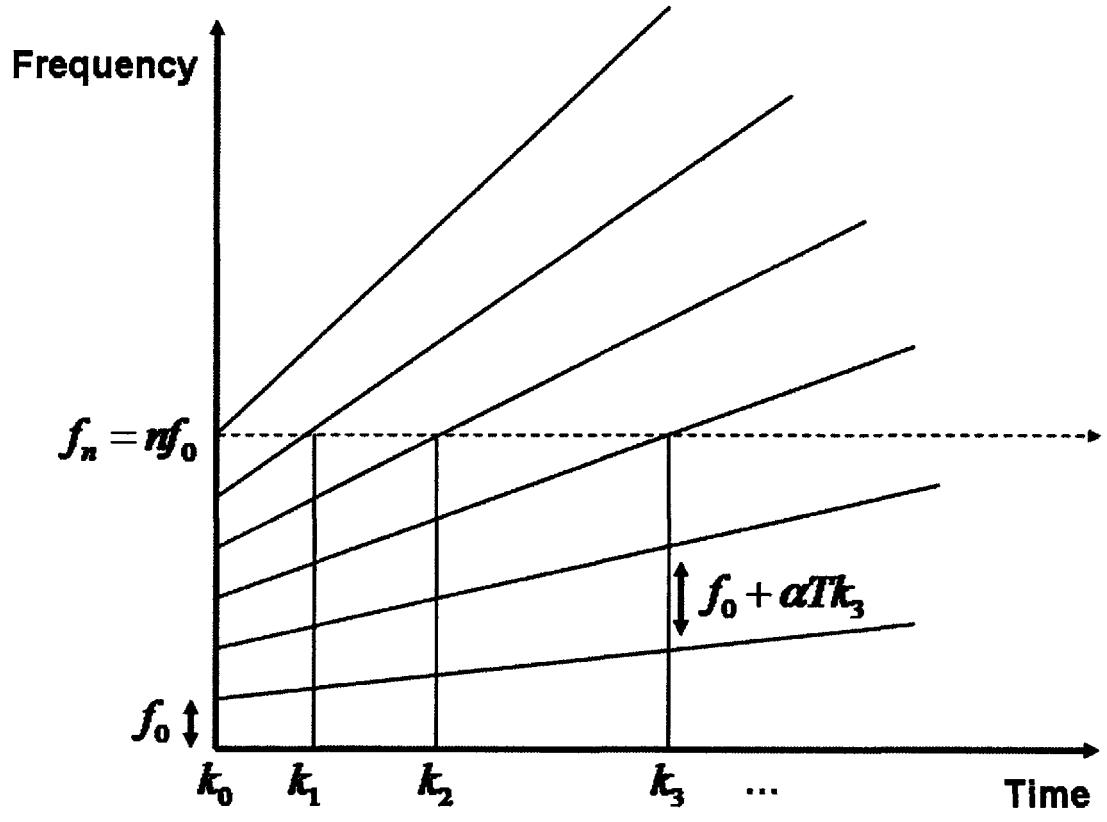

Figure 68 – Illustration showing fanned line structure invoked on the STFT when pitch is changing. The dotted horizontal line corresponds to the frequency of the n^{th} harmonic of f_0.

Consider the n^{th} frequency bin of an STFT computed on a signal with linearly changing pitch such that its corresponding frequency is $f_n = nf_0$, and f_0 is the starting pitch value (Figure 68) The temporal trajectory of this frequency bin will contain peaks with frame positions denoted as

k_i. We assume the first peak corresponds to the n^{th} harmonic at k_0. In addition, let us denote T as the frame interval (in absolute time, e.g., units of seconds) of the STFT and α as the rate of the linearly changing pitch (in units of Hz / second). To solve for the set of k_i corresponding to this n^{th} frequency bin, we observe from the fanned line structure of that we may impose the following constraint:

$$f_n = (f_0 + k_i \alpha T)(n-i) \;\; for \;\; i = 1,2,....(n-1)$$

$$such\ that \tag{0.22}$$

$$k_i = \frac{f_0 i}{(n-i)\alpha T} \;\; for \; i = 1,2,...,(n-1)$$

In words, at frame k_i, f_0 will have increased sufficiently via α such that its $(n-i)^{th}$ harmonic equals f_n. In Figure 69 (top), we show the STFT of a synthetic source signal generated at 16 kHz with $f_0 = 200$, $T = 1$, and $\alpha = 100$ Hz/second. We choose $n = 10$ such that $f_n = 2000$ Hz; this frequency bin is denoted with the arrow. In Figure 69 (bottom), we show the temporal trajectory of this frequency across time. In Figure 70, we use the result of (0.22) to compute the inter-peak spacings of the peaks observed in ("raw"). Specifically, the abscissa plots $i = 1,2,...10$ while the ordinate plots $T(k_i - k_{i-1})$. A linear and exponential fit to these inter-peak spacings are also shown. Observe that the exponential fit better matches the raw values better than the linear fit.

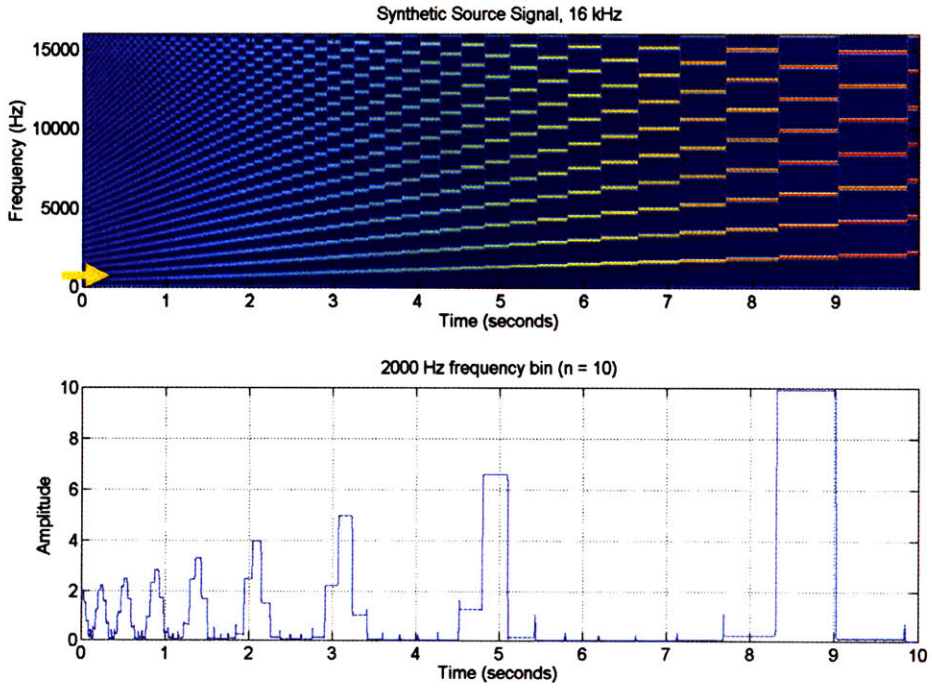

Figure 69 – STFT of synthetic source signal (top) and the temporal trajectory of the frequency bin corresponding to the 10^{th} harmonic of a 200-Hz pitch (bottom).

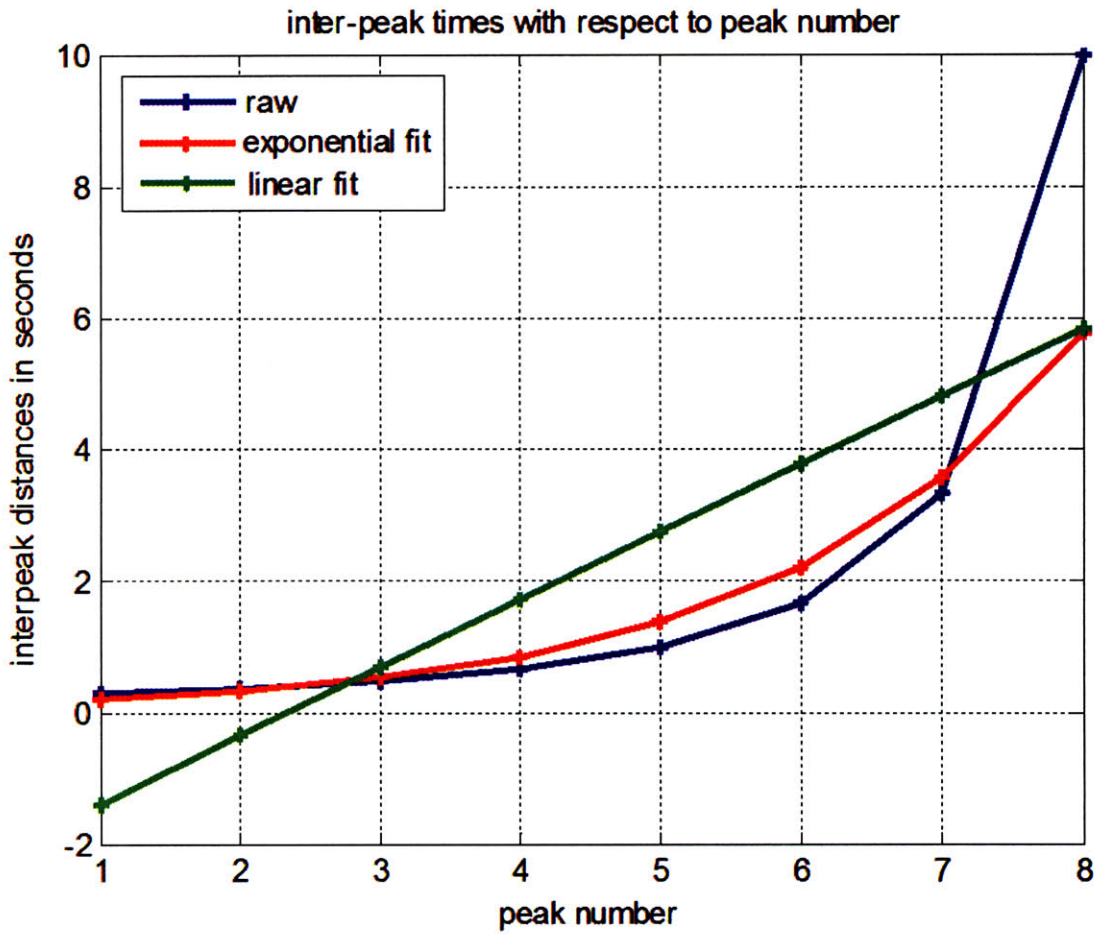

Figure 70 - Inter-peak times of the temporal trajectory computed from (0.22) with α = 100 Hz/sec, T = 1 ms, n = 10, $f_0 = 200$ Hz.

The implication of these findings in relation our 2-D processing framework is as follows. In the narrow-band STFT, pitch is represented along the frequency direction with evenly spaced harmonic peaks. Observe that this condition still holds when pitch is changing (e.g., $f_0 + \alpha T k_3$, Figure 68) . In the temporal direction, linearly changing pitch invokes an exponential-like spacing of harmonic peaks. Therefore, to coherently map this fanned line structure to a 2-D space, a transform employing a sinusoidal basis in the frequency direction and an exponential basis in the time direction is required. Such a transform is distinct from the GCT (which employs sinusoidal bases in both directions) and avoids making the assumption of parallel harmonic line structure under conditions of changing pitch.

References

[1] J. E. Markel and A. H. Gray, *Linear Prediction of Speech*. Secaucus, NJ: Springer-Verlag New York, Inc., 1983.

[2] R. W. Schafer and L. R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," *Journal of the Acoustical Society of America*, vol. 47, pp. 634-648, 1970.

[3] T. F. Quatieri, *Discrete-time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice-Hall, 2001.

[4] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1975.

[5] S. Mcadams, "Segregation of Concurrent Sounds.1. Effects of Frequency-Modulation Coherence," *Journal of the Acoustical Society of America*, vol. 86, pp. 2148-2159, 1989.

[6] R. L. Diehl, B. Lindblom, K. A. Hoemeke, and R. P. Fahey, "On explaining certain male-female differences in the phonetic realization of vowel categories," *Journal of Phonetics*, pp. 187-208, 1996.

[7] J. S. Mason and J. Thompson, "Gender Effects in Speaker Recognition," presented at International Conference on Signal Processing, Beijing, China, 1995.

[8] T. Chi, P. W. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, pp. 887-906, 2005.

[9] T. F. Quatieri, "2-D Processing of Speech with Application to Pitch Estimation," presented at International Conference on Spoken Language Processing, Denver, CO, 2002.

[10] Ezzat T., Bouvrie J., and Poggio T., "Spectrotemporal Analysis of Speech Using 2-D Gabor Filters," presented at Interspeech, Antwerp, Belgium, 2007.

[11] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 1998.

[12] F. I. Itakura and S. Saito, "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies," *Elec. and Comm. in Japan*, vol. 53-A, pp. 36-43, 1970.

[13] W. Verhelst and O. Steenhaut, "A New Model for the Short-Time Complex Cepstrum of Voiced Speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. p. 43-51, 1986.

[14] G. Kopec, A. V. Oppenheim, and J. Tribolet, "Speech Analysis by Homomorphic Prediction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, pp. p. 40-49, 1977.

[15] M. S. Rahman and T. Shimamura, "Identification of ARMA speech models using an effective representation of voice source," *Ieice Transactions on Information and Systems*, vol. E90D, pp. 863-867, 2007.

[16] D. Mehta and T. F. Quatieri, "Synthesis, analysis, and pitch modification of the breathy vowel," presented at IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, 2005.

[17] Ezzat T., Bouvrie J., and Poggio T., "Max-Gabor Analysis and Synthesis of Spectrograms," presented at International Conference on Spoken Language Processing, Pittsburgh, PA, 2006.

[18] "NSL MATLAB Toolbox: http://www.isr.umd.edu/Labs/NSL/Software.htm."

[19] J. Pickles, *An Introduction to the Physiology of Hearing*, 2nd ed. St. Louis, MO: Academic Press, Inc., 1999.

[20] G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," *Journal of the Acoustical Society of America*, vol. 24, pp. 175-184, 1951.

[21] S. Greenberg and H. Hitchcock, "Stress-accent and vowel quality in the Switchboard corpus," in *Workshop on Large-Vocabulary Continuous Speech Recognition*, 2001.

[22] D. H. Klatt, "Software for a cascade-parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, pp. 971-995, 1980.

[23] D. Mehta, "Aspiration Noise during Phonation: Synthesis, Analysis, and Pitch-Scale Modification," in *Electrical Engineering and Computer Science*, vol. Master's of Science. Cambridge: Massachusetts Institute of Technology, 2006.

[24] D. Paul, "The spectral envelope estimation vocoder," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, pp. 786-794, 1981.

[25] R. Gonzalez and R. Woods, *Digital Image Processing*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, Inc., 2002.

[26] G. Vallabha and B. Tuller, "Systematic errors in the formant analysis of steady-state vowels," *Speech Communication*, vol. 38, pp. pp. 141-160, 2002.

[27] S. Davis and S. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Signal Processing*, vol. 28, pp. 357-366, 1980.

[28] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72-83, 1995.

[29] D. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[30] T. T. Wang and T. F. Quatieri, "Exploiting Temporal Change of Pitch in Formant Estimation," in *International Conference on Acoustics, Speech, and Signal Processing (to be presented)*. Las Vegas, NV, 2008.

[31] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," presented at The DARPA Workshop on Speech Recognition, 1986.

[32] Y. Medan, E. Yair, and D. Chazan, "Super Resolution Pitch Determination of Speech Signals," *IEEE Transactions on Signal Processing*, vol. 39, pp. 40-48, 1991.

[33] Weruaga Luis. and M. Kepesi, "The fan-chirp transform for non-stationary harmonic signals," *Signal Processing*, vol. 87, pp. p. 1504 - 1522, 2007.