Toward an Interpretive Framework of Two-Dimensional Speech-Signal Processing

by

Tianyu Tom Wang

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

S.M., Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 2008

SUBMITTED TO THE HARVARD-MIT DIVISION OF HEALTH SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN BIOMEDICAL ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2011

Signature of Author ....
Harvard-MIT Division of Health Sciences and Technology
May 16, 2011

Certified by ....
Thomas F. Quatieri, Sc.D.
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 ....
Ram Sasisekharan, PhD/Director, Harvard-MIT Division of Health Sciences and Technology/Edward Hood Taplin Professor of Health Sciences & Technology and Biological Engineering.

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 author was additionally supported by the National Institutes of Deafness and Other Communicative Disorders under grant 5 T 32 DC00038.

Toward an Interpretive Framework of Two-Dimensional Speech-Signal Processing

by

Tianyu Tom Wang

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

S.M., Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 2008

Submitted to the Harvard-MIT Division of Health Sciences and Technology
on May 16th, 2011 in partial fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Biomedical Engineering

Abstract

Traditional representations of speech are derived from short-time segments of the signal and result in time-frequency distributions of energy such as the short-time Fourier transform and spectrogram. Speech-signal models of such representations have had utility in a variety of applications such as speech analysis, recognition, and synthesis. Nonetheless, they do not capture spectral, temporal, and *joint* spectrotemporal energy fluctuations (or "modulations") present in *local* time-frequency *regions* of the time-frequency distribution. Inspired by principles from image processing and evidence from auditory neurophysiological models, a variety of *two-dimensional* (2-D) processing techniques have been explored in the literature as alternative representations of speech; however, speech-based models are lacking in this framework.

This thesis develops speech-signal models for a particular 2-D processing approach in which 2-D Fourier transforms are computed on local time-frequency regions of the canonical narrowband or wideband spectrogram; we refer to the resulting transformed space as the Grating Compression Transform (GCT). We argue for a 2-D sinusoidal-series amplitude modulation model of speech content in the spectrogram domain that relates to speech production characteristics such as pitch/noise of the source, pitch dynamics, formant structure and dynamics, and offset/onset content. Narrowband- and wideband-based models are shown to exhibit important distinctions in interpretation and oftentimes "dual" behavior. In the transformed GCT space, the modeling results in a novel taxonomy of signal behavior based on the *distribution* of formant and onset/offset content in the transformed space via source characteristics. Our formulation provides a speech-specific interpretation of the concept of "modulation" in 2-D processing in contrast to existing approaches that have done so either phenomenologically through qualitative analyses and/or implicitly through data-driven machine learning approaches. One implication of the proposed taxonomy is its potential for interpreting transformations of other time-frequency distributions such as the auditory spectrogram which is generally viewed as being "narrowband"/"wideband" in its low/high-frequency regions.

The proposed signal model is evaluated in several ways. First, we perform analysis of synthetic speech signals to characterize its properties and limitations. Next, we develop an algorithm for analysis/synthesis of spectrograms using the model and demonstrate its ability to accurately

represent real speech content. As an example application, we further apply the models in co-channel speaker separation, exploiting the GCT's ability to distribute speaker-specific content and often recover overlapping information through demodulation and interpolation in the 2-D GCT space. Specifically, in multi-pitch estimation, we demonstrate the GCT's ability to accurately estimate separate and crossing pitch tracks under certain conditions. Finally, we demonstrate the model's ability to separate mixtures of speech signals using both prior and estimated pitch information. Generalization to other speech-signal processing applications is proposed.

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. Throughout the past five years, I have greatly appreciated Tom's dedication to students and his infectious enthusiasm for research. Thanks, Tom, for nurturing my analytical skills, pushing me when I needed it, and encouraging me through some of my pessimism in completing this thesis. Under your guidance, I went from being a clueless graduate student to a confident researcher and problem solver. No matter how far I may roam from speech signals, I will always value your opinion on anything and everything as a lifelong mentor, friend, and colleague.

I would like to thank my committee, Julie Greenberg, Patrick Wolfe, Herbert Voigt, and Barbara Shinn-Cunningham for the many pleasant, lively, and insightful discussions that led to completion of this thesis. Thanks to Patrick for encouraging and supporting my interests in statistical estimation techniques in the multi-pitch estimation problem. Thanks to Julie for bringing up important signal processing issues that led to refinements in this work. Thanks also to Julie for agreeing to chair my committee in light of the administrative challenges that came with being part of the Harvard-MIT Division of Health Sciences and Technology (HST) at MIT. Many thanks to Herb and Barb, for providing the solid biological insights that enriched the interpretation and development of the ideas in this thesis.

I would like to thank Clifford Weinstein and all of the members of the Human Language Technology Group at MIT Lincoln Laboratory. As a graduate student, I could not have asked for a more supportive and constructive environment to conduct my thesis research. Working at Lincoln helped me grow not only as a scientist and engineer but also as a person and professional. I would also like to thank the staff and faculty at MIT for their support during my graduate school career. Thanks to John Rosowski for being a great academic advisor. Thanks to James Glass for teaching a great course on automatic speech recognition and allowing me to use the Spoken Language Systems Group's computers during preliminary parts of this work. Thanks to George Verghese for giving me the opportunity to act as a teaching assistant for 6.011. Thanks to Louis Braida and Bertrand Delgutte for championing the HST Speech and Hearing Bioscience and Technology program (SHBT) at MIT. Finally, I would like to thank MIT Lincoln Laboratory and the National Institutes of Health for funding this work.

I would like to thank all of my friends, officemates, and roommates for their support and friendship. While research was fun in its own right, my memories of graduate school would have been a mere zillion lines of MATLAB if not for the many epic (and epic is the only appropriate word here!) adventures in between. At MIT: Nick, Daryush, Dan, Nancy, Zahi, TAB, Loomis, Nivedita, mens-club, and all of the awesome folks in HST and SHBT! From Tech: Ethan, Jeremy, Matt, En, JLee, June, and Pooja. From Pope: Kevin, Brendan, Joe, Seekely, Eric, Max, and Jon Wilson. From Athens: Pete and Steven. From Boston: Dana and Catiah.

Last but certainly not least, I would like to dedicate this thesis to my parents Zhikai Wang and Yuan Yang for all of the sacrifices they have made on my behalf and for their unending love and support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

A fundamental goal of speech-signal processing is to obtain models of empirical representations of the signal such as a speech waveform or its spectrogram). The model's ability to accurately represent speech content generally motivates application to a variety of speech processing tasks. Examples in analysis and/or synthesis include pitch and formant estimation and voice modification. Additional applications include feature extraction for speech/speaker recognition and speech enhancement. In this thesis, we formulate and evaluate models of speech content in a *two-dimensional* (2-D) representation of the signal.

1.1 Problem Statement

Traditional representations of speech are typically obtained by extracting and analyzing *short-time* segments of the speech waveform. As a canonical example, the short-time Fourier transform (STFT) performs Fourier analysis of segments of the signal to characterize the frequency/spectral components of each segment across time. This approach results in the spectrogram, or more generally for non-Fourier methods (e.g., wavelet transform [1]), a time-frequency distribution of the signal. Classical models of speech that have been developed using the STFT include the cepstral [2], all-pole [3], and sinusoidal-based representations [4].

Despite its utility in existing speech processing tasks, a critical limitation of the short-time analysis framework is its inability to analyze spectral, temporal, and spectrotemporal "modulations" in the time-frequency distribution itself. Here, we refer to "modulation" in a loose sense and characterize it qualitatively as energy fluctuations in a time-frequency distribution. Observe for instance in Figure 1-1 temporally, spectrally, and spectrotemporally-oriented fluctuations in energy for a narrowband spectrogram corresponding to vowels, onsets/offsets, and noisy content. To explicitly analyze such components, recent findings from auditory neurophysiology coupled with image processing principles, have motivated a *two-dimensional* (2-D) processing framework in which 2-D analysis is performed on the time-frequency distribution *itself* (Figure 1-2). Examples of this generalized 2-D processing framework include the modulation spectrogram proposed in [5], the physiologically-motivated model of spectrotemporal receptive fields in the mammalian cortex of [6], and early work in 2-D Fourier analysis of the spectrogram e.g., in [7].

While existing work in 2-D processing methods for speech has motivated several representations of the underlying speech signal, an outstanding difficulty lies in their *interpretation*, particularly with respect to the concept of "modulations". As will be subsequently discussed, "modulations" are often characterized qualitatively, through implicit methods (e.g., data-driven/machine learning techniques), or through an analytical construct without relation to easily interpretable characteristics of speech (e.g., pitch and formant structure in speech production). The aim of this thesis is to explicitly relate a class of 2-D representations to a concept of modulation that is also based on underlying properties of speech production characteristics.

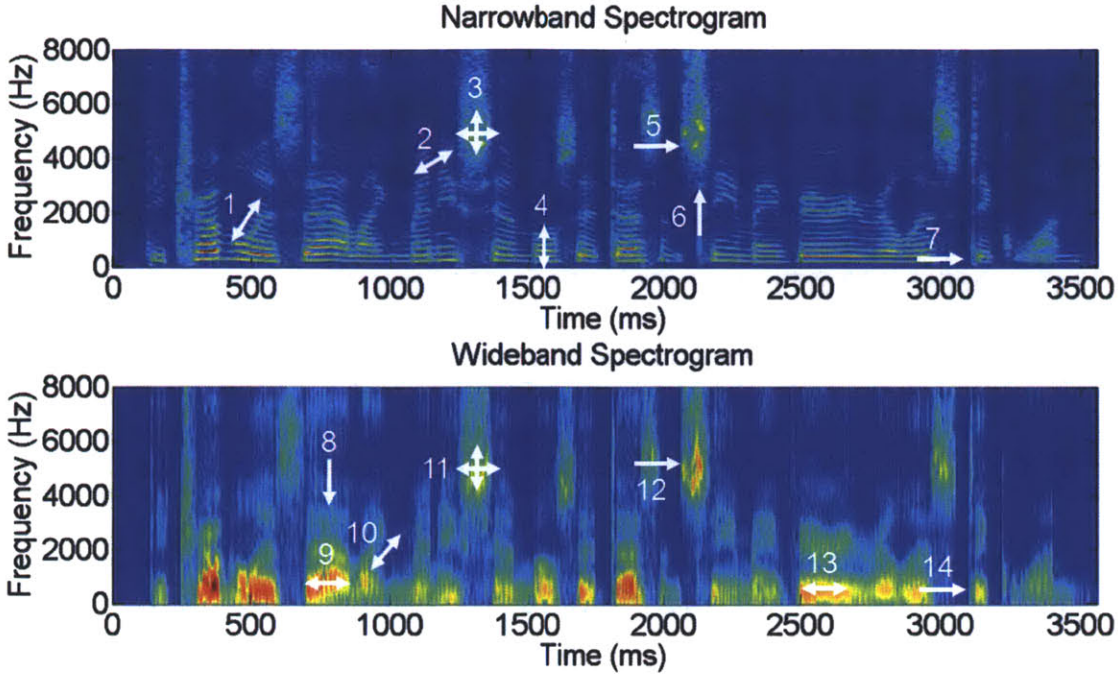

Figure 1-1. Narrow- (top) and wide- (bottom) band spectrograms[1] of a female speaker utterance, "The trouble with that is that like many symbols it doesn't seem a very realistic one". Energy fluctuations ("modulations") highlighted numerically: onset of plosive burst (/t/), (7, 14); noise in fricative /s/, (3, 11); formant structure of vowel (1, 10); harmonic and periodic content of voicing (4, 9).

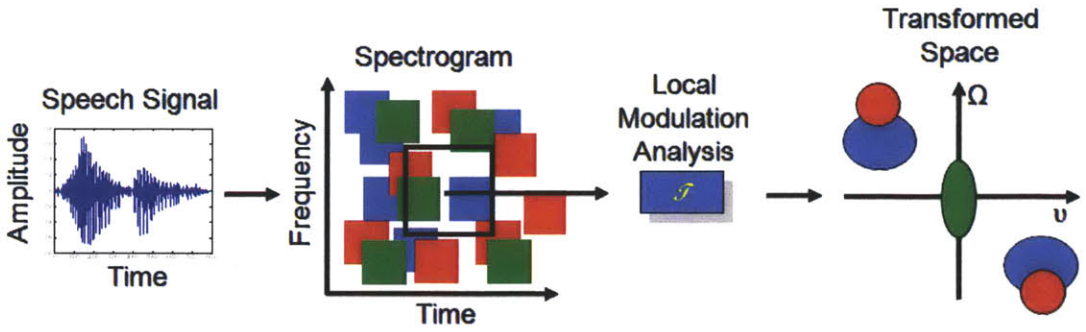

Figure 1-2. Generalized 2-D framework in which 2-D analysis is performed on local regions of a time-frequency distribution to result in a transformed 2-D space.

1.2 Framework

In this thesis, we consider a particular realization of the general 2-D framework referred to as the Grating Compression Transform (GCT). The GCT is defined as the 2-D Fourier transform of a local time-frequency region of the spectrogram. We consider two canonical spectrogram types: narrowband and wideband, and derive models for speech content in the resulting GCT space.

[1] *Unless otherwise denoted, spectrograms in this thesis are plotted by taking the fourth root of the magnitude short-time Fourier transform for display purposes to avoid image scaling issues in e.g., silent regions.*

Our choice of narrow and wideband spectrograms stems from the fact that these are the traditional time-frequency representations used in analyzing speech and reflect distinct time and frequency resolutions. In addition, they serve the basis for other time-frequency distributions that are often viewed as "mixtures" of the narrowband and wideband representations as will subsequently be discussed. We propose a 2-D sinusoidal-series modulation model in which the carrier and envelope terms are shown to reflect distinct characteristics of the speech waveform (e.g., pitch/formant content, fricatives/noise, onset/offsets). In the corresponding transformed GCT space, the models invoke a *distribution* of *copies* of the envelope content based on the specific parameters of the carrier. Furthermore, both temporally stationary and *dynamic* content in the form of pitch and formant dynamics are modeled explicitly in this representation.

The proposed signal models are evaluated in several ways. Firstly, we perform simulations on synthetic signals to highlight limitations and properties of the models. Based on these observations, we next develop and test an analysis/synthesis framework for reconstructing spectrograms of individual speakers. As our models will be shown to reflect distinct properties of speech production of individual speakers (e.g., pitch/formant structure and their dynamics), we further develop and test algorithms for application to co-channel speaker separation. Specifically, we use the GCT framework both in multi-pitch analysis/estimation as well as signal separation. In this context, we view our efforts not only as assessing the utility and applicability of the GCT for this task but also as further evaluation of the models' ability to represent speech content from individual speakers.

1.3 Summary of Contributions

The primary contribution of this thesis is the formulation of models of speech content in local time-frequency regions of both of the narrowband and wideband spectrograms as well as their Grating Compression Transform representations. The model is based on sinusoidal modulation in the time-frequency space; specifically, a speech *source*-based carrier reflecting pitch, pitch dynamics, and noisy source signals is represented by a 2-D sinusoidal series resting on a DC pedestal. This carrier is *modulated* by a slowly-varying envelope term reflecting formant structure and formant dynamics as well as onset/offset content. In the GCT space, this model invokes a distribution of copies of the envelope component at locations reflecting the specific carrier parameters. The models allow for recovery and interpolation of envelope terms that may, under certain conditions (e.g., co-channel speech mixtures) exhibit *overlap* with interfering signals. Furthermore, the models allow for explicit representation and exploitation of *dynamic* content for both pitch and formant dynamics.

The combined wideband and narrowband signal models invoke a taxonomy of speech-signal behavior in the GCT space that can be distinctly interpreted and often exhibit "dual" behavior based on speech production parameters (i.e., a noisy or voiced source signal, formant structure and formant dynamics, onset/offset structure). As will subsequently be discussed, we propose this taxonomy as an *interpretive framework* for not only the GCT but of other time-frequency distributions as well as other 2-D processing approaches. As a simple example, the auditory spectrogram is often viewed as "narrowband"/"wideband" in low/high-frequency regions, and the derived models may have implications for interpreting this alternative representation as well. To demonstrate the utility of the GCT for speech-signal processing applications, we develop algorithms to perform analysis/synthesis (i.e., reconstruction) of wideband and narrowband spectrograms. These algorithms are shown to provide accurate reconstructions of spectrograms on with average root-mean-squared errors (RMSE) of 5e-3 ~ 4e-2 when computed relative to a maximal value of unity (see Chapter 3). Furthermore, when combined with the original phase of the speech signals, the reconstructed waveforms exhibit very good speech quality with signal-to-noise ratios of 11~20 dB (see Chapter 3).

Building on our analyses and analysis/synthesis algorithms, and as an example application, we further apply the GCT to the problem of co-channel speaker separation. In particular, we develop algorithms for both signal separation (using both prior and estimated pitch information) and multi-pitch estimation. In multi-pitch estimation, we demonstrate the GCT's ability to estimate pitch tracks that are both crossing and separate under all-voiced conditions. In signal separation, we demonstrate that the narrow and wideband GCT representations result in good separation of the underlying speech signals with signal-to-noise ratios of 4~7 dB (relative to a 0 dB initial SNR). Motivated from the "dual" nature of the GCT representations of wideband and narrowband spectrograms, we further perform *fusion* of the estimates demonstrating gains of ~1 dB over either representation alone. These results provide evidence for complementary information captured by both narrowband and wideband representations. For comparison purposes, we develop extensions to a traditional frame-based sinusoidal separation system to handle silent and unvoiced regions of speech mixtures providing a baseline that is shown to exhibit performance of 6~9 dB. Though the GCT-based representation does not outperform the baseline system, *fusion* of waveforms results in a ~1 dB gain above the reference providing evidence for complementary information using the GCT for the separation task. Finally, we show that the GCT is a promising framework for this task by combining elements of the multi-pitch estimation and signal separation methods in a prototype full separation system for separation of mixtures of male and female speakers; this system is demonstrated to result in SNR gains of ~4 dB on mixtures of male and female speakers with voiced and unvoiced speech.

1.4 Thesis Outline

This thesis is organized as follows. In Chapter 2, we provide background to both the general 2-D processing framework as well as the example application of co-channel speaker separation. In Chapter 3, we derive speech-signal models for narrowband spectrograms and evaluate these models using simulations, analysis/synthesis (i.e., reconstruction) of spectrograms, as well as speaker separation using prior information. In Chapter 4, we perform the analogous steps as in but for wideband spectrograms; furthermore, in Chapter 4, we relate the wideband and narrowband representations through a taxonomy of speech signal behavior for the GCT. In Chapter 5, motivated by the narrowband model's ability to represent pitch information of individual and multiple speakers, we perform multi-pitch analysis and estimation in as a further test of the model. In Chapter 6, we develop a prototype system for speaker separation by combining signal separation and multi-pitch analysis/estimation methods. We conclude in Chapter 7 with a discussion of future directions.

Chapter 2

Background

In this chapter, we discuss background related to 2-D signal representations for speech; in addition, we outline the basic framework for the Grating Compression Transform. As additional background, we also discuss the co-channel speaker separation problem and describe several existing approaches.

2.1 Two-dimensional Modulation Representations

As described in Chapter 1, a generalized 2-D processing framework for speech is based on performing 2-D analysis of a time-frequency distribution, thereby explicitly analyzing energy fluctuations/"modulations" along spectral, temporal, and spectrotemporal dimensions (Figure 1-1). In this section, we review several realizations of this general 2-D framework.

2.1.1 Modulation Spectrogram

The *modulation spectrogram* is derived from a 2-D processing approach that analyzes temporal modulations of the spectrogram across time and frequency [5]. Specifically, from Figure 2-1, a modulation *spectrum* at a specific frequency and time in the spectrogram is obtained by computing the Fourier transform of a *time slice* from a (typically) narrowband spectrogram. The "modulation frequency" is defined as the transformed time variable via the Fourier transform (for a particular frequency band). This analysis is performed for each frequency band, and the resulting spectra can be combined to generate a 2-D function of frequency and modulation frequency for a particular point in time [8]. An alternative view can also be obtained as a function of time and frequency for a fixed modulation frequency f_m (e.g., as in [5]). Variations on the modulation spectrogram (e.g., multi-scale modulation spectrograms [9]) have also been proposed in the literature.

This framework has been applied in a variety of applications such as channel compensation [10], speech analysis [8], and co-channel speaker separation [11]. Furthermore, a model for acoustic signals has been proposed by Atlas and colleagues using properties of the Hilbert envelope [8]. Nonetheless, modulation components (i.e., an envelope component multiplying/*modulating* a carrier [8]) are defined in this framework with no explicit reference to *speech* components, thereby corresponding to a general analytical construct. Furthermore, the modulation spectrogram focuses exclusively on *temporal* modulation components for a fixed frequency band, thereby neglecting modulations that may occur *across* frequency.

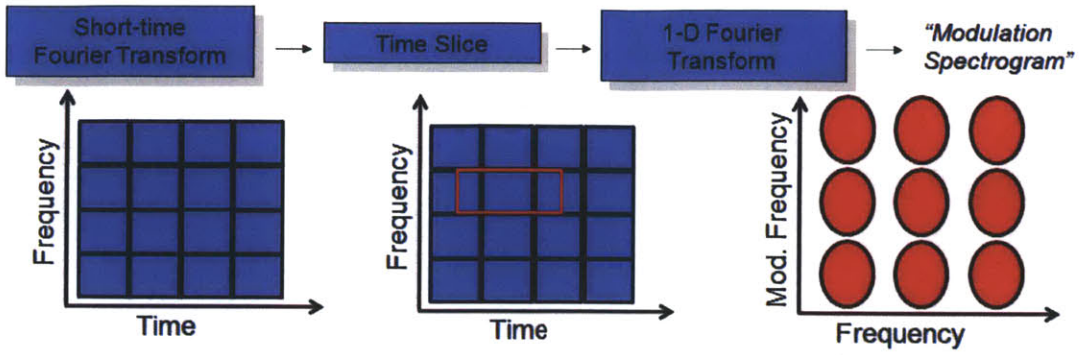

Figure 2-1. Modulation spectrogram framework; the short-time Fourier transform is computed using a fixed time-frequency tiling. Time slices for each frequency band are used in Fourier analysis to generate the modulation spectrogram.

2.1.2 Spectral Expansion

As a "dual" to the modulation spectrogram, spectral expansion was proposed in [12] as a way to characterize fluctuations along the *frequency* axis at multiple scales Figure 2-2. In particular, an auditory spectrogram is computed as the time-frequency distribution. Subsequently, spectral slices of the auditory spectrogram are analyzed using a multi-scale filterbank across different scales and corresponding modulation frequencies.

Spectral expansion has been argued analytically to exhibit robustness properties in the presence of additive noise [12]. Efforts to interpret this representation have nonetheless been limited to qualitative and phenomenological analyses in demonstrating that distinct modulations correspond to distinct spectral shapings of vowels. In addition, the representation does not explicitly capture *temporal* modulation patterns present in the auditory spectrogram. In [13], it was shown that spectral expansion can be used to derive a feature set that is a "superset" of the traditional mel-cepstral coefficients, thereby resulting in improved phoneme recognition in the presence of noise. Nonetheless, in this work, modeling was done implicitly through a traditional hidden Markov model framework in which distinct modulation patterns of phonemes were learned through training data without reference to distinct speech parameters.

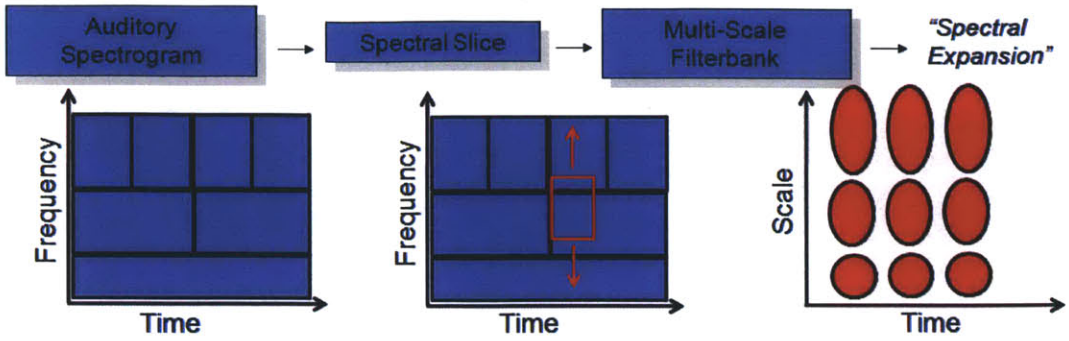

Figure 2-2. Spectral expansion framework; an auditory spectrogram computed; spectral slices are passed through a multi-scale filterbank along the frequency axis to generate a scale-time representation.

2.1.3 Spectrotemporal Auditory Model

Building on the work of [12], the auditory model of Chi, et al. [6] analyzes modulation components along both time and frequency. Specifically, an auditory spectrogram is computed on the waveform followed by 2-D filtering using a bank of 2-D filters with varying durations in time and frequency (Figure 2-3). This results in a *multi-resolution* representation along time, frequency, and temporal (υ) and frequency (Ω) modulation axes. Earlier versions of the model (e.g., in [13]) also focused exclusively on *spectral* modulation characteristics, performing analysis across the frequency axis of the auditory spectrogram, analogous to the modulation spectrogram's analysis of temporal modulation. A key component of the auditory model is its use of wavelet-like analysis both at the short-time frame and subsequent 2-D levels. In particular, the auditory spectrogram represents a non-uniform "tiling" of the time frequency space, while the subsequent 2-D filterbank contains filters with non-uniform bandwidths in the auditory space. The resulting transformed space consists of the υ and Ω axes that reflect "modulation" content oriented along time and frequency, respectively. Finally, we note also that the model contains several nonlinear components such as inner hair cell rectification in short-time analysis, and lateral inhibition across frequency bands of the auditory spectrogram, both incorporated to mimic presumed biological mechanisms in auditory processing.

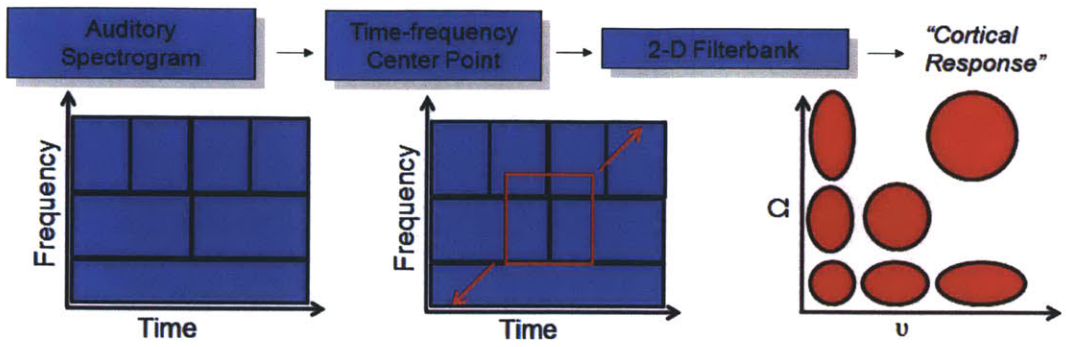

Figure 2-3. The auditory spectrogram is analyzed with a bank of 2-D filters that span distinct resolutions in the time-frequency space, resulting in a 2-D filtered/transformed space with distinct resolutions.

The described auditory model has been used for voice activity detection [14], phoneme recognition [13], and perceptual modeling [15]. These efforts have exclusively utilized data-driven learning methods to *automatically* characterize the model outputs, presumably due to the difficulty of directly interpreting the resulting space. For instance, in [16], we observed that harmonic line structure was observed to be non-uniformly spaced in the auditory spectrogram for periodic source signals such that any subsequent 2-D filterbank representation does not provide a *coherent* mapping of these components. The auditory model therefore does not motivate an model of speech signals in relation to underlying production characteristics.

2.1.4 Grating Compression Transform

From our previous discussions, we have seen that while existing methods of 2-D processing of speech has resulted in a novel paradigm for analysis, outstanding limitations include their inability to analyze the "complete" set of modulations present in the spectrogram such as temporal (across time), spectral (across frequency) and *joint* spectrotemporal (time *and* frequency). Furthermore, interpretation of these frameworks in relation to speech production characteristics such as specific parameters (e.g., pitch, pitch dynamics, formant structure) is lacking. We summarize in Table 2-1 properties of these frameworks; in addition, we list properties of an alternative framework to be explored in this thesis referred to as the Grating Compression Transform (GCT). From Table 2-1, observe for instance that spectral expansion and the modulation spectrogram are "duals" of each other in not analyzing spectral and temporal modulations, respectively. Furthermore, interpretations of these frameworks that relate to speech stem from either qualitative/phenomenological observations or are *probabilistic* in nature. Specifically, in the probabilistic setup, a statistical model is developed using training data on representations of distinct speech sounds for the purposes of speech and/or phone recognition (e.g., see [13], [17]). Finally, we have noted that signal models of the modulation spectrogram have been proposed analytically though without specific reference to *speech-based* parameters; this framework has been applied for instance in analysis of musical signals [18].

Herein we describe the 2-D processing approach taken in this thesis referred to as the Grating Compression Transform (GCT). The GCT is defined as the 2-D Fourier transform of a *localized* time-frequency region of the short-time Fourier transform magnitude (or log-magnitude, see Chapter 4) as schematized in Figure 2-4. In relation to the auditory model of the previous section, the GCT is based on a *uniform* "tiling" of the time frequency space dependent on the length of the short-time analysis window used in computing the spectrogram. Furthermore, the subsequent 2-D Fourier analysis of local regions of the spectrogram are computed based on time-

frequency regions of fixed size. Viewing the 2-D Fourier transform from the filterbank view, note that this is equivalent to filtering with a 2-D filterbank with *uniformly* sized filters in the GCT domain. The resulting transformed space has axes of υ and Ω corresponding to "modulation" frequencies in time and frequency, respectively.

Table 2-1. Summary of 2-D processing approaches aimed at analyzing energy fluctuations/"modulations" in time-frequency distributions.

	Spectral Expansion	Modulation Spectrogram	Auditory Cortex Model	Grating Compression Transform
Spectral	Yes	No	Yes	Yes
Temporal	No	Yes	Yes	Yes
Spectrotemporal	No	No	Yes	Yes
Interpretation	Speech-specific but probabilistic and/or qualitative	Analytical non-speech-specific	Speech-specific but probabilistic and/or qualitative	Speech-specific but probabilistic and/or qualitative

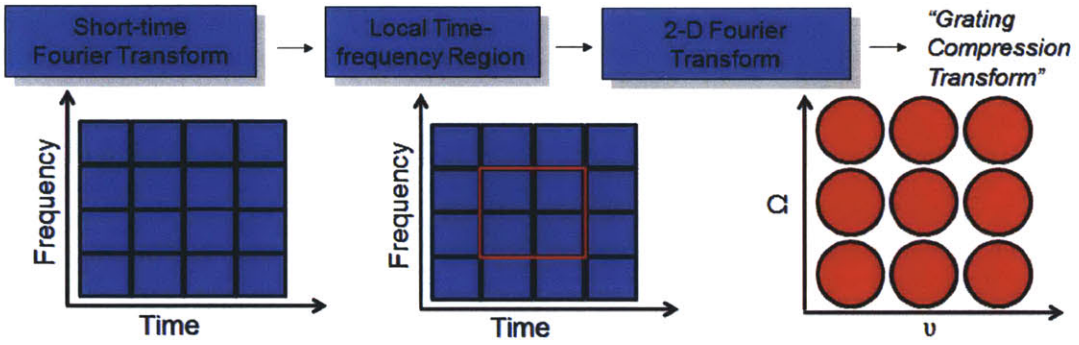

Figure 2-4. Localized time-frequency regions of a spectrogram analyzed using the 2-D Fourier transform resulting in the Grating Compression Transform (GCT) space.

Previous efforts using the GCT have demonstrated its ability to represent pitch information [7]. In our previously published work, we have also demonstrate the GCT's ability to represent and formant structure [19] in distinct regions of the resulting transformed space, multi-pitch information from distinct speakers [20], and its utility for co-channel speaker separation through spectrogram demodulation on all-voiced speech mixtures [21]. In addition, the GCT has been shown to be amenable to data-driven methods in phoneme recognition [22]. In this thesis, we develop a novel model for the GCT that characterizes modulation components explicitly in relation to speech production parameters the represents a culmination of our previous efforts and as a way to potentially interpret the results reported by others using the GCT.

2.2 Co-channel Speaker Separation

As previously discussed, we further explore in this thesis an example application using the GCT framework in co-channel speaker separation (CSS). Our emphasis is on assessing the signal model's ability to represent speech as well as its potential to address distinct aspects of the general CSS problem. As a problem definition for CSS, we consider obtaining from a *single* channel (e.g., an acoustic microphone) a mixture waveform $y[n]$ consisting of a set of speech waveforms $x_i[n]$, with each i corresponding to a distinct speaker. $y[n]$ is a weighted sum of the individual waveforms

$$y[n] = \sum_{i=1}^{N} \alpha_i x_i[n] \quad (2.1)$$

where N is the total number of speakers present in the mixtures, and α_i are scale factors that control the relative energies of each speaker in the mixture. The goal of CSS is to obtain an estimate $\hat{x}_i[t]$ of $x_i[t]$ from $y[t]$ according to some goodness criterion. In our work, we exclusively consider the case of $N = 2$, though our methods can be applied more generally for $N > 2$.

A variety of techniques have been proposed in the literature for co-channel speaker separation (CSS) under distinct formulations of the problem and constraints. Examples of these include parametric modeling (e.g., in sinusoidal analysis/synthesis [23], modulation spectrum [11]), independent components analysis (ICA) (e.g., [24]), computational auditory scene analysis (CASA) (e.g., [25]), and generative modeling (e.g., factorial hidden Markov models, FHMM [26]). ICA-based methods utilize *multiple* channels (e.g., multiple microphones) of observation of the mixture for estimating targets. Generative modeling approaches assume extensive prior knowledge for distinct target speakers; for instance, FHMM approaches separate speakers based on developing a complete model of the target speaker (using training data of the target speaker) *a priori* such that they are speaker *dependent*. We delineate our work in this thesis for application to the *single-channel* and *speaker-independent* setting. CASA approaches in this context generally involve estimating binary masks of individual time-frequency *units* of a canonical spectrogram or auditory spectrogram. No underlying model of a distinct speaker based on speech parameters is applied. Given that the GCT aims to explicitly model speech of individual speakers, we view its approach in signal separation in the context of parametric modeling methods. An example of this approach is that of the sinusoidal-based separation system proposed in [23] in which parametric models for individual speakers are fit to the resulting mixture waveforms; similar approaches have been proposed in [27] to handle unvoiced/voiced speech mixtures and using complex exponential representations rather than sinusoids. Nonetheless, the proposed set of signal models could also be used in other contexts of the CSS problem (e.g., a GCT-based representation could be used in training FHMMs for speaker-dependent separation).

2.3 Conclusions

In this chapter, we have described existing two-dimensional (2-D) speech signal processing approaches for analyzing energy fluctuations/"modulations" present in time-frequency distributions. While the variety of existing methods have motivated an alternative *framework* for processing speech, an outstanding limitation is the inability to directly interpret representations explicitly in relation to basic speech parameters such as pitch or pitch dynamics. We have also described a 2-D processing approach using 2-D Fourier analysis of local time-frequency regions of the spectrogram, a representation referred to as the Grating Compression Transform (GCT). In subsequent chapters, we aim to derive and develop models of speech in the GCT context with the aim of interpreting the GCT space in relation to basic speech parameters. This chapter has also described the problem statement of co-channel speaker separation (CSS); though we aim to develop signal processing techniques for addressing the CSS problem using the GCT, our primary aim is to further assess the GCT and its corresponding speech-signal models' ability to represent speech.

Chapter 3

Narrowband Models

In this chapter[2], we consider two-dimensional (2-D) Fourier analysis of local time-frequency regions of the narrowband spectrogram. We refer to the resulting 2-D Fourier space as the (narrowband) Grating Compression Transform (GCT)[3]. We introduce a novel *sinusoidal series-based modulation model* for speech signals using the GCT. Our model utilizes source content (e.g., noise or voicing) to *distribute* vocal tract (e.g., formant) and onset/offset content in the transformed GCT space (Figure 3-1). Specifically, the model is capable of representing a variety of speech content such as vowels, fricatives, and onset/offsets as low-frequency temporal and spectral fluctuations of the narrowband spectrogram distributed to multiple locations within the GCT space. In our analyses, we investigate properties of the model, as well as limitations using simulations on synthetic signals. Motivated from our observations, we develop and evaluate algorithms for analysis/synthesis of spectrograms that exploit the distribution of vocal tract and onset/offset energy throughout the GCT space. Finally, as a potential application, we explore co-channel speaker separation using pitch tracks of speakers obtained *a priori*. Here, the distribution of replicas of vocal tract content throughout the GCT space is essential for separation in allowing for *recovery* of corrupted components due to the present of an interfering speaker. For this application, we emphasize our focus in assessing the utility of the signal *model* rather than developing a complete separation system.

This chapter is organized as follows. Section 3.1 develops the 2-D speech-signal model for the GCT. We formulate two approaches for analysis/synthesis of spectrograms in Section 3.2; similarly, we present in Section 3.4 algorithms for co-channel speaker using *a priori* pitch estimates of distinct speakers. In Section 3.5 we describe specific methods, evaluation criteria, and present our results on both tasks. We conclude in Section 3.6 with a discussion of our results.

---

[2] Substantial portions of this chapter are taken from [53].

[3] We use "GCT" to denote the *narrowband* GCT representation in this chapter since our focus is on narrowband spectrograms. We delineate a distinction in subsequent chapters between wideband and narrowband GCTs (i.e., WGCT vs. NGCT).

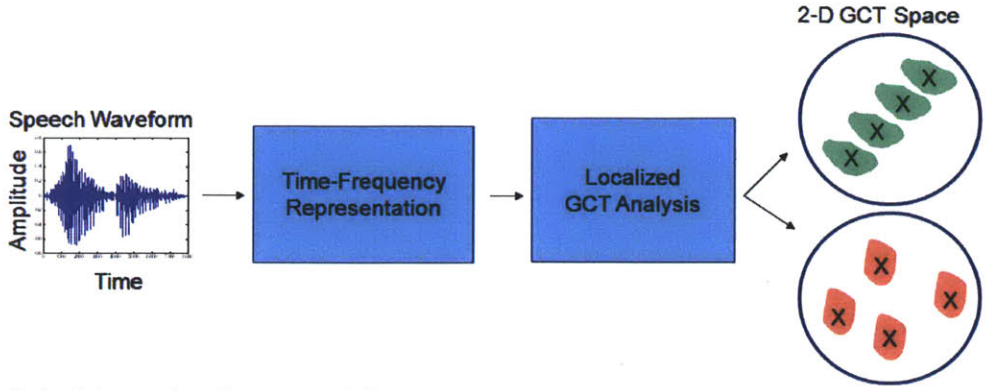

Figure 3-1. Schematic of proposed framework; the GCT space (right) where vocal tract information (shaded, red, green) is *distributed* based on properties of the voicing (top) or noise (bottom) source content ('X').

3.1 2-D Signal Modeling

3.1.1 Voiced Speech

1-D Source Model: We first develop a one-dimensional (1-D) model of the short-time Fourier transform (STFT) magnitude for periodic source signals. Consider a pure impulse train $p[n]$ with periodicity P

$$p[n] = \sum_{k=-\infty}^{\infty} \delta[n - kP]. \tag{3.1}$$

In analysis, $p[n]$ is windowed with a short-time analysis Hamming window $w[n]$, i.e.,

$$p_w[n] = w[n] \sum_{k=-\infty}^{\infty} \delta[n - kP]. \tag{3.2}$$

The *narrowband* STFT of $p_w[n]$ is

$$p_w(\omega) = \frac{2\pi}{P} \sum_{k=-\infty}^{\infty} w(\omega - \frac{2\pi k}{P}) \tag{3.3}$$

where $w(\omega)$ is the Fourier transform of $w[n]$. For narrowband spectrograms, the length N of $w[n]$ is chosen to be *at least* 2~3 times the periodicity P such that the main lobes of the $w(\omega - \frac{2\pi k}{P})$ terms approximately occupy distinct frequency regions of the spectrum and such that $p(\omega)$ exhibits harmonic structure [1] (Figure 3-2b). For analysis, we consider typical pitch values in speech of 60 to 350 Hz such that $w[n]$ can be constrained to be 32~50 ms [28]. Since $p(\omega)$ is periodic with period $\frac{2\pi}{P}$ (Figure 3-1c), it can be decomposed with a Fourier series, or equivalently, a series of cosines [2], i.e.,

$$p_w(\omega) \approx D + \sum_{k=1}^{\infty} \alpha_k \cos\left(\frac{2\pi k}{P}\omega + \psi_k\right) \tag{3.4}$$

where D corresponds to a DC term. For reasons that will subsequently become clear, we refer to $p_w(\omega)$ as a *carrier* term.

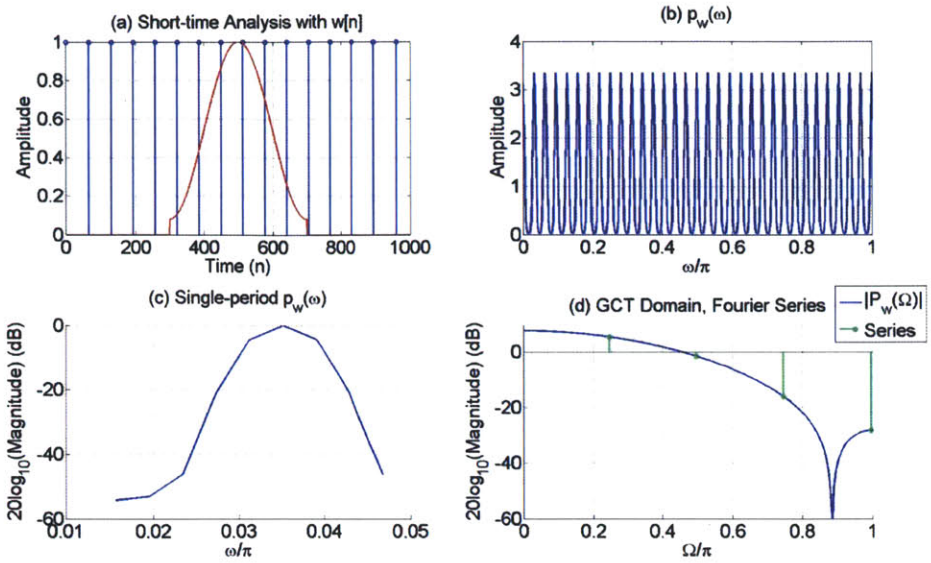

Figure 3-2. (a) Short-time analysis of a pure impulse train (blue) with a short-time Hamming window (red); (b) Short-time spectrum of (a); (c) single period of $p_w(\omega)$; (d) Magnitude of $P_w(\Omega)$ and samples at the fundamental frequency (stem, green). (c-d) plotted on a log scale for display purposes.

Figure 3-2d shows the Fourier transform of $p_w(\omega)$ which we denote as $P_w(\Omega)$ and refer to as the 1-D GCT domain. By sampling $P_w(\Omega)$ at multiples of $\frac{2\pi}{P}$ (i.e., the exponential/sinusoidal Fourier expansion of $p_w(\omega)$ in (3.3)), observe that the periodicity of the source signal is related to the position of the first Fourier coefficient with largest magnitude in the GCT domain. In Figure 3-2d, this coefficient is located at $\Omega = 0.25\pi$ such that the pitch value f_0 can be obtained as

$$f_0 = \frac{2\pi f_s}{N_{STFT}\Omega_0} = \frac{2\pi(16000)}{(512)(0.25\pi)} = 250\ Hz \tag{3.5}$$

where N_{STFT}, f_s, and Ω_0 are the length of the discrete-Fourier transform used to compute the STFT, sampling frequency of the waveform, and position of the maximum peak in the GCT domain, respectively. This characteristic of the GCT space is consistent with the duality of the Fourier transform; specifically, $P_w(\Omega)$ should approximately correspond to a Hamming window if the mainlobe content of each $w(\omega - \frac{2\pi k}{P})$ term in $p_w(\omega)$ does not interact with each other from the narrowband constraint.

For GCT analysis, we extract *localized* regions of $p_w(\omega)$ for use in Fourier analysis with a Hamming window $w_g(\omega)$ applied *along* the ω-axis such that the resulting (1-D) GCT representation $P_w(\Omega)$ is

$$P_w(\Omega) = W_g(\Omega)D + 0.5\sum_{k=1}^{\infty}\begin{bmatrix}\alpha_k e^{-j\Omega\psi_k}W_g\left(\Omega - \frac{2\pi k}{P}\right) + \\ \alpha_k e^{j\Omega\psi_k}W_g\left(\Omega + \frac{2\pi k}{P}\right)\end{bmatrix} \tag{3.6}$$

where $W_g(\Omega)$ is the Fourier transform of $w_g(\omega)$. Analogous to the argument made for *short-time analysis* in which we sought to have minimal interaction between the $w(\omega)$ terms, $w_g(\omega)$ should

be set to at least 2~3 times $\frac{2\pi}{P}$. To account for the extremal case of 350 Hz (with *closest* spacing of the impulses in the GCT domain), this constrains $w_g(\omega)$ to be between $(2)(350) = 700$ to $(3)(350) = 1050$ Hz.

1-D Vocal Tract Model: In the source-filter framework of voiced speech, the source signal is convolved with the impulse response of the formant structure $h[n]$ and glottal flow component $g[n]$

$$s[n] = h[n] * p[n] * g[n]. \tag{3.7}$$

In short-time analysis, $s[n]$ is analyzed using the window $w[n]$, i.e.,

$$s_w[n] = s[n]w[n] = (h[n] * p[n] * g[n])w[n] \tag{3.8}$$

$$s_w[n] \approx (h[n] * g[n]) * p_w[n] \tag{3.9}$$

where $p_w[n]$ is defined as in (3.2); the latter step is obtained by assuming that $w[n]$ varies slowly relative to $h[n] * g[n]$ as in [2]. The short-time spectrum magnitude $s_w(\omega)$ of the signal is

$$s_w(\omega) = p_w(\omega)a(\omega) \tag{3.10}$$

$$a(\omega) = h(\omega)g(\omega) \tag{3.11}$$

where $p(\omega)$, $h(\omega)$, $g(\omega)$ are the Fourier transform components of $p_w[n]$, $h[n]$, and $g[n]$, respectively. Here, we have combined the glottal flow and formant terms into a single component $a(\omega)$. By substituting (3.3) into (3.10) and applying a window $w_g(\omega)$ along the ω-axis,

$$s_w(\omega) \approx w_g(\omega)a(\omega)\left[D + \sum_{k=1}^{\infty} \alpha_k \cos\left(\frac{2\pi k}{P}\omega + \psi_k\right)\right] \tag{3.12}$$

The above model corresponds to *amplitude modulation* of the sinusoidal-series *carrier* of (3.4) by the *envelope* term $a(\omega)$. The 1-D GCT representation is then

$$S_w(\Omega) = DA_w(\Omega) + 0.5\sum_{k=1}^{\infty}\begin{bmatrix}\alpha_k e^{-j\Omega\psi_k}A_w\left(\Omega - \frac{2\pi k}{P}\right) + \\ \alpha_k e^{-j\Omega\psi_k}A_w\left(\Omega + \frac{2\pi k}{P}\right)\end{bmatrix} \tag{3.13}$$

$$A_w(\Omega) = A(\Omega) *_\Omega W_g(\Omega) \tag{3.14}$$

where $*_\Omega$ denotes convolution along the Ω-axis. For reasons that will subsequently become clear, we denote Ω_A as the "bandwidth" of $A(\Omega)$ such that

$$|A(\Omega)| \approx 0, \qquad |\Omega| > \Omega_A \tag{3.15}$$

Observe that if $\Omega_A < \frac{2\pi}{P} - \Omega_A$, the DC and modulated terms in (3.13) then exhibit minimal overlap with each other and occupy distinct regions along the Ω-axis as shown in Figure 3-3.

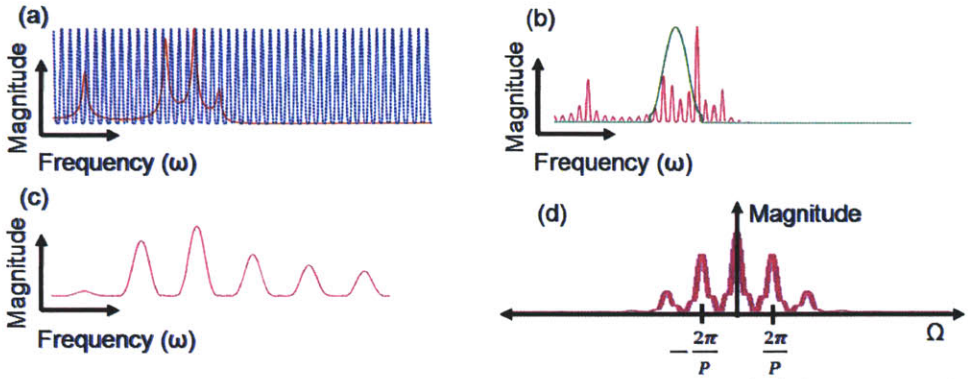

Figure 3-3. (a) Modulation model showing periodic carrier term $p_w(\omega)$ (blue, dotted) being modulated by an envelope term $a(\omega)$ (red, solid) to generate the (b) short-time spectrum $s_w(\omega)$(maroon) and analyzed by a window $w_g(\omega)$ (green) along the ω-axis; (c) local region obtained from (b); (d) 1-D GCT magnitude of (c) indicating replicas of near-DC term at multiples corresponding to the periodicity of the carrier at $\frac{2\pi}{P}$.

2-D Model: Herein we extend the 1-D model of voiced speech to both time and frequency axes of the narrowband spectrogram. In doing so, we incorporate time dependence on the formant structure. Specifically, from the theory of time-varying systems, we denote a time-dependent unit sample response $a[n,m]$ as corresponding to formant and glottal flow characteristics. We use as input $a[n,m]$ a periodic impulse train $p[n]$ as before such that the output is [1]

$$s[n] = \sum_{m=-\infty}^{\infty} a[n,m]p[n-m] \tag{3.16}$$

Analyzing this again within a short-time analysis window, we have

$$s_w[n] = w[n]\sum_{m=-\infty}^{\infty} a[n,m]p[n-m] \tag{3.17}$$

The Fourier transform of $s[n]$ can be shown to correspond to [1]

$$s[n,\omega] = a(n,\omega)\sum_{k=-\infty}^{\infty} \delta(\omega - \tfrac{2\pi}{p}k) \tag{3.18}$$

such that the Fourier transform of $s_w[n,\omega]$ is

$$\begin{aligned} s_w[n,\omega] &= \left(a[n,\omega]\sum_{k=-\infty}^{\infty} \delta[\omega - \tfrac{2\pi}{p}k]\right) *_\omega w[\omega] \\ &= \left(\sum_{k=-\infty}^{\infty} a\left[n, \tfrac{2\pi}{p}k\right]\delta[\omega - \tfrac{2\pi}{p}k]\right) *_\omega w[\omega] \end{aligned} \tag{3.19}$$

If $w[n]$ is chosen to be a "long" window such that $w(\omega)$ has a small bandwidth (e.g., in a narrowband condition), we approximate this effect by

$$\begin{aligned} s_w[n,\omega] &\approx a[n,\omega]\sum_{k=-\infty}^{\infty} w\left(\omega - \tfrac{2\pi}{p}k\right) \\ &\approx a[n,\omega](D + \sum_{k=1}^{\infty} \alpha_k \cos\left(\tfrac{2\pi k}{p}\omega + \psi_k\right)). \end{aligned} \tag{3.20}$$

where $\sum_{k=-\infty}^{\infty} w\left(\omega - \frac{2\pi}{p}k\right)$ is rewritten again a periodic sinusoidal expansion. Equation (3.20) therefore motivates a 2-D modulation model of the spectrogram for time-varying formant structure and stationary pitch, similar to that proposed in the 1-D case.

As a further extension of the model, consider a *localized* time-frequency region centered at n_c and ω_c of $s[n, \omega]$ extracted with a 2-D window $w[n, \omega]$, i.e.,

$$s_w[n, \omega] = w[n, \omega]s[n + n_c, \omega + \omega_c] \tag{3.21}$$

If the source signal $p[n]$ is *time-varying* (i.e., with changing periodicity) we propose a model of the harmonic structure as a 2-D sinusoidal series carrier, i.e.,

$$s_w[n, \omega] \approx a_w[n, \omega][K + \sum_{k=1}^{\infty} \alpha_k \cos(\phi_k[n, \omega])] \tag{3.22}$$

$$\phi_k[n, \omega] = k\Omega_s(n \cos\theta + \omega \sin\theta) + \psi_k \tag{3.23}$$

where Ω_s is the spatial frequency of the 2-D sinusoid and θ represents its orientation in the time-frequency space (Figure 3-4a). The 2-D sinusoidal carrier is motivated from observations that local time-frequency regions of harmonic content resemble *approximately parallel lines* with periodicity related pitch information (i.e., as observed in pitch estimation results with the GCT [7]). Specifically, the pitch value at the center of $s_w[n, \omega]$ in time can be obtained with the mapping

$$f_0 \approx \frac{2\pi f_s}{N_{STFT}\Omega_s \cos\theta} \tag{3.24}$$

This mapping maps the *vertical* distance between harmonic lines in the local time-frequency region to the corresponding vertical distance of the sinusoidal terms in the GCT domain (Figure 3-4).

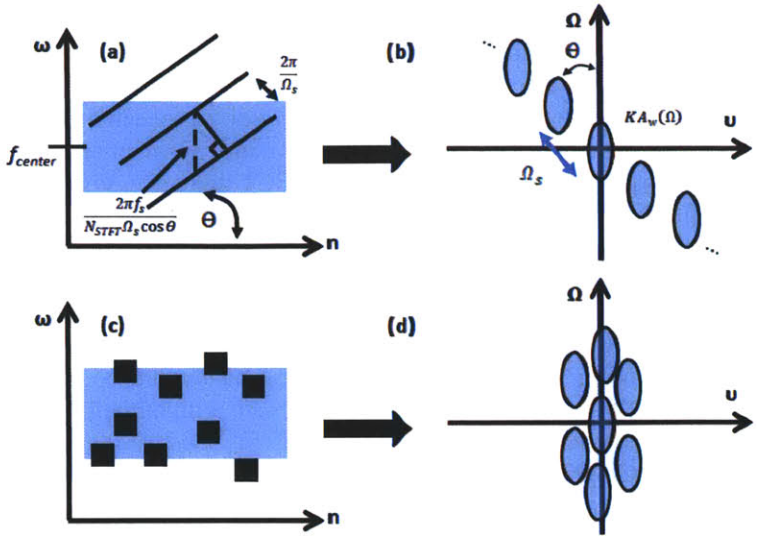

Figure 3-4. (a) Schematic of *localized* region of spectrogram with harmonic lines (solid) modulated by a local envelope (e.g., formant and/or onset/offset) structure (shaded); pitch parameters denoted; (b) GCT of (a); (c) noise structure modulating (solid, squares) envelope (shaded); (d) GCT of (c).

Furthermore, denoting Δf_0 as a change in pitch across the duration of $s_w[n,\omega]$ in time, observe from Figure 3-4b that for the k^{th} pitch harmonic in $s_w[n,\omega]$, the *absolute* change in frequency is $k\,\Delta f_0$ such that

$$\tan\theta \approx \frac{k\Delta f_0}{\Delta n} \tag{3.25}$$

Denoting f_{center} as the center frequency of $s_w[n,\omega]$, the rate of change of pitch can then be estimated as

$$\frac{df_0}{dt} \approx \frac{f_0 \tan\theta}{f_{center}} \tag{3.26}$$

Returning now to the model in (3.18), $a_w[n,\omega]$ corresponds to a windowed portion of $a[n,\omega]$ as a general envelope term. The 2-D carrier represents the harmonic line structure of narrowband spectrogram while the envelope $a_w[n,\omega]$ represents the local spectrotemporal shaping of the spectrogram (e.g., dynamic formant structure). In the GCT domain, dynamic content results in a *rotation* of the envelope component above the origin relative to a stationary formant [19] (see Appendix C). As in the 1-D case, we propose a bandlimited (Ω_{A_2}) approximation of $a_w[n,\omega]$ in which

$$A_w(\upsilon,\Omega) = A_w(\upsilon,\Omega),\ \sqrt{\upsilon^2+\Omega^2} < \Omega_{A_2}\ \approx 0, otherwise \tag{3.27}$$

The GCT is the 2-D Fourier transform of (3.22)

$$S_w(\upsilon,\Omega) = KA_w(\upsilon,\Omega) + 0.5\sum_{k=1}^{\infty}\begin{bmatrix}\alpha_k e^{-j\Omega\psi_k}A_w(\upsilon - k\Omega_s\cos\theta, \Omega + k\Omega_s\sin\theta)\\ +\alpha_k e^{j\Omega\psi_k}A_w(\upsilon + k\Omega_s\cos\theta, \Omega - k\Omega_s\sin\theta)\end{bmatrix} \tag{3.28}$$

where υ and Ω correspond to n and ω, respectively, and the carrier-modulated envelope terms exhibit minimal interaction with each other (Figure 3-4).

Simulations: Herein we evaluate the 2-D signal model for voiced speech on a synthetic signal. In a first set of simulations, we assess the extent to which dynamic formant and pitch content effects the pitch mappings proposed in (3.24) and (3.26) in the GCT domain. As a source signal, we use an impulse train with linearly rising pitch ranging from 150 Hz to 250 Hz across a 0.5-second duration. This corresponds to a moderate pitch change rate of 0.2 Hz/ms that we typically observe in speech. The impulse train is synthesized using the described parameters at an oversampled 64 kHz and then downsampled to 16 kHz; consequently, this results in a signal that is not strictly an impulse train but exhibits harmonic line structure similar to that observed in real speech in regions of changing pitch. To account for the bandwidth constraint of the formant structure (3.15) in *both* the υ and Ω directions, we synthesize a diphthong /ey/ with initial (final) formant frequencies of 669, 2349, 2973, 4000 Hz (437, 2761, 3372, 4000, Hz), and initial (final) bandwidths of 65, 90, 156, 200 Hz (38, 66, 171, 200 Hz) [28]. An adaptive all-pole filter is applied to the source to generate the speech signal.

In analysis, the STFT is computed using a 32-ms Hamming window, 1-ms frame interval, and 512-point discrete Fourier transform (DFT); GCT analysis is performed using region sizes of 875 Hz by 20 ms with an overlap factor of 4 in time and frequency directions. A 1-D low- (high-) pass filter is designed using the frequency sampling method assuming the extremal case of a 350-Hz pitch such that the pass- (stop-) band is from 0 to Ω_A:

$$\Omega_A = 0.5\frac{f_s}{350}\frac{2\pi}{N_{STFT}} = 0.5\frac{16000}{350}\frac{2\pi}{512} = 0.0893\pi \tag{3.29}$$

with a roll-off to $2\Omega_A$ for the stop- (pass-) band [2]. The 1-D filter is rotated to form a circularly symmetric 2-D filter using the frequency transformation method [29]. To assess the effect of the bandwidth constraint (3.15) on the envelope, we *low-pass* filter the spectrogram of the signal. To assess the effect of the 2-D envelope structure on pitch information, we perform peak-picking in the GCT on a *high-pass* filtered version of the spectrogram; we denote the location of this peak in the GCT as $(\widehat{\upsilon_0}, \widehat{\Omega_0})$. Here, we note that filtering can result in time-frequency units of the spectrogram (magnitude) exhibiting *negative* values such that it is no longer strictly a "magnitude"; as will be subsequently demonstrated for filtering and other operations, use of such modified "magnitudes" in approximate reconstruction of the spectrogram and waveforms can nonetheless provide good representations of speech content. To assess the mapping of both pitch and pitch-dynamic information in the GCT and effects of the formant structure, we use (3.24),(3.25), and (3.26), and the center frequency of the time-frequency region analyzed to compute the location of the first harmonic term in the GCT denoted as (υ_0, Ω_0). As an error metric, we compute the distance between these two terms $\varepsilon = \sqrt{(\widehat{\upsilon_0} - \upsilon_0)^2 + (\widehat{\Omega_0} - \Omega_0)^2}$.

We show in Figure 3-5 results of our analyses. Observe in Figure 3-5b that errors in the GCT mapping of pitch information increase with frequency up to ~0.1π for frequency regions $\omega <$ ~0.5π; this effect is presumably due to the spectrogram exhibiting *fanned* harmonic line structure for changing pitch in general that can only be *approximated* as parallel lines within local time-frequency regions [16]. Observe that fanning is more severe in high-frequency regions and consistent with this argument. In addition, the fanning effect appears to be more severe in lower-pitch values (as evidenced by larger errors at the beginning of the vowel), presumably due to the presence of more harmonics within the local time-frequency analyzed. For $\omega >$ ~0.7π (not shown), we have observed substantial errors in the mapping up to ~π; this is due to low-amplitude formant structure suppressing harmonic content in these frequency regions as can be

observed in Figure 3-5a. Finally, in Figure 4c, we show the low-pass filtered spectrogram to be compared with the true envelope spectrogram of Figure 3-5d. We scale both spectrograms such that their maximum values are 1 and compute a root-mean-squared error (RMSE) of ~0.068 between the two. Qualitatively, observe that while formant structure is generally maintained from low-pass filtering, a widening of the bandwidths occurs from filtering. The low RMSE is consistent with previous efforts in [19] that quantitatively demonstrated the GCT's utility in obtaining improved spectral representations for formant estimation. Bandwidth widening can be expected due to the bandwidth constraint of the 2-D envelope from (3.15).

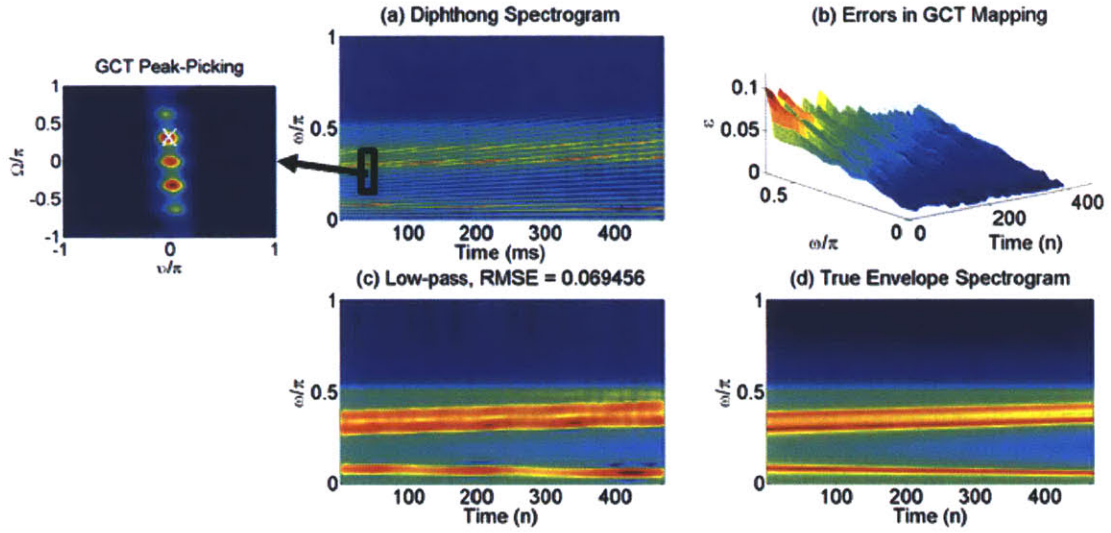

Figure 3-5. (a) Spectrogram computed for diphthong vowel, localized region (rectangle), and GCT-based peak-picking (inset; 'X'); near-origin terms in GCT ignored; (b) Radial errors of peak-picking analysis from (a); (c) low-pass filtered version of (a); (d) true formant envelope.

In a second simulation, we synthesize a 200-ms diphthong with start-to-end formant frequencies (bandwidths) of 437 2761, 3372, 4000 Hz (38, 66, 171, 200 Hz) to 669, 2349, 2972, 4000 Hz (65, 90, 156, 200 Hz). The source signal is a pure impulse train with decreasing 200-Hz pitch of -0.2 Hz/ms. The spectrogram is computed using parameters as in the previous section. We show in Figure 3-6 analysis of the diphthong in a local region of the increasing second formant. Figure 3-6c shows the WGCT of this region; for display purposes, the DC value is removed prior to computing the WGCT. Observe that the near-DC terms are *rotated* at an angle relative to the Ω-axis consistent with the increasing formant frequency; replicas of these near-DC terms are located at carrier positions reflecting the carrier as well. In Figure 3-6d, we show the results of *demodulating* the two dominant peaks in Figure 3-6c to DC; briefly, the local region is multiplied by a 2-D sinusoid generated from the parameters of the carrier corresponding to the first harmonic in the GCT domain. For further details of the method, we refer the reader to Section 3.2.1. In the result, we restrict our display to the near-DC regions of the resulting WGCT due to the presence of cross terms obtained in demodulation. Observe here that a set of *rotated* components are obtained at DC to match those in Figure 3-6c, consistent with the modulation model. Quantitatively, we compute the angle of the dominant peaks ('Theta') in the GCT in both Figure 3-6c and Figure 3-6d showing that the demodulated terms exhibit close correspondence in terms of orientation in relation to the original near-DC terms.

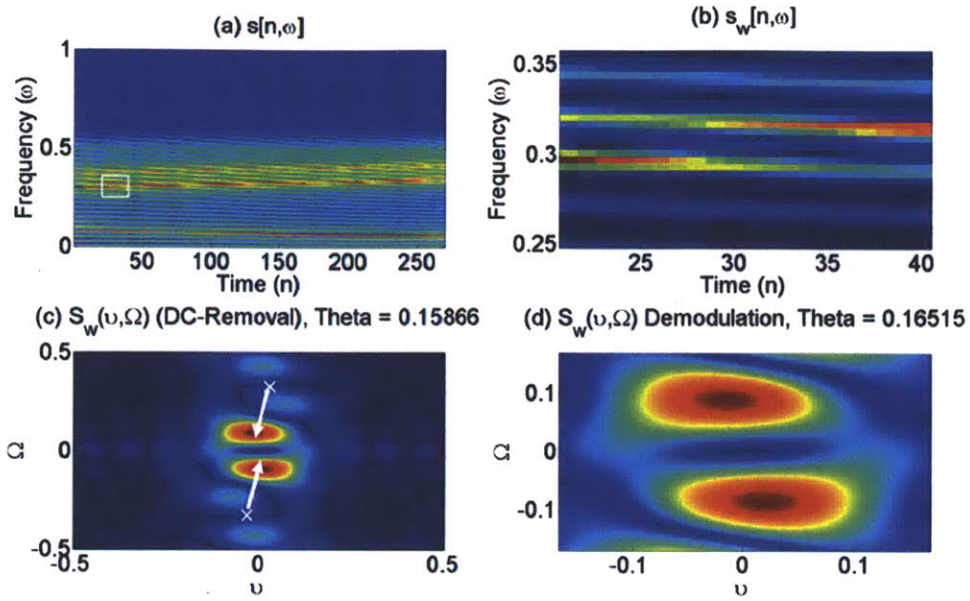

Figure 3-6. (a) Narrowband spectrogram of diphthong with local region (white); (b) local region of (a); (c) GCT of (b) with *rotated* (white line) envelope structure near origin; arrows denote demodulation of carrier terms down to DC; white 'x' denotes carrier position used in demodulation; (d) WGCT of *demodulated* version of (c) with comparable rotated components to match that in (c). In (c) and (d), DC value is removed for illustrative purposes; in (d), display limited to near-DC region due to presence of cross terms in demodulation. Theta reflects the angle of the components computed describing their orientation in (c-d) (counterclockwise relative to Ω-axis).

3.1.2 Noise Model

2-D Model of Average Behavior: This section considers modeling of noise content (e.g., fricatives) in the short-time Fourier spectral magnitude and GCT domains. Consider a zero-mean independent and identically distributed (i.i.d.) Gaussian process $\boldsymbol{w}[n]$ with standard deviation σ. Denoting the short-time Fourier transform of a realization of $\boldsymbol{w}[n]$ as $\boldsymbol{w}[n,\omega]$, we assume that 1) each time-frequency unit in $\boldsymbol{w}[n,\omega]$ is statistically independent, and 2) the complex and real components of a single time-frequency unit are independent and zero-mean, σ^2-variance distributed Gaussian random variables. Under these assumptions, we view the *magnitude* of $\boldsymbol{w}[n,\omega]$ as an i.i.d. Rayleigh process along both the n- and ω-axes since it is the magnitude of two independent Gaussian random variables [30][31]. $|\boldsymbol{w}[n,\omega]|$ can be characterized by its autocorrelation function $r_{ww}[n',\omega']$ from properties of the Rayleigh distribution, i.e.,

$$\begin{aligned} r_{ww}[n',\omega'] &= E[|\boldsymbol{w}[n,\omega]||\boldsymbol{w}[n+n',\omega+\omega']|] \\ &= \tfrac{4-\pi}{2}\sigma^2\delta[n',\omega'] + \tfrac{\pi}{2}\sigma^2. \end{aligned} \tag{3.30}$$

The power spectrum $S_{ww}(\upsilon,\Omega)$ in the GCT is therefore

$$S_{ww}(\upsilon,\Omega) = \tfrac{4-\pi}{2}\sigma^2 + \tfrac{\pi}{2}\sigma^2\delta(\upsilon,\Omega). \tag{3.31}$$

In GCT analysis, localized time-frequency regions are extracted using a 2-D window $w[n,\omega]$. From [2], we extend an analogous 1-D spectral analysis result such that the power spectrum in GCT analysis is a biased/smoothed version of (3.31), i.e.,

$$S_{ww,GCT}(\upsilon,\Omega) = \left[\frac{4-\pi}{2}\sigma^2 + \frac{\pi}{2}\sigma^2\delta(\upsilon,\Omega)\right] *_{\upsilon,\Omega} |W(\upsilon,\Omega)|^2 \\ = \frac{\pi}{2}\sigma^2|W(\upsilon,\Omega)|^2 + \frac{4-\pi}{2}\sigma^2\rho \tag{3.32}$$

$$\rho = \iint_{(-\pi,-\pi)}^{(\pi,\pi)} |W(\upsilon,\Omega)|^2 d\upsilon d\,\Omega \tag{3.33}$$

where $W(\upsilon,\Omega)$ is the 2-D Fourier transform of $w[n,\omega]$. Our analysis indicates that noise content is *distributed* across the entire GCT space along with a dominant term near the GCT origin.

2-D Model of Instantaneous Behavior: To model the instantaneous behavior of noise in each local time-frequency region, we assume that noise results in a general 2-D random process across the full spectrogram (denoted as $E[n,\omega]$). If a local region extracted with a 2-D window has size in time and frequency (denoted as $E_w[n,\omega]$) such that in the GCT domain, it has sufficiently "narrow" bandwidth, then the output of the filters to the input noise term will be *uncorrelated* since the filters will occupy distinct regions in the GCT [32]. We may therefore interpret these components using a Karhunen-Loeve expansion via any orthogonal basis set. As a choice of basis, we pick a set of sinusoids corresponding in frequency to each complex exponential used in computing the discrete-Fourier transform. Extracting a *subset* of these sinusoids based on their amplitudes in the GCT domain (e.g., through peak-picking) results in an *arbitrary* set of sinusoids in modeling $E_w[n,\omega]$; we can expect that selecting those components with the largest amplitudes (e.g., largest KL expansion coefficients) would give a good approximation such that

$$E[n,\omega] \approx D + \sum_{k=1}^{\infty} \alpha_k \cos\phi_k[n,\omega] \tag{3.34}$$

$$\phi_k[n,\omega] = \Omega_k(n\cos\theta_k + \omega\sin\theta_k) + \psi_k. \tag{3.35}$$

Ω_k, θ_k, ψ_k, and α_k again correspond to spatial frequencies, orientations, phases, and amplitudes of the 2-D sinusoids, and the D term corresponds to the DC component.

To incorporate the previous model of noise in relation to speech, consider $e[n]$ as a *realization* of an i.i.d. white Gaussian noise process $\boldsymbol{w}[n]$ such that it may be viewed as a deterministic signal. We consider $e[n]$ exciting (stationary) formant structure $h[n]$ to generate the noisy speech signal $s[n]$, i.e.,

$$s[n] = e[n] * h[n] \tag{3.36}$$

Within a local analysis window $w[n-n_0]$ beginning at time n_0, the resulting signal is

$$s_{n_0}[n] \approx w[n-n_0](h[n] * e[n]) \approx (w[n-n_0]e[n]) * h[n] \tag{3.37}$$

with corresponding Fourier transform

$$s[n_0,\omega] = H[\omega]E[n_0,\omega] \tag{3.38}$$

$$E[n_0,\omega] = e[\omega] *_{\omega} W[\omega]e^{-j\omega n_0} \tag{3.39}$$

where the windowed signal is approximated by the *windowed noise realization* convolved with $h[n]$ for $w[n]$ varying slower than $h[n]$ as in (3.9). The magnitude of this becomes

$$|s[n_0,\omega]| = |H[\omega]E[n_0,\omega]| = |H[\omega]|\left|e[\omega] * W[\omega]e^{-j\omega n_0}\right| \tag{3.40}$$

$$|s[n_0,\omega]| = |H[\omega]E[n_0,\omega]| = |H[\omega]|\left(D + \sum_{k=1}^{\infty}\alpha_k \cos\phi_k[n,\omega]\right) \tag{3.41}$$

$\left|e[\omega] * W[\omega]e^{-j\omega n_0}\right|$ is the magnitude spectrogram computed using a window $w[n]$ of the realization $e[n]$. As previously argued, we approximate this as a sum of arbitrarily spaced 2-D sinusoids. An analogous argument can be used to incorporate time-dependence in $H[\omega]$ (i.e., $H[n_0,\omega]$) if it is assumed that each windowed segment is the result of convolving an approximately stationary $h[n;n_0]$ with $e_w[n]$. The present development therefore argues for speech exhibiting noise components (e.g., a fricative) noise comprised of a carrier term in the form of arbitrarily spaced sinusoids *modulated* by an envelope term reflecting spectral shaping effects. The corresponding GCT can be shown to be analogous to that of the voiced case (3.28) though with carrier positions that are not harmonically related. The GCT will therefore similarly exhibit a distribution of envelope content as a function of the carrier parameters as schematized in Figure 3-4c-d.

Simulations: In describing the *average* behavior of noise in the GCT through models, we invoked assumptions of independent spectral magnitude values across both time and frequency in the spectrogram; this condition can be partially obtained if short-time and GCT processing is done with analysis windows that are non-overlapping. As a simulation, we compute power spectral density (PSD) estimates under such idealized conditions of non-overlapping short-time analysis windows *and* non-overlapping 2-D local time-frequency regions. Specifically, a spectrogram magnitude computed using a 32-ms window with no overlap and 512-point discrete-Fourier transform was analyzed using nonoverlapping local time-frequency region sizes of 20 ms by 875 Hz. In averaging the squared GCT magnitudes for each time-frequency region, we obtain a PSD estimate in the GCT domain. As a quantitative metric, we scale both the estimate and ideal PSDs to have a maximum value of unity to account for scaling effects and compute the root-mean-squared-error (RMSE) between them (Figure 3-7). Consistent with a good match of the estimate to the model, we observe an RMSE of 5e-4.

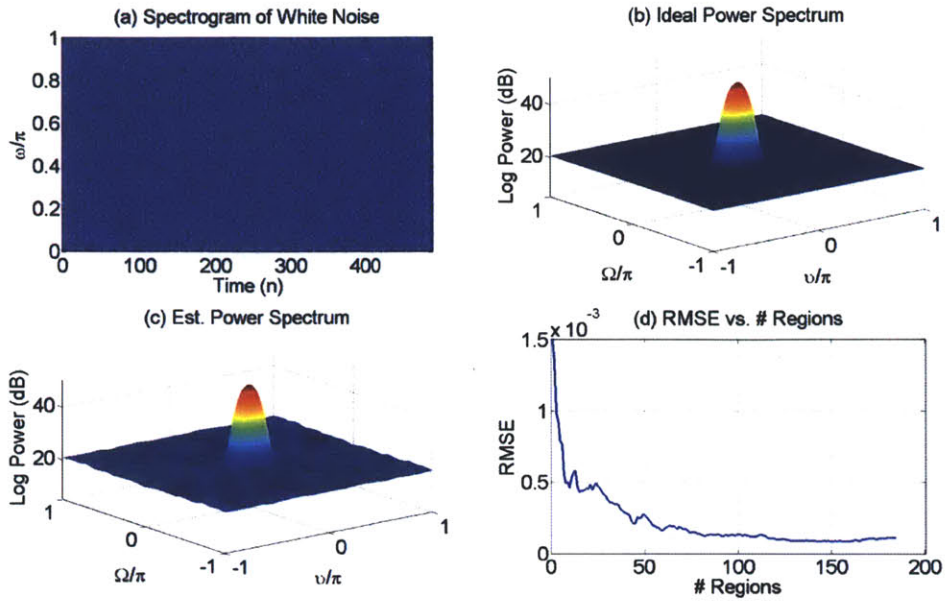

Figure 3-7. (a) Spectrogram of Gaussian white noise computed using non-overlapping window in short-time and GCT analysis; (b) ideal power spectrum; (c) estimated power spectrum from averaging; (d) RMSE vs. number of regions averaged after normalizing estimate and ideal to have maximum value of unity.

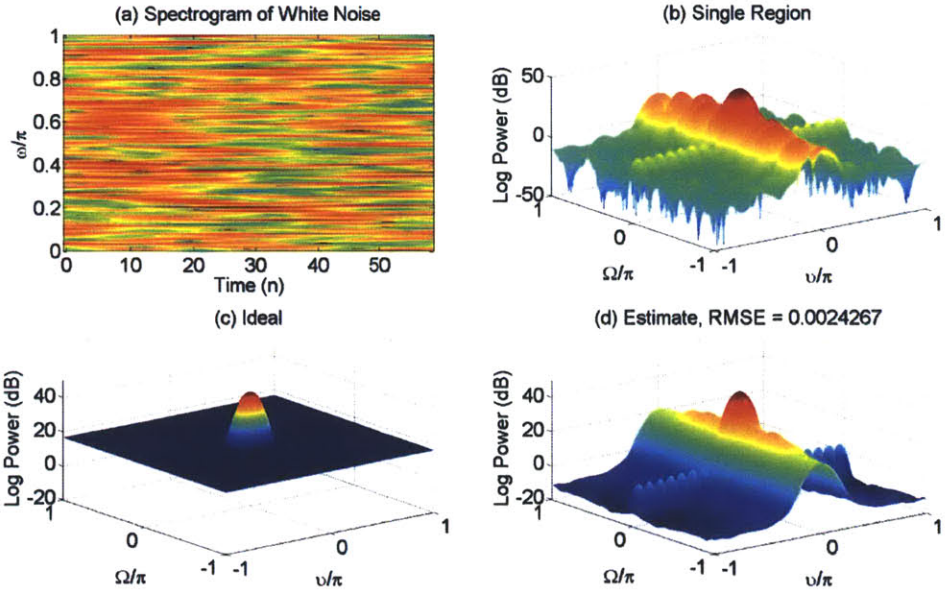

Figure 3-8. (a) Spectrogram of Gaussian white noise (log scale) computed using overlapping window; (b) Power spectrum in GCT from a single time-frequency region; (c) Ideal power spectrum; (d) estimated power spectrum from averaging.

Under typical processing conditions of the GCT, we use short-time and GCT analysis parameters with substantial overlap as described in the previous section. To assess the effect that such overlap has on the average noise representation, Figure 3-8 shows a comparison between the ideal and estimated PSD for a spectrogram and GCT computed as in Section 3.1.1. Though the model is able to capture the dominance of the DC term, observe that it fails to capture substantial spectral shaping effects in the GCT such as peaks concentrated along Ω-axis (Figure 3-8d). This is consistent with the presence of horizontal striations in the spectrogram presumably due to temporal correlation effects in short-time analysis. Figure 3-8b shows results of GCT analysis on a single region, consistent with the average behavior of Figure 3-8d. Quantitatively, this results in a RMSE (after scaling both to have maximum value of unity) of 2.42e-3, a fourfold increase relative to the ideal case. This RMSE value is nonetheless "small" relative unity, consistent with the relatively dominant near-DC component in both estimate and ideal PSDs. Observe that the near-DC term in the estimate is ~20 dB *greater* than all spectral shaping effects along the Ω-axis.

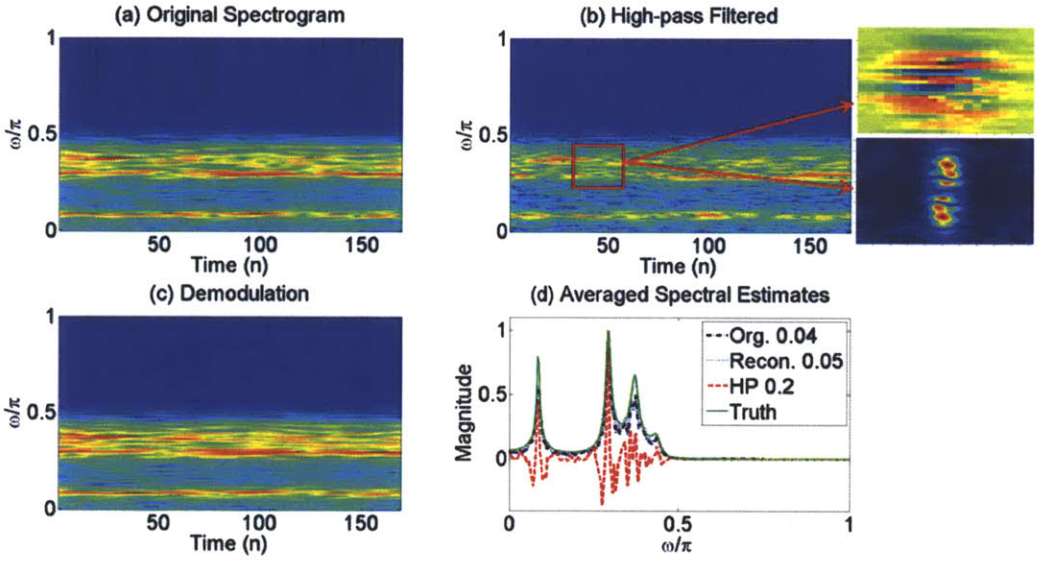

Figure 3-9. (a) Spectrogram of vowel excited by Gaussian white noise; (b) high-pass filtered version of (a) with localized region (red); inset shows local region (top) and corresponding GCT magnitude with near-DC region removed for display purposes; model invokes a distribution of envelope content at locations corresponding to the noise carrier; (c) reconstructed spectrogram; (d) averaged spectra and associated RMSE values.

The modulation-based model proposed for speech that is "noisy" (e.g., a fricative) is proposed under assumptions of "narrow" bandwidths in the GCT relative to the 2-D window used to extract local time-frequency regions. Herein we evaluate the ability of the modulation model to represent speech that is noisy. Specifically, consider a vowel formant structure with formant (bandwidth) frequencies of 669, 2349, 2972, and 3500 (65, 90, 156, and 200) Hz excited by Gaussian white noise [28]. Based on the modulation model, we expect that sinusoidal noise carrier locations observed in the GCT spectrogram as "peaks" would correspond to *modulated* versions of the underlying formant envelope. To test this hypothesis, aim to reconstruct the spectrogram by *demodulation* of envelope copies located at carrier locations. We refer the reader to Section 3.2.1 for details of this method and focus on its results. We emphasize, however, that our approach extracts only a *subset* of all peaks observed in the GCT for use as carriers in demodulation as determined by a peak-amplitude threshold in the GCT domain (Section 3.2.2).

Consequently, the method does not correspond to the trivial reconstruction condition of a simple 2-D Fourier transform inversion.

Observe from Figure 3-9 that the reconstructed spectrogram matches closely to the original indicating good representation of noise content with the 2-D sinusoidal carriers. Furthermore, in Figure 3-9, we show the results of averaging across time all the spectral slices of the spectrograms. While the average of both the reconstructed and original spectrogram result in a spectral estimate very closely matched to that of the true formant envelope, as can be expected, the average of the high-pass filtered spectrogram does *not* yield this result due to the removal of the near-origin terms of the GCT. Root-mean-squared errors are computed on the raw spectra without normalization and are shown in Figure 3-9d; these results are consistent with our qualitative observations. Our results demonstrate empirically that the modulation framework provides that a way to interpret speech that is noisy using a 2-D carrier with arbitrarily spaced sinusoids modulated by an envelope structure.

3.1.3 Onsets/Offsets

***2-D Model*:** Herein we model vertical edges observed in the spectrograms of speech (e.g., plosives) corresponding to onsets/offsets. Consider an isolated impulse $i[n]$ located at N_0

$$i[n] = \delta[n - N_0]. \tag{3.42}$$

The short-time Fourier transform of $i[n]$ is computed using a shifted Hamming window $w_m[n]$

$$I[n, \omega] = \sum_{m=-\infty}^{\infty} i[m] w_m[m - nN] e^{-j\omega n} \tag{3.43}$$

where N corresponds to the shift of the window in analysis. The STFT magnitude may be viewed as a sampled Hamming window across time, i.e.,

$$|I[n, \omega]| = \left|\sum_{m=-\infty}^{\infty} \delta[m - N_0] w_m[N_0 - nN] e^{-j\omega N_0}\right| \tag{3.44}$$

$$|I[n, \omega]| = |I[n]| = w_m[N_0 - nN] \tag{3.45}$$

where N corresponds to the sampling rate of the window and corresponds to the frame rate of the STFT. In the GCT domain, a time-domain impulse corresponds to the 2-D Fourier transform of a downsampled Hamming window. GCT analysis in a local region using a 2-D window $w[n, \omega]$ such that

$$I(\upsilon, \Omega) = W(\upsilon, \Omega) *_{\upsilon} W_m{}^* \left(\frac{\upsilon}{N}\right) e^{j\upsilon N_0} \tag{3.46}$$

where $*_{\upsilon}$ denotes convolution along the υ-axis and $W(\upsilon, \Omega)$ is the 2-D Fourier transform of $w[n, \omega]$. The $W_m{}^* \left(\frac{\upsilon}{N}\right)$ term in (3.46) will therefore have a (Hamming window) main lobe accompanied with side lobe structure along the υ-axis; the presence of $W(\upsilon, \Omega)$ will additionally expand the bandwidth of $I(\upsilon, \Omega)$ based on the 2-D bandwidth of the main lobe of $w[n, \omega]$.

In the context of real speech, onset/offsets generally occur in the presence of either voicing or noise content (e.g., a voicing onset, a stop burst) [28]. In analyzing a region near an onset with a local window, we may therefore obtain only a *portion* of the onset term in time followed by a flat envelope upon entering the onset

$$O[n] = R[n]|I[n - n_0]| + R[n - n_0 - w_i] \tag{3.47}$$

$$R[n] = 1, 0 \leq n \leq n_0 + w_i$$
$$= 0, otherwise \quad (3.48)$$

where w_i corresponds to some value less than the length of $|I[n - n_0]|$ (Figure 3-10). Since $O[n]$ is only a function of n, we can expect the resulting GCT representation to be similarly concentrated along the υ-axis as in the ideal impulse case. We can view onset/offsets as an *envelope* term $A_w(\upsilon, \Omega)$ in (3.28) with the associated bandwidth constraints as in (3.15) for formant structure. This envelope can be modulated by noise (e.g., plosive burst) or harmonic carriers (e.g., voicing onset) represented by a sinusoidal series carrier.

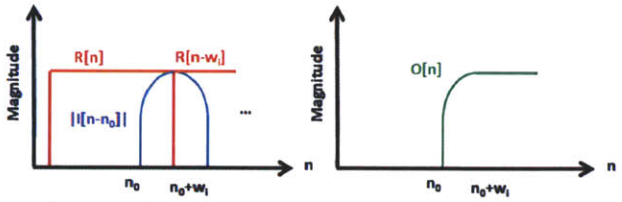

Figure 3-10. Schematic in time for generation (left) of resulting onset envelope (right) term including a voiced/noise onset; here w_i is chosen to be exactly half of an onset. Beyond (prior to) $n_0 + w_i$, harmonic/noise structure is present (absent) and is viewed as the carrier component modulated by $O[n]$.

Simulations: Figure 3-11 shows results of synthesizing and reconstructing voicing and noise onset/offsets using demodulation. Specifically, in the GCT domain, we approximately reconstruct the spectrogram by *demodulation* of envelope copies located at envelope locations as a function of carrier positions. We refer the reader to Section 3.2.1 for details of this method and focus on its results as in the noise case. The reconstruction in Figure 3-11b exhibits widening of the onsets as may be expected from the bandlimited nature of the analysis/synthesis method. Nonetheless, this widening is consistent with the envelope obtained in low-pass filtering the original signal in Figure 3-11c and as can be shown in filtering the reconstruction in Figure 3-11d. RMSE values are computed after scaling the estimate and original spectrograms to have maximum value of unity and are consistent with our qualitative observations.

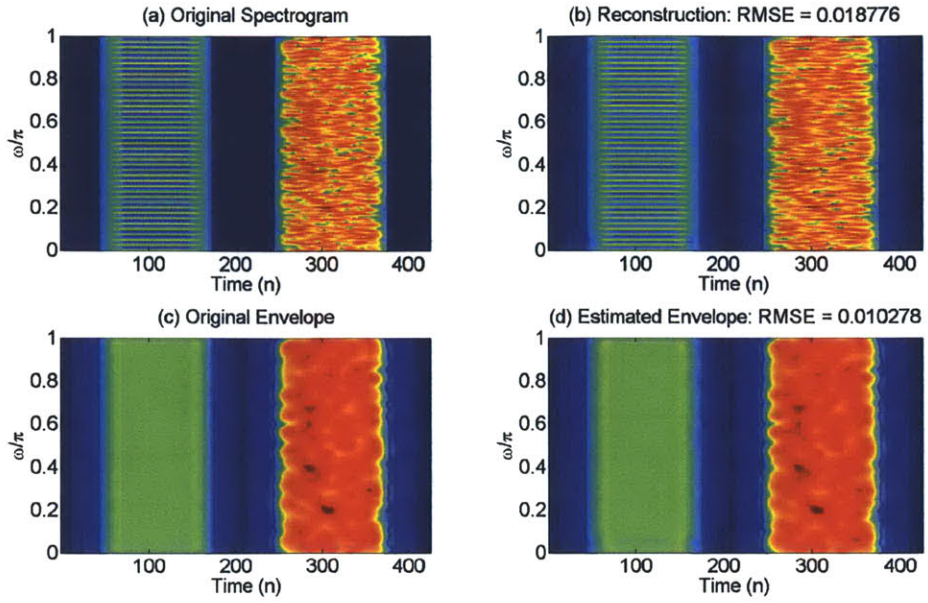

Figure 3-11. (a) Spectrogram of voicing and noise onset/offset; (b) reconstruction of (a); (c) low-pass filtered version of (a) demonstrating onset/offset envelopes; (d) as in (c) but for the reconstruction in (b); associated RMSEs computed after normalization in all cases.

3.2 Spectrogram Analysis/Synthesis

Herein we describe algorithms for analysis/synthesis of the spectrogram to assess the utility of the model in representing real speech. Overall, the algorithm consists of analyzing and synthesizing local time-frequency regions of the spectrogram followed by overlap-add to obtain a spectrogram estimate; an "upper limit" of waveform reconstruction is obtained by combining the spectrogram estimate with the phase of the original signal (Figure 3-12).

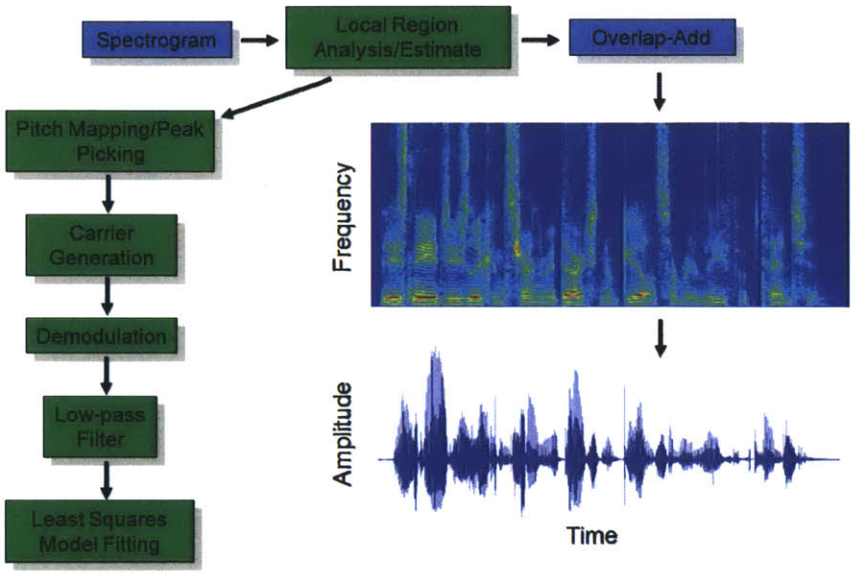

Figure 3-12. Flow diagram illustrating analysis/synthesis methodology.

3.2.1 Framework

Consider a narrowband spectrogram magnitude $s_{full}[n,m]$ computed for an utterance spoken by a single speaker. Furthermore, recall that the GCT is assumed to exhibit concentrated terms near the origin corresponding to an envelope. This occurs for voiced and unvoiced (e.g., noise source) speech exhibiting vowel formant structure as well as onsets/offsets.

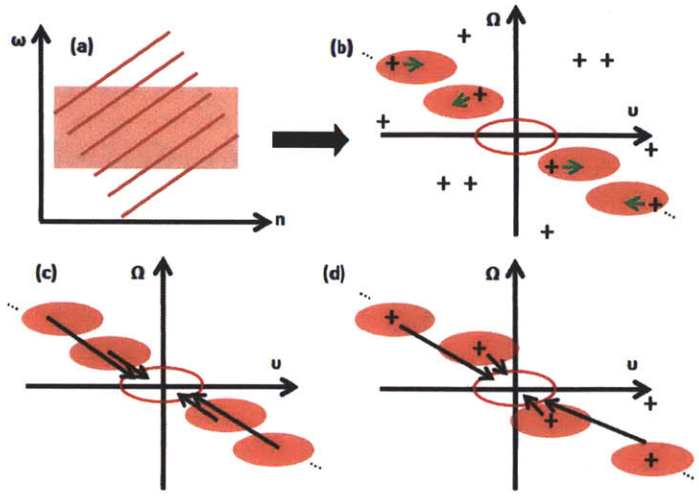

Figure 3-13. (a) Single speaker in voiced region with harmonic (lines) and formant structure (shaded). (b) GCT representation with formant envelope at origin denoted (hollow) as the aim for reconstruction and hypothesized carrier terms; potential carrier locations from peak-picking ('+'); reassignment of hypothesized carrier term locations (arrow, green); (c) demodulation (direct) by hypothesized carriers (shaded, arrows) to recover envelope at GCT origin (hollow); (d) demodulation (bootstrap) using reassigned carrier locations.

We adopt the experimental framework of [21] in estimating these terms using their replicas located at distinct carrier positions through demodulation[4] (Figure 3-13, Figure 3-13). As an algorithmic convenience in anticipation of peak-picking in the GCT domain, we first apply a 2-D high-pass filter $h_{hp}[n,\omega]$ to $s_{full}[n,\omega]$ to obtain $s_{hp}[n,\omega]$ and aim to estimate $s_{full}[n,\omega]$ (denoted as $\hat{s}_{full}[n,\omega]$). We obtain $\hat{s}_{full}[n,\omega]$ using 2-D overlap-add (OLA) that combines estimates of each *localized* region from GCT analysis under a least-squared error (LSE) constraint in 2-D. Denoting $w[n,m]$ as the 2-D window used in GCT analysis,

$$\hat{s}_{full}[n,\omega] = \frac{\sum_m \sum_l w[Tm-n,Fl-\omega]\hat{s}_{ml}[n,\omega]}{\sum_m \sum_l w_{ml}{}^2[Tm-n,Fl-\omega]} \tag{3.49}$$

where T and F (m and l) denote the indices (step sizes) across the time and frequency dimensions, and $\hat{s}_{ml}[n,\omega]$ corresponds to the estimate of a *local* region in GCT analysis.

3.2.2 Estimation of a Single Local Region

The $\hat{s}_{ml}[n,\omega]$ are obtained from sinusoidal-series based demodulation and fitting using each local region of the original and high-pass filtered spectrograms denoted as $s_{ml}[n,\omega]$ and $s_{hp,ml}[n,\omega]$, respectively. In Figure 3-14 we show a flow graph of the estimation process for a

[4] We refer to "demodulation" here as demodulation (i.e., *multiplication*) by a sinusoidal carrier as in standard amplitude demodulation.

single local time-frequency region consisting of an envelope estimation and least-squared error (LSE) fitting procedure. For clarity of discussion of the entire estimation process, we describe first the overall algorithm assuming the carrier parameters from the carrier estimation step (green, Figure 3-14). Subsequently, we discuss this component in detail.

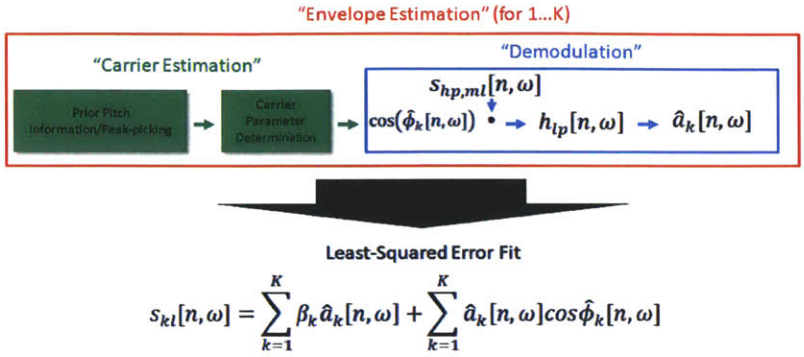

Figure 3-14. Estimation of a single local region consisting of an envelope estimation (red) step that is performed for each carrier position; the envelope estimation step consists of generating the carrier parameters from prior pitch information and/or peak-picking (green) and a demodulation step consisting of synthesizing the sinusoidal carrier, multiplying it by the high-pass filtered local region, and low-pass filtering. The collected set of envelopes are used in a least-squared error fit (black).

Overall Algorithm: In envelope estimation, carrier parameters for a single sinusoidal carrier are determined using prior pitch and pitch-dynamic information (for voiced speech) and/or peak-picking (e.g., for noise carriers) (green, Figure 3-14); subsequently, a demodulation step consists of synthesizing a single sinusoidal carrier $\cos(\hat{\phi}_k[n,\omega])$ using these parameters, multiplying it by the high-pass filtered local region $s_{hp,ml}[n,m]$, and low-pass filtering by $h_{lp}[n,\omega]$ to generate a single envelope estimate $\hat{a}_k[n,\omega]$ (blue, Figure 3-14), i.e.,

$$\hat{a}_k[n,\omega] = h_{lp}[n,\omega] *_{n,\omega} \left[s_{hp,ml}[n,m]\cos(\hat{\phi}_k[n,\omega])\right] \tag{3.50}$$

For interpretative purposes in the demodulation step summarized by (3.50), the GCT (i.e., 2-D Fourier transform) of (3.50) is

$$\hat{A}_k(\upsilon,\Omega) = H_{lp}(\upsilon,\Omega)\left[\begin{array}{c} S_{hp,ml}(\upsilon,\Omega) *_{\upsilon,\Omega} \\ 0.5(e^{-j\psi_k}\delta(\upsilon+\upsilon_k,\Omega-\Omega_k) + e^{j\psi_k}\delta(\upsilon-\upsilon_k,\Omega+\Omega_k)) \end{array}\right] \tag{3.51}$$

where υ_k, Ω_k, and ψ_k are the two spatial frequencies and phase of the 2-D sinusoidal carrier, respectively. In the GCT domain, demodulation corresponds to a convolution of impulses spaced based on carrier parameters with the 2-D Fourier transform of the high-pass filtered local time-frequency region. This can be shown to result in a component at the GCT origin reflecting the replica of the envelope term located at the carrier position ('*' in Figure 3-15) and *cross terms* at twice the spatial frequencies that are replicas of the original near-DC components in $S_{hp,kl}(\upsilon,\Omega)$ (i.e., $2\upsilon_k$ and $2\Omega_k$) ('**' in Figure 3-15) [2]. Due to the presence these cross terms, a low-pass filter $H_{lp}(\upsilon,\Omega)$ is applied to *isolate* the demodulated envelope term at the GCT origin (green rectangle, Figure 3-15). Furthermore, $\hat{a}_k[n,\omega]$ is a bandlimited (in the GCT domain) version of the original envelope content as proposed in the model (3.22).

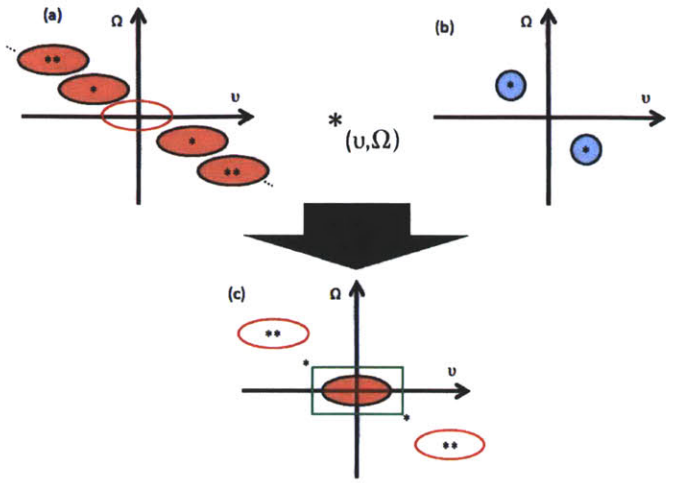

Figure 3-15. Schematic illustrating demodulation steps in the GCT domain to obtain a single $\hat{a}_k[n,\omega]$ term with a convolution of (a) $S_{hp,ml}(v,\Omega)$ with (b) a set of impulses reflecting a single sinusoidal carrier $0.5(e^{-j\psi_k}\delta(v+v_k,\Omega-\Omega_k)+e^{j\psi_k}\delta(v-v_k,\Omega+\Omega_k)$ resulting in (c) a demodulated envelope at the origin; the latter result is low-pass filtered (green) to remove effects of cross terms in demodulation to obtain $\hat{a}_k[n,\omega]$. In (a), filled ovals represent replicas of the envelope located at distinct carrier positions; the original envelope at the GCT origin is partially removed by high-pass filtering (unfilled oval); '*' denotes the carrier position used for generating the carrier in (b) while '' reflects twice the carrier spatial frequencies.**

The resulting set of $\hat{a}_k[n,\omega]$ obtained in performing the envelope estimation steps K times are then used to fit gain parameters β_i in relation to $s_{kl}[n,\omega]$ (i.e., the original local region of $s_{full}[n,\omega]$) in a least-squared error (LSE) procedure (black, Figure 3-14)

$$s_{ml}[n,\omega]=\sum_{k=1}^{K}\beta_k\hat{a}_k[n,\omega]+\sum_{k=1}^{K}\hat{a}_k[n,\omega]cos\hat{\phi}_k[n,\omega] \tag{3.52}$$

where K corresponds to the number of carriers. In matrix form, (3.52) corresponds to $A\underline{\beta}=\underline{b}$, i.e.,

$$A=\begin{bmatrix}\hat{a}_1(1) & \cdots & \hat{a}_K(1)\\ \vdots & \ddots & \vdots\\ \hat{a}_1(Z) & \cdots & \hat{a}_K(Z)\end{bmatrix} \tag{3.53}$$

$$\underline{\beta}=\begin{bmatrix}\beta_1\\ \vdots\\ \beta_K\end{bmatrix},\quad \underline{b}=\begin{bmatrix}s_{ml}(1)-\sum_{k=1}^{K}\hat{a}_k(1)\cos\hat{\phi}_k(1)\\ \vdots\\ s_{ml}(Z)-\sum_{k=1}^{K}\hat{a}_k(Z)\cos\hat{\phi}_k(Z)\end{bmatrix} \tag{3.54}$$

where we have indexed $\hat{a}_k[n,\omega]$ and $\cos\hat{\phi}_k[n,\omega]$ in column form as $\hat{a}_k(z)$ and $\cos\hat{\phi}_k(z)$, and Z corresponds to the total number of points within the local region $s_{kl}[n,\omega]$. (3.52) is overdetermined when $Z > K$ such that we may solve for $\underline{\beta}$ in the least-squared error sense, i.e., $\underline{\hat{\beta}}=(A^TA)^{-1}A^T\underline{b}$.

Substituting the estimated gain terms into the right-hand side of (3.52) results in the estimate $\hat{s}_{ml}[n,\omega]$ of $s_{ml}[n,\omega]$.

(3.52) effectively uses the results of demodulating by *multiple* carriers to solve for the near-GCT-origin terms of the region, thereby exploiting the series-model's *distribution* of copies of the envelope content across the GCT space. Note that this method represents a generalization of the single-sinusoidal demodulation technique presented in [21]. As an example, if higher-order carrier terms corresponding to the envelope are poor estimates, the fitting procedure will weigh these accordingly to minimize their contribution to the final estimate. Finally, as previously noted, the described set of operations in demodulation, filtering, and least-squares fitting can result in negative values of the spectrogram "magnitude" estimate. Despite this limitation, we utilize these estimates directly in evaluating goodness of fit to the true magnitude as well in waveform reconstruction.

Carrier Estimation: For carrier estimation, we first describe a method for obtaining a set of peak locations in the GCT of $s_{hp,kl}[n,\omega]$ ($S_{hp,kl}(\upsilon,\Omega)$). These peaks will be used subsequently in determining carrier positions of both voiced and unvoiced speech for use in demodulation.

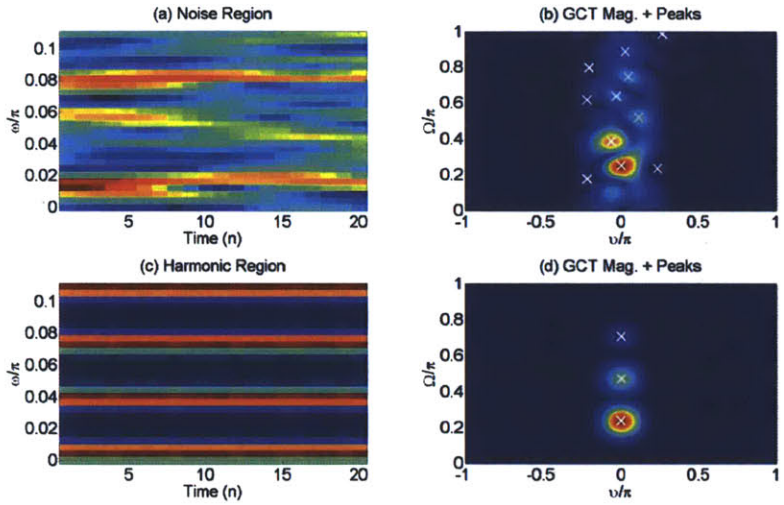

Figure 3-16 (a) Noise region; (b) GCT magnitude of (a) with peaks (white, 'x'); (c) harmonic region ; (d) as in (b) but for (d); for display purposes GCTs are shown only for $0 < \Omega/\pi < 1$.

GCT-domain Peak Picking: The magnitude of $S_{hp,kl}(\upsilon,\Omega)$ is used to obtain a set of peak locations by first computing first-differences between each coordinate's 8 nearest neighbors

$$\left|S_{(\upsilon\pm),(\Omega\pm)}(\upsilon,\Omega)\right| = \left|S_{hp,kl}(\upsilon(\pm d\upsilon),\Omega(\pm d\Omega))\right| - \left|S_{hp,kl}(\upsilon,\Omega)\right| \tag{3.55}$$

where $d\upsilon$ and $d\Omega$ represent step sizes in the discrete representation of $S_{hp,kl}(\upsilon,\Omega)$. The resulting estimates are used to generate individual binary masks, e.g.,

$$B_{(\upsilon\pm),(\Omega\pm)} = \left|S_{(\upsilon\pm),(\Omega\pm)}(\upsilon,\Omega)\right| > 0. \tag{3.56}$$

An additional binary mask is formed by thresholding the magnitude of $\left|S_{hp,kl}(\upsilon,\Omega)\right|$

$$B_{mag} = \left|S_{hp,kl}(\upsilon,\Omega)\right| > \gamma max\left[\left|S_{hp,kl}(\upsilon,\Omega)\right|\right] \tag{3.57}$$

where $\gamma = 10^{-\frac{30}{20}}$, i.e., remove points 30 dB below the maximum value of the GCT magnitude. In addition, a binary mask is formed based on removing peak candidates located in the region $0 < |\Omega| < 2\Omega_A$, where $2\Omega_A = \frac{f_s}{350}\frac{2\pi}{N_{STFT}} = 0.176\pi$ with Ω_A defined as in (3.15), i.e.,

$$B_{\Omega_A}(v,\Omega) = \begin{matrix} 0, for\ 0 < |\Omega| < 2\Omega_A \\ 1, otherwise \end{matrix} \tag{3.58}$$

The latter mask removes all components *above* the maximum pitch value of 350 Hz assumed to be present in the speech analyzed as well as components along the v axis. A final binary mask

$$B_{final} = B_{mag} \cap \left(\cap_{i=1}^{8} B_{v,\Omega}\right) \cap B_{\Omega_A} \tag{3.59}$$

is applied to $|S_{hp,kl}(v,\Omega)|$, where $B_{v,\Omega}$ corresponds to the 8 masks from (3.55). The resulting mask obtains peak locations that are the maximum amongst their 8 nearest neighbors and exhibit a minimal magnitude of 30 dB below the maximum value of the GCT magnitude. For our subsequent discussion, denote this set of locations as the set $P := \{v_{i,P}, \Omega_{i,P}\}\ i = 1, \ldots, N_p$, where N_p is the total number of peaks obtained for the region. Figure 3-16 illustrates the results of this peak-picking in both the noise and harmonic-source case, indicating its ability to obtain harmonic and non-harmonically related peaks.

For demodulation, we consider an *a priori* pitch estimate of the full utterance across time denoted as $f_0(n)$; these estimates are computed such that zero values in $f_0(n)$ denote unvoiced or silent time points. For $s_{hp,kl}[n,\omega]$ ranging from time n_0 to n_1 and centered at f_{center}, we obtain the 2-D carrier frequencies for the voiced case using a direct and bootstrapping technique. The former method is motivated from the proposed signal model while the latter is motivated from observed limitations and properties of the model.

Voiced Case - Direct: When $f_0(n)$ contains *at least* one non-zero pitch value, we view the region as a partially/fully voiced region such that the 2-D sinusoidal parameters can be obtained from $f_0(n)$. Specifically, given the non-zero values of $f_0(n_0)$ through $f_0(n_1)$, we compute their average to represent the pitch value of the entire region ($f_{0,ml}$). Next, we compute one-step differences across the non-zero values of $f_0(n)$ and compute the mode of this set of differences as an estimate of the pitch track slope ($\frac{df_0}{dt}_{ml}$). $f_{0,ml}$, $\frac{df_0}{dt}_{ml}$, and f_{center} are substituted into (3.24) through (3.26) to solve for Ω_s and θ. ψ is computed as

$$\psi = angle\{S_{ml,hp}(v_1,\Omega_1)\} \tag{3.60}$$

$$v_1 = \Omega_s \cos\theta, \Omega_1 = \Omega_s \sin\theta \tag{3.61}$$

v_1 and Ω_1 are then scaled by $k = 1, \ldots, N_D$ where

$$N_D = min\left(floor\left\{\frac{\pi}{v_1}\right\}, floor\left\{\frac{\pi}{\Omega_1}\right\}\right) \tag{3.62}$$

to generate the set $D := \{v_{i,D}, \Omega_{i,D}\}\ i = 1, \ldots, N_D$. The phase parameters $\psi_{i,d}$ are obtained from (3.61) with $v_{i,D}$ and $\Omega_{i,D}$.

The present approach for determining the carrier parameters uses the signal model for voiced speech *directly*. We use this method as a reference to assess the utility of the model in estimates of the original spectrogram despite the limitations highlighted in Section 3.1.1.

Voiced Case - Bootstrapping: As previously shown, the voiced signal model can exhibit errors in the mapping of the pitch and pitch derivative information to the GCT. Motivated from these observations, we propose an alternative *bootstrapping* method for estimating the carrier parameters directly from the GCT using the previously described set of peaks $P := \{v_{i,P}, \Omega_{i,P}\}\ i = 1, \ldots, N_P$. Specifically, we assign $(v_{1,B_0}, \Omega_{1,B_0})$ (i.e., the first harmonic in the GCT domain) as corresponding to the location mapped pitch location using (3.24) and (3.26). A set $B_0 := (v_{1,B_0} i, \Omega_{1,B_0} i)$ for $i = 1, \ldots, N_{B_0}$ is computed by scaling the $(v_{1,B_0}, \Omega_{1,B_0})$ location as in the set D. Next, a distance matrix $D_{B_0-P}(i,j)$ is computed between the sets B_0 and P, i.e.,

$$D_{B_0-P}(i = 1, \ldots, N_{B_0}, j = 1, \ldots, N_P) = \sqrt{(v_{i,B_0} - v_{j,P})^2 + (\Omega_{i,B_0} - \Omega_{j,P})^2}. \quad (3.63)$$

Each location $(v_{i,B_0}, \Omega_{i,B_0})$ in B_0 is then *reassigned* to the closest location in P according to the minimum radial distance (i.e., in D_{B_0-P}) using an iterative algorithm as follows:

initialize A={}
for 1 through N_{B_0}
1) Find the minimal value of $D_{B_0-P}(i,j)\ \forall (i,j) \notin A$.
2) Reassign the i^{th} *hypothesized carrier parameter to the* j^{th} *carrier parameter.*
3) Add all indices of the i^{th} *row and* j^{th} *column of* D_{B_0-P} *to A.* (3.64)

Figure 3-13 schematically shows this algorithm in the GCT space. This method maintains uniqueness of the reassigned locations and restricts the set of carriers in each region to be those estimated from the GCT *itself*. This contrasts the direct approach in which they are obtained from *mapping* the pitch information to the GCT domain. Carrier phase values are obtained by substituting the reassigned locations into (3.61).

Unvoiced Case: If $f_0(n)$ exhibits all zero values, we adopt the noise carrier modulation model (Section 3.1.2) as represented by a sum of non-harmonically related sinusoids. We therefore use all of the locations obtained from peak-picking the GCT as the carrier parameters.

3.3 Reference Approach using Sinusoidal Series Only

The previous section has described a demodulation framework for reconstruction of high-pass filtered spectrograms. Herein we describe an alternative approach that assesses the value of the model in relation to the envelope terms $\hat{a}_k[n,\omega]$ in (3.50). If we assume $\hat{a}_k[n,\omega] = 1$ for all k, the demodulation procedure is in essence fitting a sinusoidal series model to the local time-frequency region, i.e., (3.52) becomes

$$s_{ml}[n,\omega] = \sum_{k=0}^{K} \beta_k \cos\hat{\phi}_k[n,\omega] \quad (3.65)$$

where we use the $k = 0$ term to correspond to an arbitrary DC value. In practice, the local region is windowed such that the fit is performed on windowed versions of the sinusoids. An analogous set of least-squared error equations may be solved to obtain the fit as in the demodulation case and overlap-add may be performed to obtain the spectrogram estimate. The purpose of using this reference method is to distinguish the contribution of the bandwidth of $\hat{a}_k[n,\omega]$ to reconstruction.

In using the sinusoidal series we use carriers obtained from both the direct mappings as well as the bootstrapping methods as in the demodulation techniques.

3.4 Co-channel Speaker Separation

This section describes two algorithms using the proposed signal model for co-channel speaker separation. We emphasize that our aim is in evaluating utility of the signal *model* and assume *a priori* knowledge of the pitch trajectories of individual speakers. To develop a complete separation system, pitch estimates may be obtained from existing multi-pitch estimation methods proposed in the literature [33][34] and those based on the GCT (see Chapter 5, Chapter 6).

Consider a spectrogram computed on an additive mixture of two speakers denoted as $s_{mix,full}[n,\omega]$. We approximate a local region $s_{mix}[n,\omega]$ of $s_{mix,full}[n,\omega]$ as the sum of two distinct models of speech corresponding to each speaker, i.e.,

$$s_{mix}[n,\omega] \approx \sum_{i=1}^{2}\left[a_i[n,\omega](D_i + \sum_{k=1}^{N_i}\alpha_{i,k}\cos\phi_{i,k}[n,\omega]\right] \tag{3.66}$$

$$\phi_{i,k}[n,\omega] = \Omega_{i,k}\left(n\cos\theta_{i,k} + \omega\sin\theta_{i,k}\right) + \psi_{i,k} \tag{3.67}$$

where i corresponds to the i^{th} speaker and k corresponds to the k^{th} term in the sinusoidal series. As in the single-speaker case, $a_i[n,\omega]$, $\cos\phi_{i,k}[n,\omega]$, and $\phi_{i,k}[n,\omega]$ correspond to the envelope, carrier, and carrier parameter terms, respectively, and D_i and $\alpha_{i,k}$ are arbitrary gain terms. The GCT of (3.66) is then

$$S_{mix}(v,\Omega) = \sum_{i=1}^{2}\sum_{k=1}^{N_k}\left(D_iA_i(v,\Omega) + M_{i,k}(v,\Omega)^{+} + M_{i,k}(v,\Omega)^{-}\right) \tag{3.68}$$

$$M_{i,k}(v,\Omega)^{\pm} = 0.5\alpha_{i,k}e^{\pm j\psi_{i,k}}A_{i,k}\left(v \mp \Omega_{i,k}\cos\theta_{i,k}, \Omega \pm \Omega_{i,k}\sin\theta_{i,k}\right) \tag{3.69}$$

Observe from (3.68) that the $D_iA_i(v,\Omega)$ terms can be expected to exhibit overlap at the GCT origin. This multi-speaker model is an approximation and is based on an assumption of linearity in the spectral magnitude domain; as we subsequently show, this approximation can lead to good separation results under certain conditions. Nonetheless, future work aims to explore explicitly limitations of this assumption for improved separation performance as well as the role of phase in the GCT context.

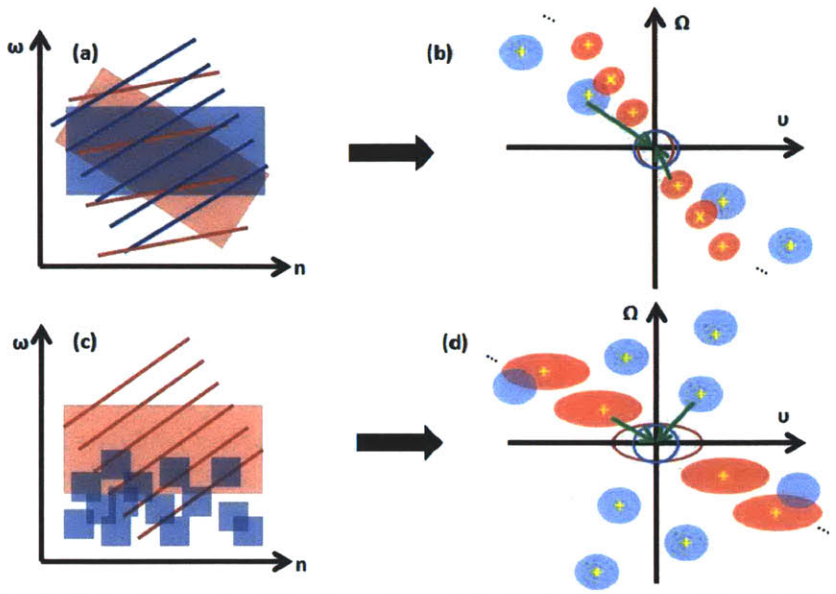

Figure 3-17. (a) Voiced-on-voiced local region with distinct speakers' (red, blue) harmonic (lines) and formant structure (shaded); (b) GCT of (a) showing demodulation (green arrow) of a single term for each speaker; '+' and 'x' are both used in direct method; 'x' is excluded in exclusion method due to a lower harmonic number in the GCT; overlapped and removed near-origin terms are shown as hollow components to be recovered (c) voiced-on-unvoiced region; (d) GCT of (c); noise terms overlapped with voiced carriers are always excluded in demodulation in this case.

Figure 3-17 illustrates schematically the mapping of multi-speaker content in the GCT domain for voiced and unvoiced speech. As shown from (3.68), the $D_i A_i(\upsilon, \Omega)$ of the two speakers will be overlapped at the GCT origin; however, their replicas located at speaker-specific carrier locations are distributed throughout the GCT space and can exhibit separability. An algorithm for speaker separation can then be developed that is similar to performing analysis/synthesis of a single speaker. Specifically, localized time-frequency regions of the *mixture* spectrogram can be modeled using (3.66) as a basis for estimating the envelope and carriers modulated by the envelope for individual speakers. Envelope terms at the GCT origin are estimated from their *replicas* at distinct carrier locations for individual speakers. Performing this across localized regions and combining the estimates using overlap-add as in (3.49) can be used to obtain the spectrogram estimate for individual speakers. Subsequently, we describe two algorithms for computing estimates of individual speakers within local time-frequency regions of the mixture spectrogram.

3.4.1 Direct Method

Figure 3-18 summarizes an approach we refer to as the *direct* method of estimation. Only minor differences exist between this algorithm and that of analysis/synthesis for a single speaker. As in that discussion, we describe first the overall algorithm under the assumption of known carrier parameters and discuss the carrier estimation step subsequently in details.

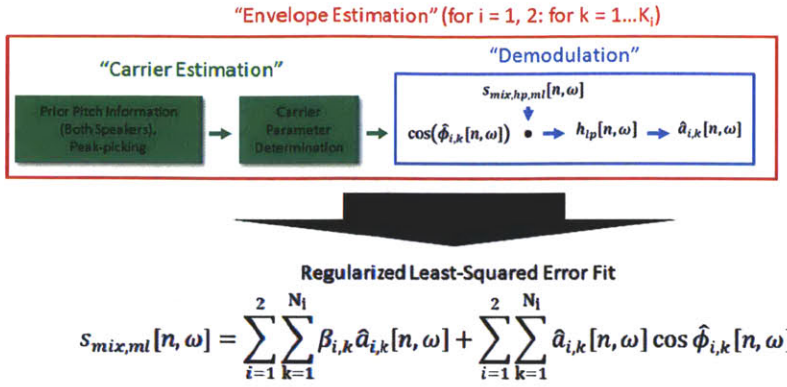

Figure 3-18. Algorithm for estimation of envelope and carriers modulated by envelopes of individual speakers from a mixture spectrogram.

Overall Algorithm: The envelope estimation procedure is performed for each speaker $i = 1,2$ across K_i carrier terms. In the carrier estimation step, prior pitch information from each individual speakers is used to determine the carrier parameters. In demodulation, a local region of the high-pass filtered *mixture* spectrogram (denoted as $s_{mix,hp,ml}[n,\omega]$) is multiplied by the speaker specific sinusoidal carrier $\cos(\hat{\phi}_{i,k}[n,\omega])$ and low-pass filtered by $h_{lp}[n,\omega]$ to obtain a speaker-specific envelope estimate $\hat{a}_{i,k}[n,\omega]$. These envelope estimates are then used in an least-squared error (LSE) fitting procedure fit to the local time-frequency region of the mixture spectrogram (denoted as $s_{mix,ml}[n,\omega]$, i.e.,

$$s_{mix}[n,\omega] = \sum_{i=1}^{2}\sum_{k=1}^{N_i}\beta_{i,k}\hat{a}_{i,k}[n,\omega] + \sum_{i=1}^{2}\sum_{k=1}^{N_i}\hat{a}_{i,k}[n,\omega]\cos\hat{\phi}_{i,k}[n,\omega] \quad (3.70)$$

The resulting $\beta_{i,k}$ gain terms are combined with their respective carrier and envelope estimates to obtain an estimate of speaker i, i.e.,

$$\hat{s}_i[n,\omega] = \sum_{k=1}^{N_i}\beta_{i,k}\hat{a}_{i,k}[n,\omega] + \sum_{k=1}^{N_i}\hat{a}_{i,k}[n,\omega]\cos\hat{\phi}_{i,k}[n,\omega] \quad (3.71)$$

(3.70) represents an analogous system of overdetermined equations as in (3.52). Specifically, we have that

$$s_{env}[n,\omega] = s_{mix}[n,\omega] - \sum_{i=1}^{2}\sum_{k=1}^{N_i}\hat{a}_{i,k}[n,\omega]\cos\hat{\phi}_{i,k}[n,\omega] \approx a_1[n,\omega] + a_2[n,\omega] \quad (3.72)$$

$$s_{env}[n,\omega] = \sum_{i=1}^{2}\sum_{k=1}^{N_i}\beta_{i,k}\hat{a}_{i,k}[n,\omega]. \quad (3.73)$$

For conciseness, we denote A_m, $\underline{b}_m$, and $\underline{\beta}_m$ as the matrix containing all values of $\hat{a}_{i,k}[n,\omega]$, $s_{env}[n,\omega]$, and $\beta_{i,k}$ such that a least-squares formulation of the gain parameters $\beta_{i,k}$ is again $A_m^T A_m \underline{\beta}_m = A_m^T \underline{b}_m$ as in the single speaker case. (3.70) uses multiple envelope representations from demodulation to solve for the *sum* of $a_1[n,\omega]$ and $a_2[n,\omega]$.

Regularized Least Squares Fitting: In contrast to the single-speaker case, the matrix ${A_m}^T A_m$ arising from the least-squares formulation of the equation can become singular/near-singular

under certain conditions. For instance, if both the pitch values *and* dynamic information of the pitch of the two speakers are similar, we can expect to get 2-D carriers that are nearly identical. We detect the extent to which the resulting matrix is singular using

$$c_2 = \frac{|\lambda_{max}|}{|\lambda_{min}|} \tag{3.74}$$

where λ_{max} and λ_{min} are the eigenvalues of ${A_m}^T A_m$ that have maximum and minimum absolute values. If the matrix has a value of c_2 smaller than a threshold γ, we solve the least squares problem directly. Otherwise we solve a *regularized* least squares problem [35]

$$({A_m}^T A_m + \xi I)\underline{\beta}_m = {A_m}^T \underline{b}_m \tag{3.75}$$

$$\xi = \frac{\lambda_{max} - \gamma\lambda_{min}}{\gamma - 1} \tag{3.76}$$

where ξ is a *diagonal loading factor*, and I is the identity matrix, thereby forcing $c_2 = \gamma$ for $({A_m}^T A_m + \xi I)$ and a solution $\hat{\underline{\beta}}_m = ({A_m}^T A_m + \xi I)^{-1} {A_m}^T \underline{b}_m$. The estimated gains are substituted into single speaker models to obtain $\hat{s}_i[n, \omega]$, for each speaker i

$$\hat{s}_i[n, \omega] = \sum_{k=1}^{N_i} \hat{\beta}_{i,k} \hat{a}_{i,k}[n, \omega] + \sum_{k=1}^{N_i} \hat{a}_{i,k}[n, \omega] \cos \hat{\phi}_{i,k}[n, \omega] \tag{3.77}$$

Carrier Estimation: To obtain the carrier parameters for each speaker for use in demodulation, we assume a priori pitch estimates $f_1[n]$ and $f_2[n]$ corresponding to the two speakers across a local region $s_{mix,ml}[n, \omega]$ in time. We consider three cases as determined by $f_1[n]$ and $f_2[n]$ in $s_{mix,hp}[n, \omega]$: voiced-on-voiced, voiced-on-unvoiced, unvoiced-on-unvoiced.

Voiced-on-voiced: If both $f_1[n]$ and $f_2[n]$ exhibit at least one non-zero value of pitch in $s_{mix,hp}[n, \omega]$, we consider the region to be consist of two voiced components. In contrast to the single-speaker case, peak picking in the GCT domain can lead to peaks with ambiguous assignments since 1) the effects of the envelope in the modulation model can shift the location of the peak away from the ideal pitch information mapping and 2) local formant structure from multiple speakers *overlaps/interacts* for carriers that are located close to each other in the GCT space. Due to this ambiguity, we map pitch information to the GCT domain *directly* as in the direct method for single-speaker resynthesis.

Voiced-on-unvoiced: If either $f_1[n]$ or $f_2[n]$ exhibit at least one non-zero value and the other exhibits all zero values in $s_{mix,hp}[n, \omega]$, we consider the region as voiced-on-unvoiced. Carrier positions for the voiced speaker are again obtained using the direct mapping. Peak-picking is then done on the GCT domain to obtain noise carrier locations. Similar to the voiced-on-voiced case, these carriers exhibit ambiguity in assignment. To account for this, we remove carriers obtained in peak-picking that are within a threshold of $\Omega_e = 2\Omega_A = 0.1786\pi$ away from the mapped locations of the *voiced* carriers as motivated from our bandwidth constraint in Section 3.1.1 for the envelope. The remaining carriers obtained in peak-picking are assigned to the unvoiced speaker. In the present formulation, it is possible that no carriers are assigned to the unvoiced speaker if all peak positions are pruned away in relation to the mapped pitch conditions. In this case, we subtract the least-squares fit of $s_{mix}[n, \omega]$ from a single speaker (i.e., the speaker *with* carriers assigned) as an estimate of the unassigned speaker, e.g., if speaker 1 has no carriers

$$\hat{s}_1[n,\omega] = s_{mix}[n,\omega] - \hat{s}_2[n,\omega]. \tag{3.78}$$

Unvoiced-on-unvoiced: If both $f_1[n]$ and $f_2[n]$ consist entirely of zeros, the region is unvoiced-on-unvoiced. In this case, a set of carriers is obtained from directly peak-picking the GCT as in the unvoiced case for a single speaker. In this case, we set estimates of individual speakers as half the amplitude of the least-squares fit to both speakers, i.e.,

$$\hat{s}_{1,u}[n,\omega] = \hat{s}_{2,u}[n,\omega] = \frac{\hat{s}_{12}[n,\omega]}{2} \tag{3.79}$$

where $\hat{s}_{12}[n,\omega]$ is the fit to the region assuming a single speaker and $\hat{s}_{1,u}[n,\omega] = \hat{s}_{2,u}[n,\omega]$ are new estimates of each speaker.

3.4.2 Exclusion and Re-estimation Method

As alluded to in our previous discussion, overlap of envelope content corresponding to distinct speakers in the GCT at *carrier* locations may reduce the effectiveness of speaker separation using the *direct* approach. In particular, demodulating a term that contains envelope content from *both* speakers can lead to erroneous estimates of the envelope. A schematic of such overlap is shown in Figure 3-19. Herein we describe a method for *excluding* carrier terms in demodulation for envelope estimation and subsequent *re-estimation* of carrier terms. This method is applied to the voiced-on-voiced condition while the voiced-on-unvoiced and unvoiced-on-unvoiced conditions are solved as in the direct method. Figure 3-19 illustrates this algorithm in detail.

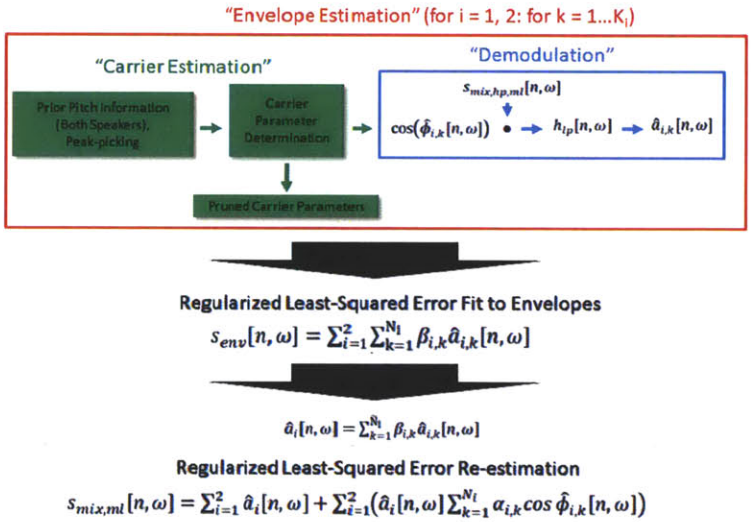

Figure 3-19. Schematic illustrating exclusion/re-estimation method; the full set of carrier positions are used to generate modulated envelopes subtracted from $s_{mix,ml}[n,\omega]$ to form $s_{env}[n,\omega]$; $s_{env}[n,\omega]$ is fit using a *subset* of the demodulated envelopes (denoted by $\widetilde{N}_i$ in the envelope fitting step). Envelope estimates are then combined with the full set of carriers in a final re-estimation step.

We consider again the set of carrier positions obtained from the direct pitch mapping of the two speakers denoted as $S1 := (\omega_1 k, \Omega_1 k), k = 1,2,\dots,N_1$ and $S2 := (\omega_2 l, \Omega_2 l), l = 1,2,\dots,N_2$. We compute the distance between all elements of $S1$ and $S2$, i.e.,

$$D_{S1,S2}(k = 1,\dots,N_1, l = 1,\dots,N_2) = \sqrt{(\omega_1 k - \omega_2 l)^2 + (\Omega_1 k - \Omega_2 l)^2} \tag{3.80}$$

For each element in $D_{S1,S2}(k,l) < 2\Omega_A = 0.1786\pi$, we obtain *subsets* $\widetilde{S1}$ and $\widetilde{S2}$ from $S1$ and $S2$ by removing carrier values with strictly larger carrier *numbers*, e.g., removal of k^{th} carrier if $k > l$, removal of l^{th} if $l > k$. An example of such removal is shown in Figure 3-16, in which the 2nd harmonic of the red speaker is excluded while the first harmonic of the blue speaker is kept. This results in the subsets $\widetilde{S1}$ and $\widetilde{S2}$ with sizes $\widetilde{N}_i \leq N_i, i = 1,2$. The use of harmonic order for pruning is motivated from our observations in Section 3.1.1 indicating that the magnitudes of the sinusoidal carrier series in the GCT *decrease* with increasing number (e.g., Figure 3-2). Since carriers of a speaker are pruned when their carrier numbers are *strictly* greater than that of the other speaker, the method guarantees *at least* one carrier per speaker in the resulting subsets.

For demodulation, as in the direct method, the *full* set of carriers are first used to generate modulated envelope representations $\hat{a}_{i,k}[n,\omega]\cos\hat{\phi}_{i,k}[n,\omega], k = 1,2,\ldots,N_1+N_2, i = 1,2$ and subtracted from $s_{mix}[n,\omega]$ to approximate the sum of the envelopes alone (3.72). In the exclusion method, we then use the *subset* of the $\hat{a}_{i,k}[n,\omega]$ obtained from the carrier positions in $\widetilde{S1}$ and $\widetilde{S2}$ to estimate gain parameter values and subsequently the envelope estimates of $a_1[n,\omega]$ and $a_2[n,\omega]$ denoted as $\hat{a}_1[n,\omega]$ and $\hat{a}_1[n,\omega]$, i.e.,

$$s_{env}[n,\omega] = s_{mix,ml}[n,\omega] - \sum_{i=1}^{2}\sum_{k=1}^{\widetilde{N}_i}\hat{a}_{i,k}[n,\omega]\cos\hat{\phi}_{i,k}[n,\omega] = \sum_{i=1}^{2}\sum_{k=1}^{\widetilde{N}_i}\beta_{i,k}\hat{a}_{i,k}[n,\omega] \quad (3.81)$$

$$\hat{a}_i[n,\omega] = \sum_{k=1}^{\widetilde{N}_i}\beta_{i,k}\hat{a}_{i,k}[n,\omega], i = 1,2 \quad (3.82)$$

where $\widetilde{N}_i \leq N_i$ and $s_{env}[n,\omega]$ is defined as in (3.72). Note that the summation of envelopes on the right-hand side of (3.81) for each speaker ranges from $k = 1$ through $k = \widetilde{N}_i$ (rather than $k = N_i$ in the direct method) indicating that only envelopes obtained from the *subset of pruned carriers* $\widetilde{S1}$ and $\widetilde{S2}$ are used in envelope estimation. $\beta_{i,k}$ are solved using the previously described least-squares and (when necessary) regularized least-squares methods. This method aims to obtain envelope estimates from carrier terms while attempting to minimize the effects of overlapped envelope structure of distinct speakers at *carrier* locations in the GCT space.

Envelope estimates are then combined with the *full* set of carriers for each speaker to solve for the carrier amplitudes denoted as $\alpha_{i,k}$ in *re-estimation* step,

$$s_{mix}[n,\omega] = \sum_{i=1}^{2}\hat{a}_i[n,\omega] + \sum_{i=1}^{2}\left(\hat{a}_i[n,\omega]\sum_{k=1}^{N_i}\alpha_{i,k}\cos\hat{\phi}_{i,k}[n,\omega]\right) \quad (3.83)$$

(3.83) is solved again using the least squares methods, and the resulting gains are combined with the envelope and carriers of individual speakers. The current *exclusion/re-estimation* procedure exploits the distribution of replicas of the envelope term in the GCT space to account for overlapped versions of envelopes both at the origin as well as at carrier positions.

3.4.3 Reference Method and Fusion

The previous sections have discussed GCT-based approaches to speech-signal separation using prior pitch information. As a reference signal representation for the co-channel speaker separation task, we apply a standard frame-based sinusoidal-based separation method developed first in [23] and extended in [36] with similarities to the method developed in [27] that requires a priori pitch information [23]. Specifically, the system of [23] is modified in several ways to address issues of singular least-squared error matrices as well as handle unvoiced speech. We

refer the reader to Appendix B for discussion of the specific modifications and [23] for the general setup of the method.

In addition to the reference method, we sought to assess potential benefits of *fusing* the output of the GCT-based method with the sinusoidal-based method. Fusion allows us to assess the extent to which the GCT-based methods can provide additional "complementary" information in separation, distinct from that using the reference sinusoidal-based method. As one motivation for fusion, recall that the GCT signal model explicitly represents temporal pitch dynamic content (i.e., the rotation of harmonic structure θ in Figure 3-4). It is conceivable therefore that speakers exhibiting similar pitch values but *different* pitch dynamics can be better separated in the GCT space in contrast frame-based sinusoidal method that relies on pitch values only in estimating parameters to the individual speakers. As a simple example of this, Figure 3-20 schematizes a narrowband region in which two speakers have similar pitch values as measured by the vertical distance of components in the GCT domain but different pitch dynamics.

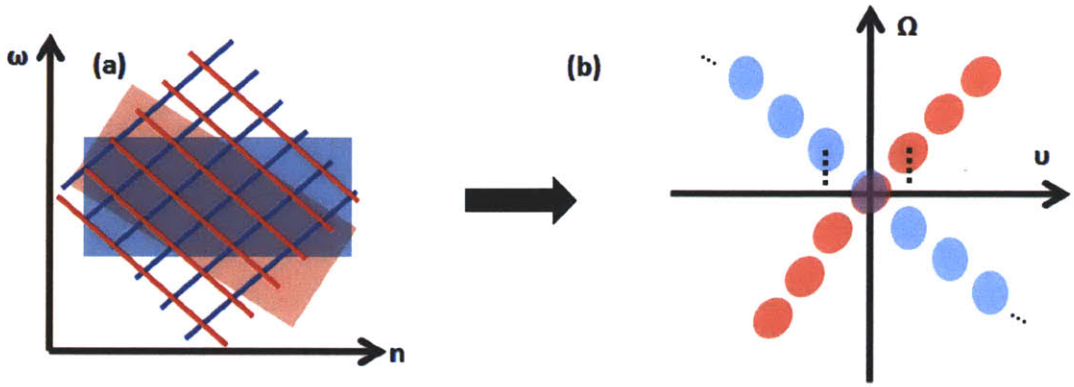

Figure 3-20. (a) Schematic of two speakers in a local time-frequency region with similar pitch values but distinct pitch dynamics; (b) in the corresponding GCT domain, separation of replicas of envelope content (overlapped at the GCT origin) can still be maintained due to this distinction in pitch dynamics despite similar pitch values as represented by vertical distance of components to v axis (dashed lines).

As an additional estimate of individual speakers, we therefore consider a simple fusion method using a weighted sum of the GCT-based and sinusoidal-based systems, i.e.,

$$\hat{x}_{fused}[n] = \alpha\hat{x}_{narrow}[n] + (1-\alpha)\hat{x}_{Sine}[n] \tag{3.84}$$

where $0 \leq \alpha \leq 1$. α is tuned on a development set of narrowband estimates to maximize the overall average signal-to-noise ratio. Here, $\hat{x}_{narrow}[n]$ refers to the GCT estimate and $\hat{x}_{Sine}[n]$ refers to the reference sinusoidal-based (frame-based) method.

3.5 Evaluation

Herein we discuss methods and results for spectrogram analysis/synthesis and speaker separation.

3.5.1 Data Set

In our experiments, we use a subset of the TIMIT corpus sampled at 16 kHz [37]. For analysis/synthesis of spectrograms, 10 males and 10 females speaking 2 distinct utterances are used for a total of 40 examples. We use two data sets for development and final testing. The final test set is generated from 4 males and 4 females speaking 2 sentences. Sentences are

additively mixed after truncating to the minimum length of the two with 0 dB overall signal-to-signal ratio; care was taken such that each mixture contains distinct speakers and sentences. This results in 24 male-male (MM), 24 female-female (FF), and 64 female-male (FM) mixtures. A set of 15 mixtures (5 each MM, FM, FF) from a distinct set of 3 males and 3 females were used in development. The Wavesurfer software package was used to estimate the pitch trajectories of all sentences individually prior to analysis [38].

As described in Section 2.2, the speaker separation problem has been addressed under a variety of different formulations/constraints imposed on the problem. Our focus is on single-channel, speaker independent methods. In this context, data in the existing literature has generally been further constrained to reflect the separation capabilities of proposed systems. As an example, the sinusoidal-based (that is frame-based) separation system was evaluated strictly on *all-voiced* data without the inclusion of unvoiced speech and/or pauses between words for both the target and interferer (e.g., "Nanny may know my meaning", "Why were you away a year Roy?") [23]. Alternatively, in auditory-based approaches such as that by Wu, et al. [39], the underlying target was chosen to be strictly voiced but interfering speech was allowed to have voiced and unvoiced components. Furthermore, analyses of the data used in these experiments indicated that the underlying pitch trajectories of speakers exhibited no crossings or mergings. For these reasons, we chose in this thesis to evaluate our algorithms on what we believe to be a more general database accounting for voiced/unvoiced *mixtures* as well as allowing for pitch trajectory merging and crossings. An example is shown in Figure 3-21 highlighting the presence of both conditions for a mixture of two female speakers. Observe that at time 400 ms that the pitch tracks of both speakers are very close in frequency, thereby approaching a "merging" of the pitch tracks. Similarly, at time ~900 ms, observe that speaker 2's pitch track exhibits a crossing with that of speaker 1; specifically, speaker 2 has decreasing pitch at this time point while speaker 1 has increasing pitch.

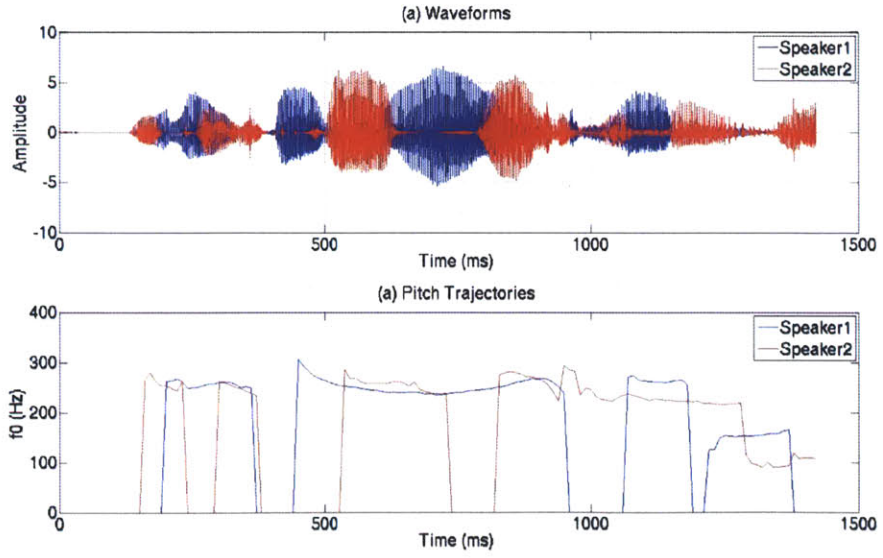

Figure 3-21. (a) Waveforms of two female speakers ("Anything wrong Captain?" + "With his club foot, he might well..."); (b) Pitch tracks of speakers exhibiting unvoiced and voiced speech (pitch values of zero indicate silence/unvoiced speech) and pitch crossings.

3.5.2 Spectrogram Analysis/Synthesis

Specific Methods: A narrowband spectrogram $s_{full}[n,\omega]$ is computed on the signal $x_{single}[t]$ using a 32-ms Hamming window, 1-ms frame interval, a 512-point discrete Fourier transform (DFT). In GCT analysis, local regions of size 875 Hz by 20 ms are extracted using a 2-D modified Hamming window with overlap factor of 4 to ensure the constant-overlap-add criterion in 2-D [1]. The GCT is computed using a 512-point 2-D DFT. A high-pass (low-pass) 2-D filter $h_{hp}[n,\omega]$ ($h_{lp}[n,\omega]$) is designed using a 1-D filter designed using the frequency sampling method followed by the frequency transformation method [2][29]. $h_{hp}[n,\omega]$ ($h_{lp}[n,\omega]$) has order 80 with pass-band (stop-band) beginning at $\Omega_A = 0.0893\pi$ (Section 3.1.1) and stop-band (pass-band) roll-off to $2\Omega_A$. $h_{hp}[n,\omega]$ is applied to $s_{full}[n,\omega]$ to obtain $s_{full,hp}[n,\omega]$. $s_{full,hp}[n,\omega]$ is multiplied by a set of sinusoidal carriers generated from the pitch track and each product is low-pass filtered by $h_{lp}[n,\omega]$ to obtain a set of $\hat{a}_k[n,\omega]$ for use in demodulation (Section 3.2.1).

Two quantitative metrics are used to compare the results of spectrogram analysis/synthesis. In the first, we directly compute the root-mean-squared-error between the estimated spectrogram and the reference spectrogram across all time-frequency points, i.e.,

$$RMSE = \sqrt{\frac{1}{NM}\sum_{n=1}^{N}\sum_{\omega=1}^{\omega_N}\left[s_{full}[n,\omega] - \hat{s}_{full}[n,\omega]\right]^2} \tag{3.85}$$

where ω_N denotes the total number of DFT frequency bins in the spectrogram. In addition, $s_{full}[n,\omega]$ is combined with the phase of the original single-speaker sentence to resynthesize a waveform of a single speaker using the least-squared error (LSE) overlap-add method (LSE-OLA) [1] to generate $\hat{x}_{single}[t]$. Note that $\hat{x}_{single}[n]$ represents an "upper limit" of resynthesis performance of the waveform using the 2-D signal model due to the use of the *original* phase in reconstruction. Based on this waveform estimate, we compute a signal-to-noise (SNR) ratio of $\hat{x}_{single}[n]$ defined as

$$SNR = 10log\left(\frac{\sum_t x^2{}_{single}[n]}{\sum_t\left[x_{single}[n]-\hat{x}_{single}[n]\right]^2}\right). \tag{3.86}$$

Results: In Figure 3-22, we show the results of analysis/synthesis of a single utterance spoken by a male. Observe that while both the direct and bootstrapping techniques with demodulation preserve the general characteristics of the harmonic, noise, and onset/offset structure of the original spectrogram. An outstanding limitation in both cases appears to be modest widening of the harmonic and formant structure, presumably due to bandwidth constraints of the model (3.15). Relative to the direct approach, bootstrapping appears to introduce fewer distortions in harmonic structure, presumably due to more accurate demodulation results (e.g., at time $n = 1000$). Similar relative effects were observed on male speakers. In contrast, observe that the two-dimensional sinusoidal-series fit (see Section 3.3) exhibits widening of the formant bandwidths as well as that of onsets/offsets due to the lack of an envelope term in reconstruction. In Figure 3-23, we show a representative example of this effect for formant structure; observe that while both demodulation capture the sharp harmonic and nearby content at $\omega \approx 0.08\pi$, the sinusoids fail to do so. Similarly, Figure 3-24 shows the time slices of a vowel onset; while both demodulation approaches capture the sharpness of this onset, the sinusoidal-only model results in substantial smoothing effects. Quantitatively, Table 3-1 lists the average root-mean-squared errors of the methods, indicating that the bootstrapping technique provides gains over the direct method of ~0.02 RMSE while both demodulation methods outperform the sinusoidal-only

approach. These results demonstrate the significance of the envelope term in the modulation model in representing onset/offset content as well as formant structure.

Table 3-1 further lists the average signal-to-noise ratios of the waveforms when the estimated magnitudes are combined with the true phase of the original signal; recall that these represent "upper limits" of waveform reconstructions due to use of the original phase potentially recovering magnitude information lost due to limitations of the signal model. Overall, the waveform reconstruction results in SNR of at least ~11 dB. In informal listening, listeners (non-authors) reported indistinguishable waveforms between the original and bootstrap and direct methods. However, the sinusoidal-series fit was described as substantially more "muffled" and "rough" in relation to the original. These results are consistent with accurate magnitude modeling using both the direct and bootstrapping methods. Waveform quality is presumably maintained by a combination of this modeling and use of the phase of the original signal in waveform reconstruction. Our results demonstrate empirically that the relative poor estimates of spectrogram magnitudes using the 2-D sinusoidal-series fits cannot be recovered using phase information.

3.5.3 Co-channel Speaker Separation

Specific Methods: In speaker separation, the mixed signal $x_{mix}[n]$ is processed as in the single-speaker case for short-time and GCT analysis. The reconstructed spectrograms are combined with the phase of the mixed signal, and the waveform $\hat{x}_i[n]$ is reconstructed using the LSE-OLA method as in the single-speaker case. Since the mixtures were created at 0 dB signal-to-signal ratio, we compute the signal to interferer *gain* as a goodness metric,

$$Gain_{SNR,i} = 10log\left(\frac{\Sigma_t x^2{}_i[n]}{\Sigma_t[x_i[n]-\hat{x}_i[n]]^2}\right) \tag{3.87}$$

where $x_i[t]$ is the original (unmixed) utterance. In the least-squares formulation, the c_2 parameter (i.e., the threshold for determining diagonal loading (3.75)), is tuned on the development set with values $[10^0, 10^1, \ldots, 10^6]$ for both the direct and exclusion methods. The best c_2 values based on the average SNR gain was then applied to the final test set; in our development, we observed that c_2 values ranging from 10^0 through 10^3 provided similar results and set $c_2 = 10^2$. Finally, in fusion, we swept α values in step sizes of 0.01 on separation results from the development set obtained using the GCT-based "exclusion/re-estimation" approach and sinusoidal-based method. An "optimal" α value of 0.49 was obtained and used on the test set.

Results: In Figure 3-22, we show the results of analysis/synthesis of a single utterance spoken by a male. Observe that while both the direct and bootstrapping techniques with demodulation preserve the general characteristics of the harmonic, noise, and onset/offset structure of the original spectrogram. An outstanding limitation in both cases appears to be modest widening of the harmonic and formant structure, presumably due to bandwidth constraints of the model (3.15). Relative to the direct approach, bootstrapping appears to introduce fewer distortions in harmonic structure, presumably due to more accurate demodulation results (e.g., at time $n = 1000$). Similar relative effects were observed on male speakers. In contrast, observe that the sinusoidal-only model exhibits widening of the formant bandwidths as well as that of onsets/offsets due to the lack of an envelope term in reconstruction. In Figure 3-23, we show a representative example of this effect for formant structure; observe that while both demodulation methods capture the sharp harmonic and nearby content at $\omega \approx 0.08\pi$, the sinusoidal method fails to do so. Similarly, Figure 3-24 shows the time slices of a vowel onset; while both demodulation approaches capture the sharpness of this onset, the sinusoidal-only model results in substantial smoothing effects.

Quantitatively, Table 3-1 lists the average root-mean-squared errors of the methods, indicating that the bootstrapping technique provides gains over the direct method of ~0.02 RMSE while both demodulation methods outperform the sinusoidal-only approach. These results demonstrate the significance of the envelope term in the modulation model in representing onset/offset content as well as formant structure.

Table 3-1 further lists the average signal-to-noise ratios of the waveforms when the estimated magnitudes are combined with the true phase of the original signal; recall that these represent "upper limits" of waveform reconstructions due to use of the original phase potentially recovering magnitude information lost due to limitations of the signal model. Overall, the waveform reconstruction results in SNR of at least ~11 dB. In informal listening, listeners (non-authors) reported indistinguishable waveforms between the original and bootstrap and direct methods. However, the sinusoidal-only method was described as substantially more "muffled" and "rough" in relation to the original. These results are consistent with accurate magnitude modeling using both the direct and bootstrapping methods. Waveform quality is presumably maintained by a combination of this modeling and use of the phase of the original signal in waveform reconstruction in the demodulation methods. Our results demonstrate empirically that the relatively poorer estimates of spectrogram magnitudes using the sinusoidal methods cannot be recovered using phase information.

In Figure 3-25 and Figure 3-26, we show results of speaker separation for a female-female (FF) and female-male (FM) mixture with a male target, respectively. For the FF case, observe that both the direct and exclusion methods can provide faithful reconstruction of the target speaker and suppression of the harmonic structure of the interferer Relative to the direct method, observe at time $n = 750$ that the exclusion method can provide more suppression of the harmonic structure of the interferer, presumably due to exclusion in demodulating the envelope terms. In the FM case, observe that both the direct and exclusion method provide less suppression of interfering speakers at e.g., time $n = 2200$; this is likely caused by inaccurate carrier parameters being used in demodulation such that harmonic structure of the interfering speaker is maintained. In addition, the overall reconstruction of the target speaker is qualitatively worse than in the FF case. An explanation for this performance difference across genders is that male speakers generally have lower pitch than females. Low pitch values are further away from the GCT origin such that there are fewer carrier terms for demodulation. Consistent with this explanation observe quantitatively that SNR gains increase from the MM (~3 dB), MF (male target), MF (female target), and FF cases (up to ~6 dB).

As an oracle experiment, we also show in Table 3-3 results of combining demodulated estimates with their corresponding *true* phase spectra (instead of the phase of the mixture signal); here, observe that we can obtain substantially higher SNR values up to ~8 dB due to the combined effects of magnitude and phase in waveform reconstruction. Finally, in informal listening, (non-author) listeners reported generally faithful reconstruction of target speakers with suppression, though not complete removal, of interferers; specifically, listeners reported "weaker" (in amplitude) and "muffled" interferers in the resulting estimates.

In relation to the reference method, the GCT does not appear to perform as well as the frame-based sinusoidal-based reference system as shown in Table 3-3, with up to ~3 dB poorer performance (in the FF case). This effect may be due in part to the GCT's use of magnitude only to obtain separation estimates as evidenced by the previously noted performance improvement with addition of the true phase. Nonetheless observe that fusion of the GCT-based and sinusoidal-based-based separation can result in up to ~1 dB gain relative to the sinusoidal-based-based system. These results provide evidence for complementary information provided by the GCT

signal representation in signal separation. As an example of improved separation, Figure 3-27 shows a mixture of two males as well as estimates from fusion and the reference sinusoidal method. Observe at times ~1.25e4 and ~1.65e4 (in samples) that the fused estimates more closely resembles the reference waveform of the target in reducing the periodic-like components obtained from the sinewave system. A potential explanation of these gains may be attributed to the GCT's use of temporal pitch dynamics in providing separability of speakers despite similar pitch values.

As another quantitative metric, we compute in Table 3-2 root-mean-squared errors (RMSE) between the estimated spectrograms themselves and the reference spectrograms for reference in relation to potential application of magnitude-only representations (e.g., features in automatic speech recognition). RMSE values are computed after scaling both the estimate the reference spectrograms by their maximum values to account for overall scaling effects obtained from overlap-add. We observe that the exclusion method provides gains over the direct method using the GCT, consistent with the SNR results. RMSE values obtained for spectrograms recomputed using the fused and sinusoidal-based estimated waveforms do not exhibit such a correspondence in relation to their relative SNRs. We believe this discrepancy to reflect effects of recomputing the spectrogram and effects of the *mixture phase* used in waveform reconstruction. In informal listening, (non-author) listeners reported that the sinusoidal-based system does not result in complete removal of interferers, similar to the GCT. One reported distinction is that sinusoidal-based results in "whispered" interferers while the GCT results in interferers with lower amplitude. In addition, the fused result in Figure 3-27 was reported to not exhibit interfering periodic components consistent with the qualitative observations of the waveforms. The combined results of analysis/synthesis and separation demonstrate that the GCT-based signal model can provide good representations of speech and is a promising one for the co-channel speaker separation task.

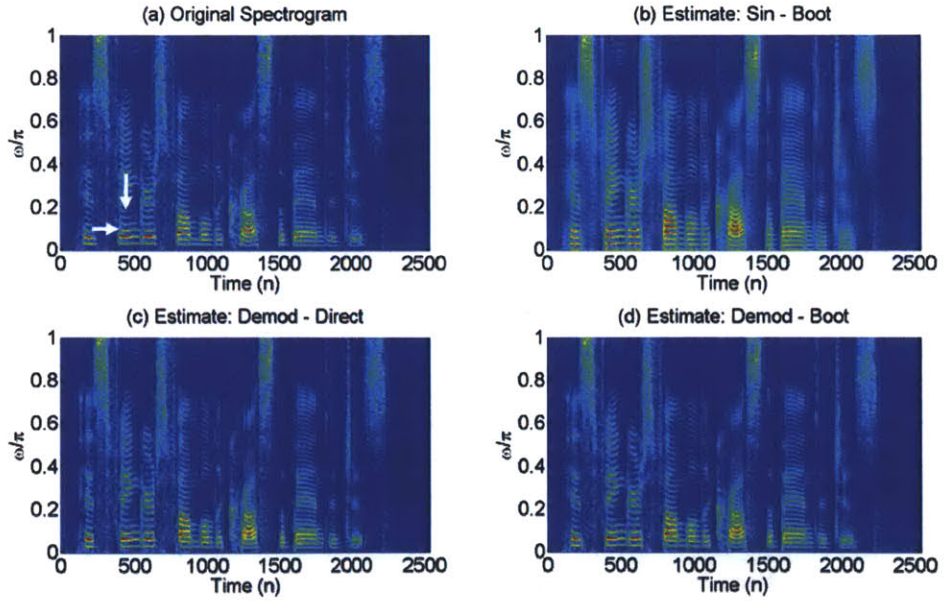

Figure 3-22. (a) Original spectrogram of a female utterance; (b) Estimate from 2-D sinusoidal-series fit using the bootstrapped carrier positions; (b) direct mapping demodulation method; (c) bootstrapped carriers demodulation method; in (a), we denote vertical and horizontal arrows as frequency and time slices, respectively.

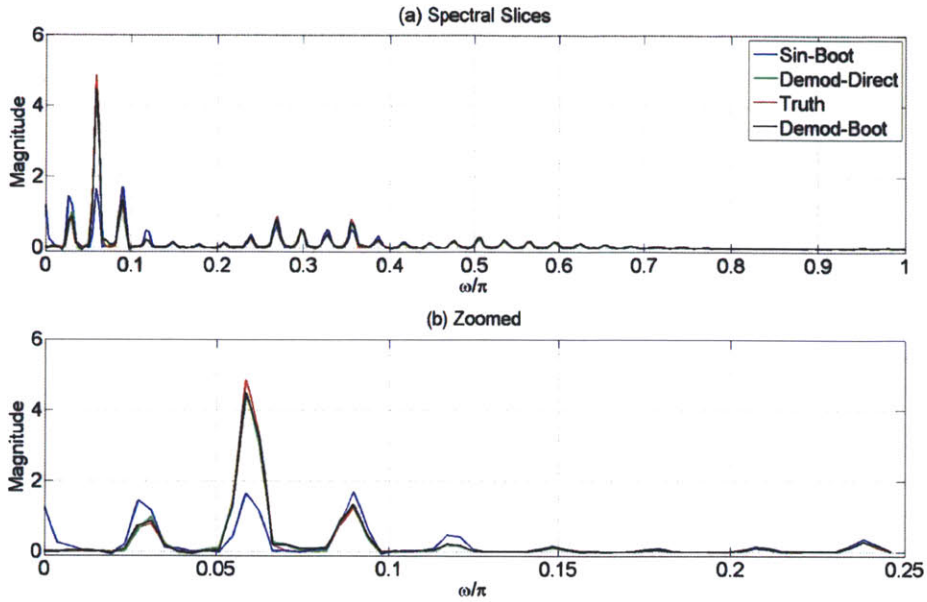

Figure 3-23. (a) Full spectral slices from original (red), sinusoids-series fit only (blue), direct (green) and bootstrap (black) demodulation methods; (b) Local frequency region of (a). Extracted at time 440 from Figure 3-22.

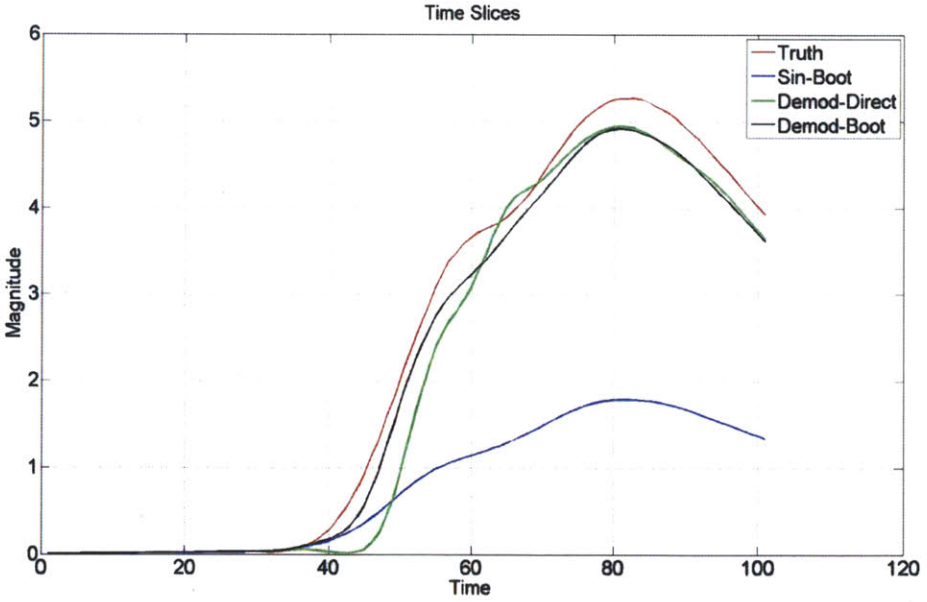

Figure 3-24. Time slices from Figure 3-22 extracted at frequency $\omega = 0.062\pi$, times 350-450.

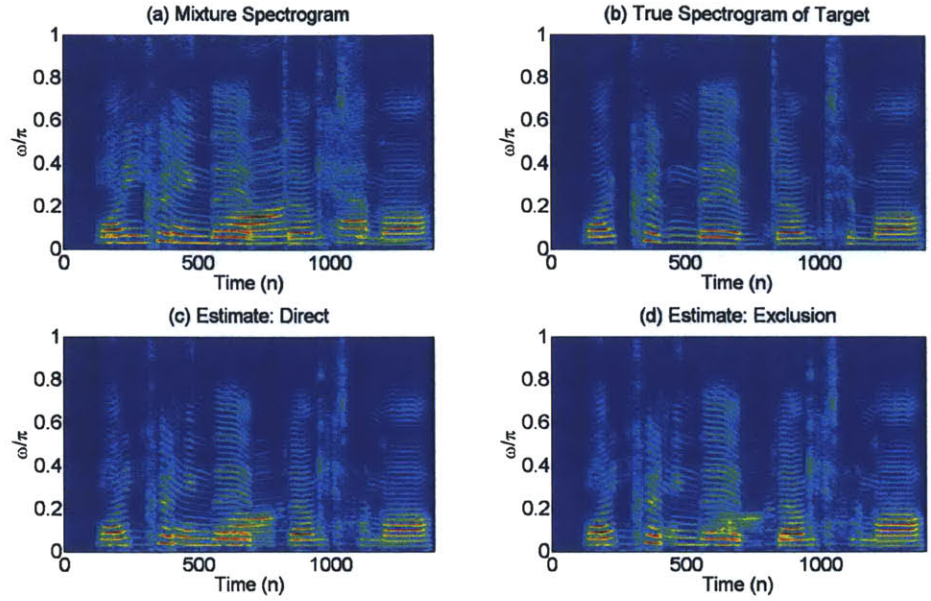

Figure 3-25. (a) Spectrogram of mixture (FF) ("Why couldn't they have dumped him off", "Anything wrong, Captain?"); (b) Original target spectrogram (utterance containing "Why"); (c) Reconstruction from direct method; (d) reconstruction from exclusion.

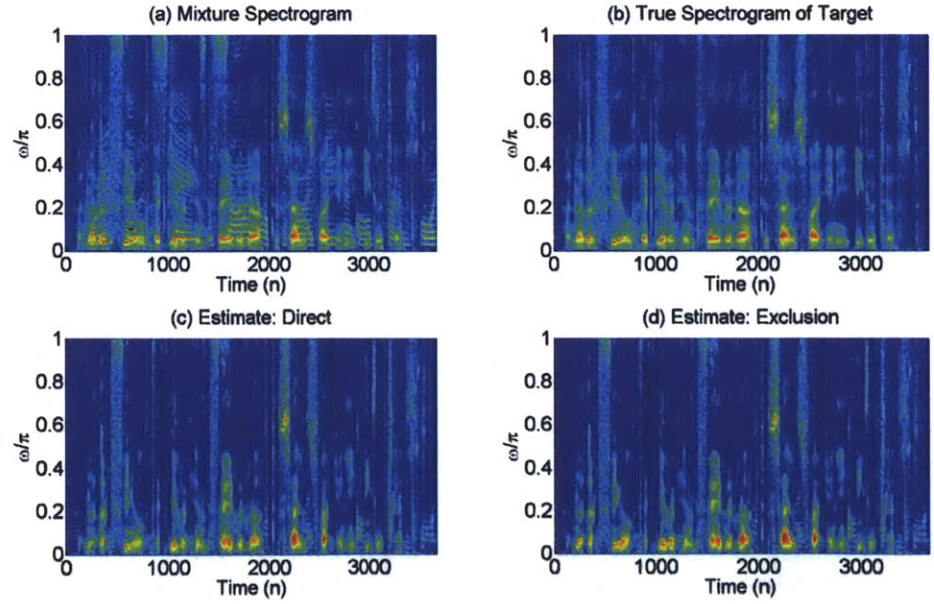

Figure 3-26. As in Figure 3-25 but a mixture of female ("Forty-seven states assign or provide vehicles for employees") and male ("Another field had given him fame enough to satisfy any egotist") with male target.

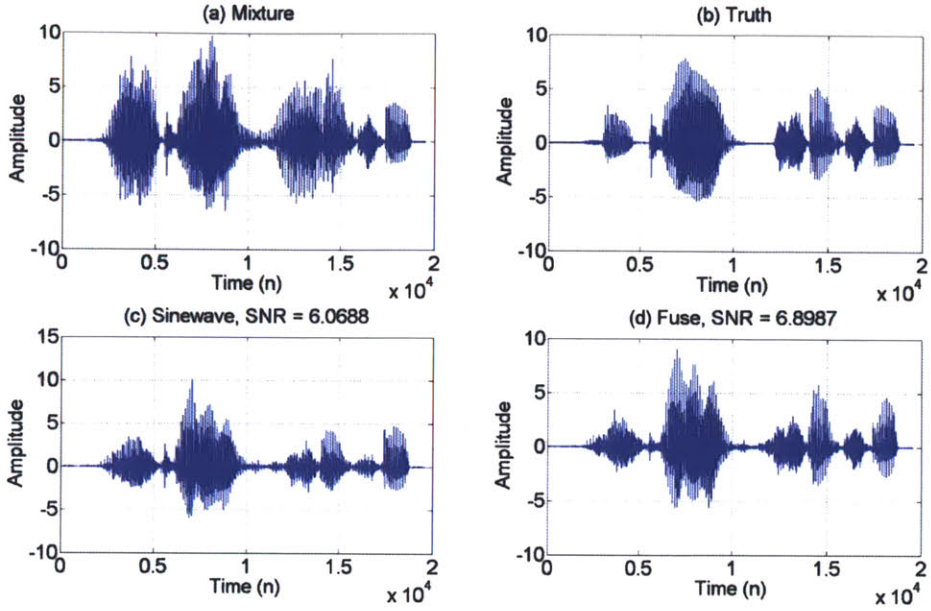

Figure 3-27. (a) Mixture of two males ("They'll tow the line" + "He drove essential"); (b) reference target waveform ("He drove essential"); (c) estimate from sinusoidal representation; (d) fused estimate.

Table 3-1. Average RMSE and SNRs for analysis/synthesis of spectrograms and standard errors; here, "Sine – Direct" and "Sine – Boot" correspond to the 2-D sinusoidal series fit using the direct and bootstrapped carrier positions.

	Sine – Direct	Sine – Boot	Demod. – Direct	Demod. – Boot
RMSE (Males)	1.80e-2 [7.28e-4]	1.77e-2 [7.29e-4]	1.17e-2 [6.34e-4]	1.06e-2 [6.12e-4]
RMSE (Females)	1.39e-2 [7.68e-4]	1.34e-2 [8.48e-4]	5.76e-3 [4.51e-4]	4.18e-3 [3.44e-4]
SNR (dB) (Males)	6.33 [0.27]	6.46 [0.27]	10.52 [0.30]	11.36 [0.33]
SNR (dB) (Females)	8.50 [0.35]	8.74 [0.34]	17.77 [0.60]	20.84 [0.59]

Table 3-2. Average RMSEs for speaker separation, standard errors [] on test set.

	Direct	Exclusion	Fusion	Sine-based
MM	2.66e-2 [1.06e-3]	2.62e-2 [1.16e-3]	1.62e-2 [6.85e-4]	1.61e-2 [6.38e-4]
FF	1.93e-2 [5.86e-4]	1.58e-2 [3.65e-4]	1.02e-2 [2.29e-4]	9.29e-3 [2.63e-4]
FM – Male	2.37e-2 [6.85e-4]	2.22e-2 [6.96e-4]	1.29e-2 [3.68e-4]	1.20e-2 [3.15e-4]
FM – Female	1.84e-2 [5.13e-4]	1.66e-2 [4.41e-4]	1.19e-2 [3.14e-4]	1.28e-2 [3.54e-4]

Table 3-3. Average SNRs (dB) for speaker separation (dB), standard errors [] on test set; (T) denotes true phase results; here, "Sin-based" refers to frame-based sinusoidal-based separation method.

	Direct	Exclusion	Direct (T)	Exclusion (T)	Sine-based	Fusion
MM	3.70 [0.14]	3.80 [0.15]	5.22 [0.17]	5.23 [0.18]	4.73 [0.13]	5.58 [0.13]
FF	4.86 [0.14]	6.31 [0.14]	8.02 [0.14]	8.53 [0.14]	9.18 [0.18]	8.81 [0.14]
FM – Male	4.69 [0.10]	4.91 [0.11]	5.59 [0.12]	5.97 [0.13]	6.81 [0.13]	7.34 [0.10]
FM – Female	5.06 [0.09]	5.83 [0.11]	8.33 [0.11]	8.27 [0.13]	6.32 [0.10]	7.36 [0.14]

3.6 Conclusions

This chapter has proposed a 2-D signal model of speech content in local time-frequency regions of the narrowband spectrogram. In particular, a sinusoidal-series carrier as a function of pitch, pitch dynamic, and noisy source information is *modulated* in time and frequency by a slowly-varying envelope term reflecting formant/dynamic formant and onset/offset content. Consequently, a transformed 2-D Grating Compression Transform space (GCT) exhibits distributed *copies* of the 2-D Fourier transform of the envelope at both the GCT origin and at positions corresponding to the carrier parameters. We have explored the properties and limitations of the model through simulations on synthetic data. In addition, we have developed algorithms that seek to exploit the distributive nature of replicas of envelope content in the GCT based on demodulation and interpolation in local time-frequency regions of spectrograms for both analysis/synthesis and co-channel speaker separation using prior pitch information. In analysis/synthesis, our algorithms demonstrate that the signal model is capable of very accurately representing speech content in recovering a variety of energy fluctuations observed in the spectrogram. Results of co-channel speaker separation using the GCT-based methods alone demonstrate that it is a promising one for this task in providing global SNR gains (relative to 0 dB initial signal-to-signal ratios) up to ~6 dB. When fused with a standard frame-based sinusoidal method, gains in SNR are observed up to ~1 dB providing evidence for complementary information by the GCT representation for the separation task.

Chapter 4

Wideband Models and a Taxonomy of Speech-Signal Behavior

In Chapter 3, we developed speech-specific signal models for localized 2-D Fourier analysis of the *narrowband* spectrogram [7], a representation referred to as the (narrowband) Grating Compression Transform (NGCT). More generally, it is of interest to apply 2-D Fourier analysis to time-frequency distributions that can be viewed as mixtures of both narrowband and *wideband* spectrograms. Examples of such mixed-resolution distributions include the auditory, super-resolution, and cone-kernel spectrograms [40][1]. Towards this end, we develop in this chapter signal models for the counterpart *wideband* spectrogram in the context of the GCT (WGCT) (Figure 4-1), thereby providing a more complete interpretation of speech signal behavior in both the GCT framework and potentially other 2-D processing schemes and time-frequency distributions.

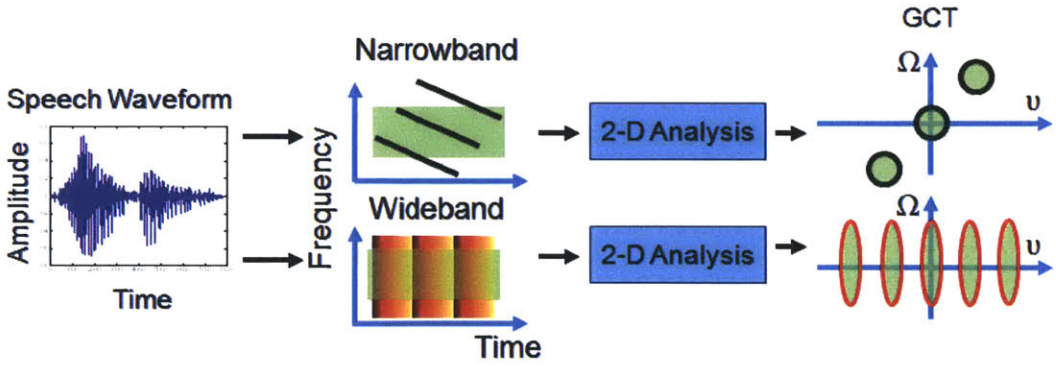

Figure 4-1. Schematic of general 2-D processing framework with short-time analysis followed by localized 2-D analysis for narrow (top) and wideband (bottom) representations.

In our development, we show that the WGCT is distinct from the NGCT in interpretation, thereby motivating a novel *taxonomy* of speech-signal behavior in 2-D processing of speech. We also show that the WGCT can be used in speech-signal processing via sinusoidal-series-based demodulation as in Chapter 3 to motivate spectrogram analysis/synthesis methods. To assess the ability of the model to represent speech content, we evaluate these methods for reconstruction of wideband spectrograms and as an example application, build on previous work in Chapter 3 in using the WGCT for co-channel speaker separation with prior pitch information. In this context, we emphasize our focus on assessing the signal models' representations of speech rather than developing a complete separation system. Chapter 6 describes efforts toward developing a

complete separation system.

This chapter is organized as follows. Section 4.1 reviews the GCT framework. Section 4.2 develops a 2-D speech-signal model for stationary voiced speech; Section 4.3 describes extensions to non-stationary voiced speech while Section 4.4 discusses models for speech-based noise and onset/offset content. Section 4.5 presents a taxonomy of speech-signal behavior in the WGCT and NGCT. Section 4.6 describes approaches to spectrogram reconstruction and speaker separation. Sections 4.7 and 4.8 present our results and conclusions, respectively. Section 4.9 is an appendix for reference to derivations referenced in Section 4.2.

4.1 Framework

We first review the Grating Compression Transform (GCT) framework developed in Chapter 3. Consider the short-time Fourier transform (STFT) of a speech signal $y[n]$ using a window $w[n]$

$$Y(n,\omega) = \sum_{m=-\infty}^{\infty} w[m-n]y[m]e^{-j\omega m}. \tag{4.1}$$

In Chapter 3, we considered $w[n]$ with length (L) 2~3 times the pitch period P of voiced speech present in $y[n]$, resulting in a narrowband spectrogram. This window choice leads to *harmonic line structure oriented across frequency*. For local time-frequency regions of $|Y(n,\omega)|$

$$|Y(n,\omega)|_{local} \approx w[n,\omega]H(n,\omega)E(n,\omega) \tag{4.2}$$

where $w[n,\omega]$ is the 2-D window, $H(n,\omega)$ is the vocal tract formant *envelope*, and $E(n,\omega)$ is a 2-D sinusoidal-series *carrier* dependent on pitch and pitch dynamic content. In the GCT domain, the model results in *distribution* of replicas of the envelope (Figure 3-1). Similar behavior was argued for unvoiced speech and onset/offset content.

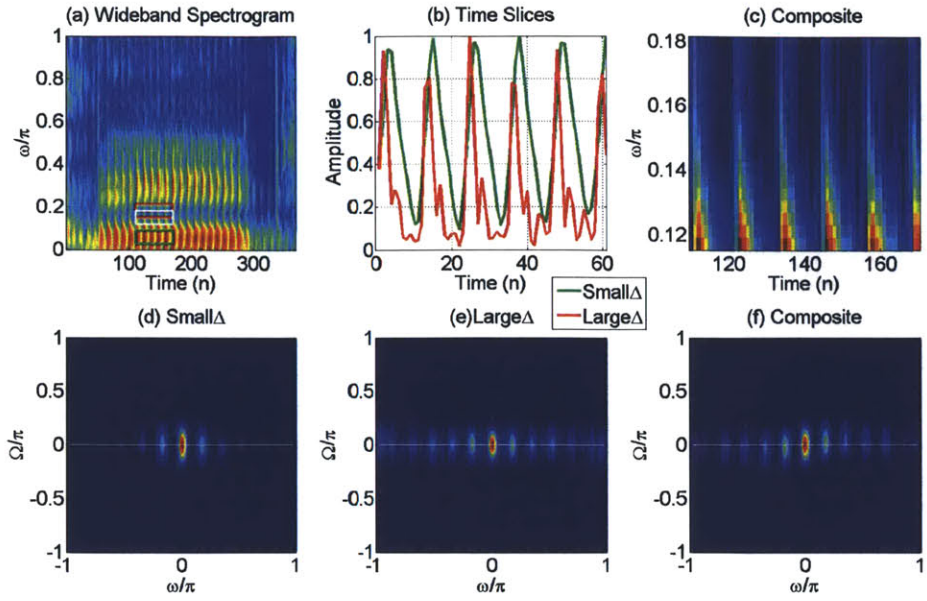

Figure 4-2. (a) Wideband spectrogram of real speech male utterance "needs" illustrating analysis near the first formant ($\omega \approx 0.05$) (b) small Δ (red), large Δ (green) and an (c) "edge" case (white); (d – f) WGCT representation of three regions; note off-axis terms in (e); (d – f) computed for regions including time slices in (b); see discussion of simulations for WGCT computation details.

This chapter considers $w[n]$ with $L < P$, such that $w[n]$ analyzes $y[n]$ within a single period P voiced speech [1]. This window choice leads to *harmonic-like grating structure oriented across time* in a *wideband spectrogram.* A model for wideband spectrograms of voiced speech is proposed in [1]

$$|Y(n,\omega)| = E[n]\widetilde{H}(\omega) \tag{4.3}$$

where $E[n]$ is a time-dependent "energy" term and $\widetilde{H}(\omega)$ is a "smoothed" version of the true formant envelope. Figure 4-2 shows analysis of several local time-frequency regions of a wideband spectrogram computed for voiced speech. We observe distinct behaviors in each region and their corresponding WGCTs based on the proximity to the first formant. Subsequently, we argue for a set of models with the general form of (4.3) to characterize these behaviors.

4.2 Stationary Voiced Speech Modeling

4.2.1 Single-Formant Modeling

Consider a simple model of speech in which an impulse train

$$p[n] = \sum_{k=0}^{N_k} \delta[n - kP] \tag{4.4}$$

with periodicity P and N_k terms excites a single formant modeled as a decaying sinusoid

$$h[n] = \xi_f e^{-\alpha_f n} \cos(\omega_f n)\, u[n]. \tag{4.5}$$

with corresponding Fourier transform

$$H(\omega) = \frac{0.5\xi_f}{\alpha_f + e^{j(\omega-\omega_f)}} + \frac{0.5\xi_f}{\alpha_f + e^{j(\omega+\omega_f)}}. \tag{4.6}$$

ξ_f, α_f, and ω_f are the amplitude, decay rate (corresponding to formant bandwidth), and formant frequency, respectively. We analyze the resulting signal

$$y[n] = \sum_{k=0}^{N_k} h[n - kP] \tag{4.7}$$

using the short-time Fourier transform (STFT) with $w[n]$ of length $L < P$ to satisfy the wideband constraint.

Consider the *filterbank* view of the STFT such that at an analysis frequency $\omega = \omega_f + \Delta$ [1],

$$Y(n, \omega) = (y[n]e^{-j\omega n}) *_n w[n] \tag{4.8}$$

$$Y(n, \omega) = \left(\sum_{k=0}^{N_k} h[n - kP]\, e^{-j(\omega_f+\Delta)n}\right) *_n w[n]. \tag{4.9}$$

By linearity of convolution, a single term in the summation is

$$Y(n, \omega; k) = \left(h[n - kP]e^{-j(\omega_f+\Delta)n}\right) *_n w[n] \tag{4.10}$$

with corresponding Fourier transform

$$Y(\omega', \omega; k) = e^{-jkP(\omega'-\omega_f-\Delta)} W(\omega') \left(\frac{0.5\xi_f}{\alpha_f + e^{j(\omega'-\Delta)}} + \frac{0.5\xi_f}{\alpha_f + e^{j(\omega'-2\omega_f-\Delta)}}\right). \tag{4.11}$$

n maps to ω' through the Fourier transform and is distinct from ω. Since $W(\omega')$ is concentrated near $\omega' = 0$ and nearly zero in magnitude far away from $\omega' = 0$ origin (i.e., at $\omega' = 2\omega_f + \Delta$),

$$Y(\omega', \omega; k) \approx e^{jkP(\omega_f+\Delta)} e^{-j\omega' kP} W(\omega') \frac{0.5\xi_f}{\alpha + e^{j(\omega'-\Delta)}}. \tag{4.12}$$

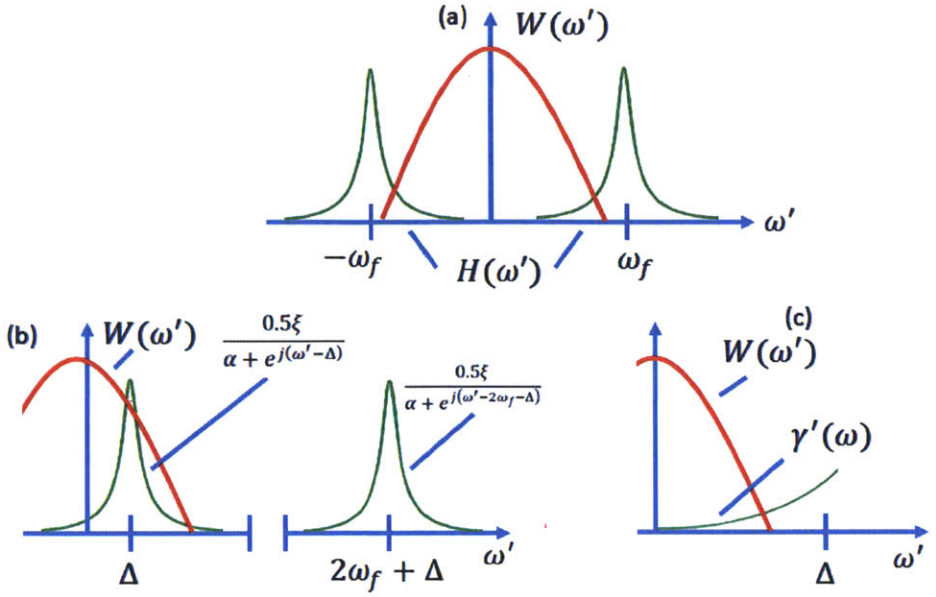

Figure 4-3. (a) Fourier transform of impulse response (green) and window (red); (b) small Δ case with majority of demodulated formant near origin is within window filter; modulated formant at $\omega' = 2\omega_f + \Delta$ excluded by window filter; (c) large Δ case with tail of formant content within bandwidth of window filter; $\omega' = 2\omega_f + \Delta$ component not shown.

We consider two limiting conditions of "small" or "large" values of Δ and derive *modulation* representations in both cases (Figure 4-3).

Small Δ: Applying the inverse Fourier transform to (4.12),

$$Y(n,\omega;k) \approx \left(0.5\xi_f e^{jkP(\omega_f+\Delta)}\right) w[n] *_n \left(e^{-\alpha_f(n-kP)}u[n-kP]e^{j\Delta(n-kP)}\right). \tag{4.13}$$

For *small* Δ, $e^{-jn\Delta}$ fluctuates slowly in time, and we therefore approximate it as $e^{-jn\Delta} \approx cos(\Delta) + jsin(\Delta)$. Furthermore, we assume that $\cos(\Delta)$ dominates $e^{-jn\Delta}$ for small Δ and $jsin(\Delta) \approx 0$. We then have

$$Y(n,\omega;k) \approx \gamma(\omega)e^{jkP(\omega_f+\Delta)}\epsilon[n-kP] \tag{4.14}$$

$$\gamma(\omega) = 0.5\xi_f cos\,(\omega - \omega_f) \tag{4.15}$$

$$\epsilon[n-kP] = w[n] *_n e^{-\alpha_f(n-kP)}\,u[n-kP]. \tag{4.16}$$

In (4.15), we have rewritten $\cos(\Delta)$ as $\cos(\omega - \omega_f)$ since $\Delta = \omega - \omega_f$. *Note that if $\Delta = 0$, (4.14) holds with equality with $\gamma(\omega) = 0.5\xi_f$.* Returning to the summation over k, we obtain

$$Y(n,\omega) \approx \sum_{k=0}^{N_k} e^{jkP(\omega_f+\Delta)}\gamma(\omega)\epsilon[n-kP]. \tag{4.17}$$

If $\epsilon[n-kP]$ decays to zero within each period, the *magnitude* of the sum may be approximated as the sum of the magnitudes, i.e.,

$$|Y(n,\omega)|_{local} \approx w[n,\omega]\gamma(\omega)E_d[n] \tag{4.18}$$

$$E_d[n] = \sum_{k=0}^{N_k} \epsilon[n - kP] = \sum_{l=0}^{N_s} \beta_l \cos\left(\frac{2\pi l}{P}n + \psi_l\right) \tag{4.19}$$

where $\gamma(\omega)$ is assumed to be non-negative for "small"-Δ e.g., $0 \leq |\Delta| \ll \frac{\pi}{2}$, and we have rewritten $E_d[n]$ as a sinusoidal series expansion. Here, we have also introduced a 2-D analysis window $w[n,\omega]$ to emphasize analysis in a *local* time-frequency region of the wideband spectrogram.

Our derivation argues for a modulation model as in (4.3) with a sinusoidal series carrier $E_d[n]$ representing source periodicity and formant bandwidth/decay rate (Figure 4-2b) and envelope $\gamma(\omega)$ representing frequency-dependent scaling of the formant peak in the spectral domain. It can be shown from an alternative Fourier transform view of the STFT that one interpretation of this scaling is smoothing of the true formant spectrum with the Fourier transform of the window:

$$\gamma(\omega) \approx \widetilde{H}(\omega) = |W(\omega) *_{\omega} H(\omega)|. \tag{4.20}$$

We refer the reader to Section 4.9 for a discussion of this derivation and illustrate subsequently through simulations its limitations. Note that if the bandwidth of $W(\omega)$ is substantially greater than that of $H(\omega)$, then the bandwidth of $\widetilde{H}(\omega)$ effectively becomes that of the window.

Since $E_d[n]$ and $\gamma(\omega)$ are separable in (4.18), its 2-D Fourier transform (i.e., the WGCT) is

$$Y(\upsilon,\Omega) = W(\upsilon,\Omega) *_{\upsilon,\Omega} \left[\eta(\Omega)\left(K\delta(\upsilon) + \sum_{l=1}^{N_s} 0.5\beta_l e^{\mp j\psi_l \upsilon}\delta\left(\upsilon \pm \frac{2\pi l}{P}\right)\right)\right] \tag{4.21}$$

where n and ω map to υ and Ω, respectively, and $\eta(\Omega)$ ($W(\upsilon,\Omega)$) is the Fourier transform of $\widetilde{H}(\omega)$ (i.e., a 2-D window in the time-frequency space $w[n,\omega]$). $\eta(\Omega)$ is the WGCT representation of the smoothed formant envelope in a *local* time-frequency region. Copies of $\eta(\Omega)$ are weighted by β_l coefficients (representing the bandwidth of the formant) along the υ-axis at multiples of $\frac{2\pi}{P}$ (representing the source periodicity). This product is further smoothed in υ and Ω with the Fourier transform of the 2-D analysis window. Note that formant bandwidth along the ω-axis "lost" due to smoothing by the short-time analysis window is "recovered" in *time* and represented in the carrier.

The present and subsequent formulation motivates a modulation/demodulation framework for speech signal processing similar to Chapter 3. Since copies of $\eta(\Omega)$ are distributed in the WGCT space via the carriers, they may be *demodulated* to reconstruct the $\eta(\Omega)$ term at the WGCT origin if this component is corrupted e.g. from an interfering signal.

Large Δ: For large Δ, the approximation in (4.14) does not hold. $\omega = \omega_f + \Delta$ is "far away" from ω_f, and we alternatively assume that the frequency response of the formant is approximately a single complex value $\gamma'(\omega)$ (i.e., a flat spectrum) (Figure 4-3). The frequency domain interpretation of this from (4.11) is

$$Y(\omega',\omega;k) = \gamma'(\omega)(e^{-jkP(\omega'+\omega)}W(\omega'). \tag{4.22}$$

Inverting (4.22) and invoking the summation as in (4.17),

$$Y(n,\omega;k) = \sum_k^{N_k} \gamma'(\omega) e^{-jkp\omega} \delta(n - kP) *_n w[n] = \sum_k^{N_k} \gamma'(\omega) e^{-jkp\omega} w[n - kP] \quad (4.23)$$

Since $L < P$, the summed terms do not overlap in time. In a local-time frequency region analyzed with a 2-D window $w[n,\omega]$, the magnitude of the sum can be rewritten as a sum of magnitudes, i.e.,

$$|Y(n,\omega)|_{local} = w[n,\omega]|\gamma'(\omega)|E_w[n] \quad (4.24)$$

$$E_w[n] = \sum_k^{N_k} w[n - kP] \quad (4.25)$$

resulting again in a *modulation* model of the spectrogram with a source periodicity-dependent carrier $E_w[n]$ and an envelope $|\gamma'(\omega)|$ which we again interpret as $\widetilde{H}(\omega)$ from (4.20). A WGCT representation analogous to (4.21) is given by (Figure 4-4)

$$Y(\upsilon,\Omega) = W(\upsilon,\Omega) *_{\upsilon,\Omega} \left[\eta'(\Omega)\left(K\delta(\upsilon) + \sum_{l=1}^{N_s} 0.5\beta'_l e^{\mp j\psi'_l \upsilon} \delta\left(\upsilon \pm \frac{2\pi l}{P}\right)\right)\right] \quad (4.26)$$

where $\eta'(\Omega)$ is the Fourier transform of $|\gamma'(\omega)|$ and β'_l and ψ' parameters of the sinusoidal series representation of $E_w[n]$. While the WGCT domain contains copies of $\eta'(\Omega)$ reflecting smoothed formant structure in local time-frequency regions as in the small Δ case, carrier positions and corresponding gain terms reflect source periodicity only. Note that $\eta'(\Omega)$ are not constrained along the Ω-axis and are therefore schematized as ellipses oriented vertically in the GCT as in Figure 4-4.

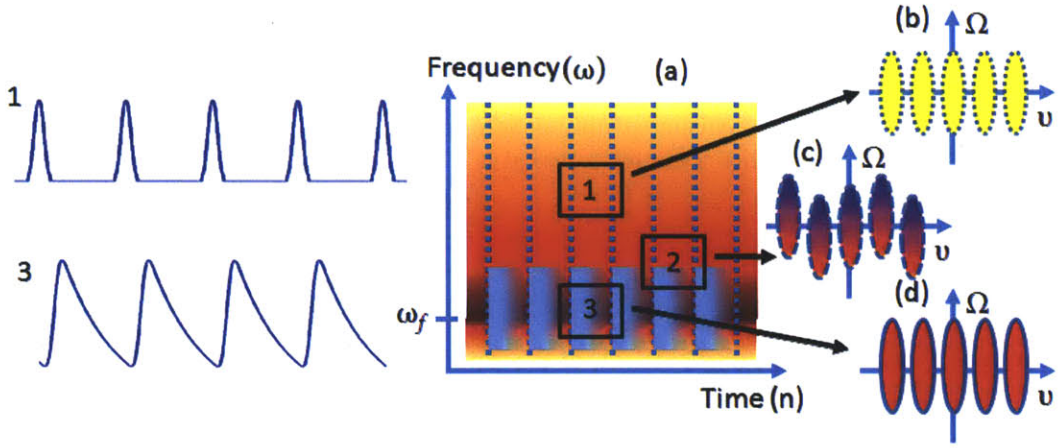

Figure 4-4. (a) Wideband spectrogram schematic illustrating analysis of a single formant in distinct frequency regions (1) large Δ, (2) small Δ, (3), "in between" case; periodicity and bandwidth-dependent carrier (blue, shaded), periodicity-dependent carrier (dotted lines) and composite carrier; (b-d) WGCT of regions 1 – 3 with distinct modulated envelopes delineated: small Δ – red, large Δ – yellow, "in between" – graded.

Composite Carrier: Our discussion thus far has described limiting cases of Δ modulation models in time-frequency regions of wideband spectrograms. To account for values of Δ "in between", we propose a "composite" carrier

$$E_c[n,\omega] = E_d[n]R[\omega] + E_w[n]R[\omega - \omega_0] \quad (4.27)$$

$$R[\omega] = \begin{matrix} 1, 0 < \omega < M \\ 0, otherwise \end{matrix} \tag{4.28}$$

where M ranges from 0 to the full length M_{full} region in frequency, and ω_0 is a shift in frequency. A similar composite carrier can be obtained by interchanging $E_w[n]$ and $E_d[n]$. $E_c[n,\omega]$ may be modulated by $\widetilde{H}(\omega)$ to invoke a modulation interpretation as in the limiting cases. A *generalized* modulation model in local time-frequency regions is

$$|Y(n,\omega)|_{local} = w[n,\omega]\widetilde{H}(\omega)(E_d[n]R[\omega] + E_w[n]R[\omega - \omega_0]). \tag{4.29}$$

The 2-D Fourier transform (WGCT) of (4.29) is (Figure 4-4)

$$Y(\upsilon,\Omega) = W(\upsilon,\Omega) *_{\upsilon,\Omega} \sum_{i\epsilon(d,w)} \eta_{R,i}(\Omega) \begin{pmatrix} K_{R,i}\delta(\upsilon) + \\ \sum_{l=1}^{N_s} 0.5\beta_{l_{R,i}}\delta\left(\upsilon \pm \frac{2\pi l}{P}\right) \end{pmatrix} \tag{4.30}$$

where $\eta_{R,i}(\Omega)$ is the Fourier transform of $\widetilde{H}(\omega)R(\omega)$ $(i = d)$ and $\widetilde{H}(\omega)R(\omega - \omega_0)$ $(i = w)$. $K_{R,i}$ and $\beta_{l_{R,i}}$ are the sinusoidal series coefficients of the two carrier types. The WGCT contains a scaled *sum* of $\eta_{R,i}(\Omega)$ terms at the origin and carrier locations. If the bandwidth of $\eta_{R,i}(\Omega)$ denoted as υ_η are such that $0.5\upsilon_\eta < \frac{2\pi}{P}$, then their modulated copies will occupy distinct regions along the υ-axis (Figure 4-4). Note that this model does not impose constraints on the bandwidth along the Ω-axis.

The WGCT also invokes a mapping of pitch f_0 information

$$\upsilon_0 = f_0 \frac{2\pi}{f_s} \tag{4.31}$$

where f_s is the sampling frequency of the waveform. If the time width of the local time-frequency region is be 2~3 times the pitch period [1], the resulting the WGCT exhibits distinct copies of the envelope at multiples of $\frac{2\pi k}{P}$; f_0 is *inversely* related to the number of terms in the WGCT.

4.2.2 Multiple Formants

For multiple formants, we generalize (4.11) as the summation

$$Y(\omega', \omega; k) = W(\omega') \sum_{f=1}^{N_f} e^{jkP(\omega' + \omega_f + \Delta)} \left(\frac{0.5\xi}{\alpha_f + e^{j(\omega' - \Delta)}} + \frac{0.5\xi}{\alpha_f + e^{j(\omega' - 2\omega_f - \Delta)}} \right) \tag{4.32}$$

where N_f is the number of formants. Assuming that the ω_f are well-separated in frequency, we approximate $Y(\omega', \omega; k)$ as being dominated by a *single* formant in local frequency regions. Consequently, identical arguments can be applied as in the previous sections to arrive at modulation models for individual formants. This invokes a sum-of-magnitudes approximation for the magnitude

$$|Y(n,\omega)|_{local} \approx w[n,\omega] \sum_{f=1}^{N_f} E_c[n,\omega; f]\widetilde{H}_f(\omega) \tag{4.33}$$

where $E_c[n,\omega;f]$ and $\widetilde{H}_f(\omega)$ are formant specific. This model interprets local regions of the wideband spectrogram magnitude as a *sum of modulation products*. The WGCT $Y(v,\Omega)$ is a summation of terms corresponding to (4.33) for each formant

$$Y(v,\Omega) = W(v,\Omega) *_{v,\Omega} \sum_{f=1}^{N_f} \sum_{i\epsilon(d,w)} \eta_{R,i}(\Omega;f) \begin{pmatrix} K_{R,i,f}\delta(v) + \\ \sum_{l=1}^{N_s} 0.5\beta_{l_{R,i,f}} \delta\left(v \pm \frac{2\pi l}{P}\right) \end{pmatrix} \quad (4.34)$$

where $\eta_{R,i}(\Omega;f)$, $K_{R,i,f}$, and $\beta_{l_{R,i,f}}$ are now formant-dependent versions of those in (4.30). We expect this approximation to be best for frequency regions for when Δ is either "small" or "large" (i.e., very near or very far from the formant peak) e.g., $\omega_f - \Delta < \omega < \omega_f + \Delta$, analogous to the single-formant case. Furthermore, at frequency regions far away from formant frequencies, the summation implies dominance by a "large Δ" model (4.24) corresponding to a single formant. Nonetheless, if formants interact within a local frequency region, the model can be expected to less accurate. In our subsequent analyses, we show the effects of such interactions.

4.2.3 Simulations

Single Formant: Herein we illustrate properties of the *carrier* models proposed for the previously described small and large Δ conditions. We synthesize a decaying sinusoid $h'[n]$[5] with $\omega_f = 0.1\pi$ corresponding to a periodicity of 20 samples, $\xi = 1$, and $\alpha_f = 0.01$ (4.5); $h'[n]$ is excited with a pure impulse train $p'[n]$ with periodicity $P = 77$ to generate $y'[n]$. Signals are synthesized at 16 kHz with resulting pitch (formant) frequency of 210 Hz (800 Hz). Wideband spectrograms are computed using a Hamming window $w[n]$ with length $L = 40 \equiv 2.5$-ms Hamming; to account for an extremal case of a 350-Hz pitch, L can be chosen in general to be less than $\sim\frac{1}{350} = 2.9$ ms. A single-sample frame rate and 2048-point discrete Fourier transform (DFT) is applied to both $y'[n]$ and $p'[n]$ to obtain $|Y'(n,\omega)|$ and $|P'(n,\omega)|$. WGCT analysis was performed using region sizes of 37.5 ms by 500 Hz extracted with a 2-D Hamming window followed by a 512-by-512-point 2-D DFT. Analogous to the choice of L, 2~3 times the lowest pitch period of 60 Hz constrains the time width to ~33 to 50-ms. We refer the reader to our subsequent discussion to motivate the choice of frequency widths.

For "small-Δ", we extract time slices from $|Y(n,\omega)|$ at $\omega = \omega_f$ and $\omega = \omega_f + \Delta$ with $\Delta = 0.0313\pi$ (corresponding to 250 Hz). $\omega = \omega_f$ ($\Delta = 0$) represents the *idealized carrier* in the modulation model as discussed in (4.14); $\Delta = 0.0313\pi$ represents a "small-Δ" condition. Time slices are normalized to have unity amplitude and shown in Figure 4-5b. We plot absolute differences between the slices and compute the root-mean-squared error (RMSE) across time. Consistent with the model, both time slices resemble decaying exponentials smoothed by the window with the $\Delta = 0.0313\pi$ case having RMSE of ~0.09 relative to the $\Delta = 0$ case. This discrepancy is reflects limitations of the small-Δ assumptions used in modeling.

For "large-Δ", Figure 4-5c shows a time slice extracted at $\omega = 0.5\pi$ (i.e., "far away" from ω_f). We also plot a time slice $|P'(n,\omega)|$ corresponding to periodically summed windows, i.e., the idealized carrier/excitation component $E_w[n]$. The $\omega = 0.5\pi$ time slice closely matches $E_w[n]$ with RMSE of ~0.05.

[5] We delineate $h'[n]$ (and other signals $p'[n]$, etc.) as the *simulated* versions of their corresponding general forms using a specific set of parameters.

In a second set of simulations, we explore properties of the smoothed formant interpretation of the envelope term of the modulation model (4.20). We replicate a time slice $|Y(n, \omega = \omega_f)|$ across all frequencies to generate an *idealized* 2-D carrier $E'(n, \omega)$; we also compute the time average of all spectral slices in $|Y(n, \omega)|$ and replicate this across time to obtain an *idealized*/reference smoothed envelope term $\tilde{H}(n, \omega) \approx H_r(n, \omega)$. Subsequently, we compute

$$E'_e(n, \omega) = \frac{|Y'(n,\omega)|}{H_r(n,\omega)}, H_e(n, \omega) = \frac{|Y'(n,\omega)|}{E'(n,\omega)}. \tag{4.35}$$

$E'_e(n, \omega)$ and $H_e(n, \omega)$ denote *estimates* of the carrier and envelope components assuming the idealized versions of their counterparts in the factorization (4.3) (i.e., the idealized envelope $H_r(n, \omega)$ and idealized carrier $E'(n, \omega)$), respectively. Figure 4-6 shows $E'_e(n, \omega)$ and time slices corresponding to the decaying and window-based carriers in regions near and far from the ω_f respectively, as can be expected since $H_r(n, \omega)$ varies with frequency only. In addition, $H_e(n, \omega)$ is reasonably matched to $H_r(n, \omega)$ in frequency regions near ω_f though not for ω away from ω_f. This is consistent with our use of the exponential decaying carrier in computing $H_e(n, \omega)$. In addition, $H_e(n, \omega)$ exhibits temporal fluctuations in energy at ω_f. This effect reflects the fact that the assumed envelope $H_r(n, \omega)$ best matches in time regions away from excitation impulses (see Appendix I). Quantitatively, normalized spectral slices of $H_e(n, \omega)$ exhibit an RMSE relative to $H_r(n, \omega)$ up to ~0.09.

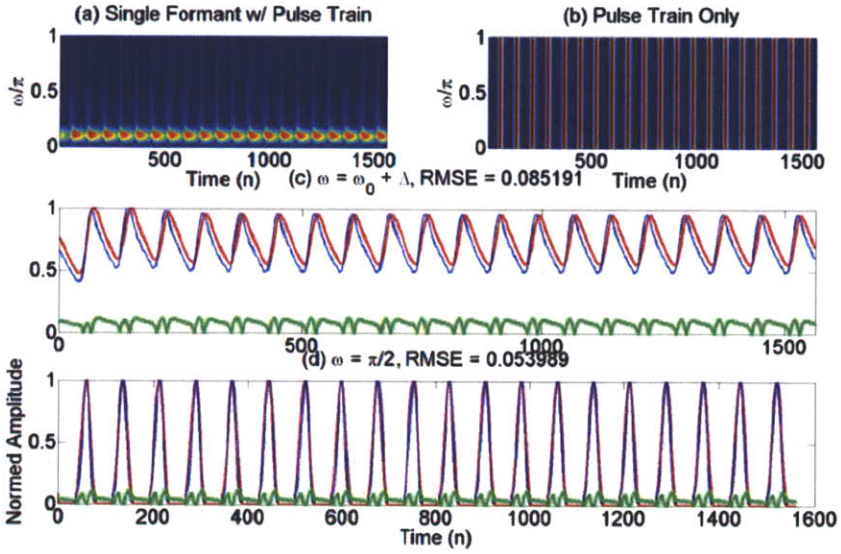

Figure 4-5. Wideband spectrogram (plotted on linear scale) of (a) decaying sinusoid excited with a pure impulse train and (b) pure impulse train; note that a time slice of (b) corresponds to periodically summed copies of the short-time analysis window; (c) time slice of (a) located at the formant peak (red) and for a small Δ value away from the peak (blue); absolute difference (green) between the two curves; (d) as in (c) but for the idealized pure impulse train time slice (red) and actual time slice located "far away" from the formant peak (blue).

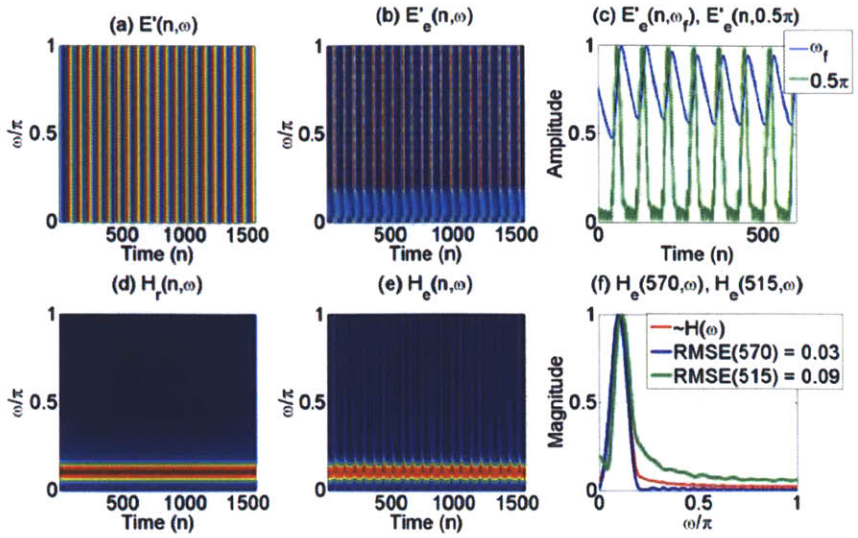

Figure 4-6. (a) $Y(n,\omega_f)$; (b) $E'_e(n,\omega)$; (c) time slices of (b); (d) $H_r(n,\omega)$; (e) $H_e(n,\omega)$; (f) spectral slices of (d) and (e); RMSEs in (f) computed between normalized spectral slices of (e) and the idealized term in (d); in (f), RMSE(570) denotes extraction of a spectral slice at *time* 570.

In a final set of simulations, we assess model properties in frequency regions in between the limiting cases of "small" and "large" Δ. Figure 4-7a shows a local region of $E'_e(n,\omega)$ (4.35) (Figure 4-6b) centered at $\omega = 0.23\pi$ in which two carriers appear to interact within the same local region. The corresponding WGCT contains components *off* the horizontal axis, violating the assumption of a strictly time-dependent carrier ($E_d[n], E_w[n]$). From (4.27), we set each half of the region in frequency to $E_d[n]$ and $E_w[n]$. Observe that the resulting WGCT of this signal does indeed exhibit off-axis similar to those in Figure 4-2f. In Figure 4-7c, we show the result of summing $E_w[n]$ and $E[n]$ *without* applying $R[\omega]$; the resulting WGCT does *not* exhibit off-axis terms, indicating that the displacement effects of $R[\omega]$ corresponding to phase terms in the WGCT are crucial in modeling this behavior. This can understood from (4.30) by noting that the Fourier transform of $E_c[n,\omega]$ has the same form but with $\eta_{R,i}(\Omega)$ replaced by the Fourier transforms of $R[\omega]$ and $R[\omega-\omega_0]$, thereby invoking dependence along Ω in the WGCT domain.

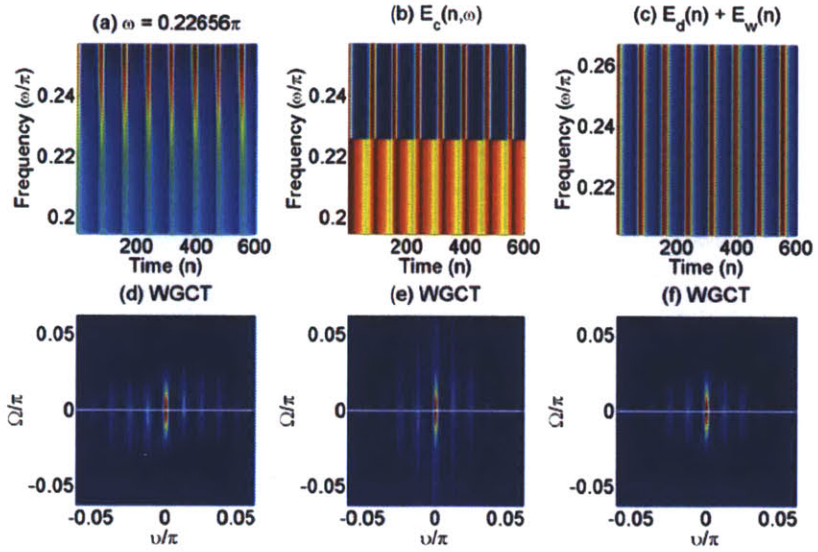

Figure 4-7. (a) Local region from $E'_e(n,\omega)$ centered at $\sim 0.23\pi$; (b) "composite" carrier; (c) carrier obtained from direct summation; (d-f) WGCT of (a-c), respectively with $\Omega = 0$ (line) denoted; plotted on linear scales.

Multiple Formants: Herein we explore properties of the sum-of-modulation-products model (4.33) for multiple formants. In addition, we implicitly investigate effects of downsampling the spectrogram on the model, as is typically done in implementation. Furthermore, we motivate a choice of region size in WGCT analysis along the frequency dimension.

A synthetic vowel generated using a pure impulse train $p[n]$ with a 250-Hz pitch is filtered with a stationary formant structure with frequencies (bandwidths) 669, 2349, 2972, 3500 Hz (65, 90, 156, 200 Hz) to generate $y[n]$ (i.e., a female /ae/ vowel, [28]). Spectrograms are computed as in the previous section though a frame rate of 10 samples (i.e., $\frac{L}{4}$). In addition, we apply a high-pass filter to the spectrogram and aim to recover localized regions using demodulation with bootstrapping as alluded to in Section 4.2.1. For each point along the ω-axis, we extract a region of the filtered spectrogram of time length 37.5 ms and vary the frequency width to obtain local regions (Figure 4-8d). Using demodulation as in Chapter 3, we obtain an estimate of the original local region; we refer the reader to Section 4.6.1 for details of the method and focus here on the results. We compute the root-mean-squared error (RMSE) between the estimate and original 2-D region extracted after both are scaled to have maximum value of unity for comparison purposes.

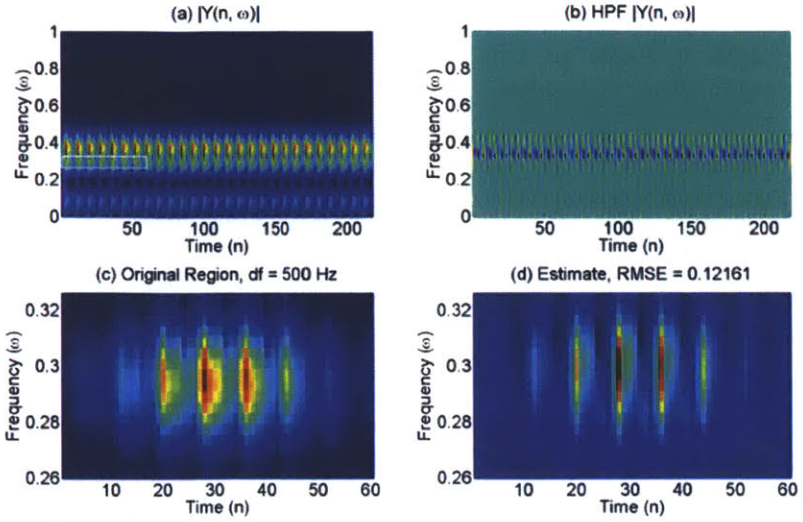

Figure 4-8. (a) Spectrogram of vowel with local region highlighted (white); (b) high-pass filtered version of (a) for use in reconstruction; (c) original local region; (d) estimate of (d) using demodulation; all figures plotted on linear scale.

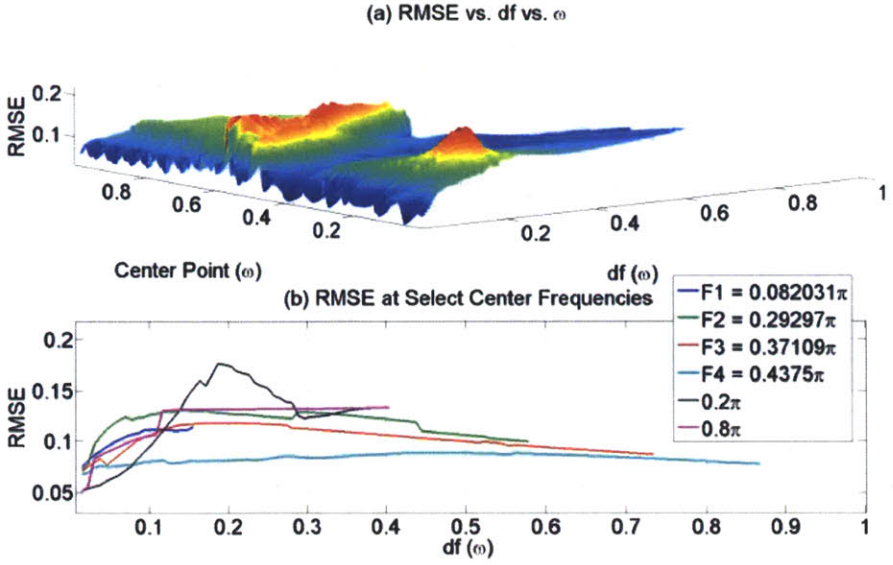

Figure 4-9. (a) RMSE as a function of frequency widths and frequency points analyzed; (b) RMSE for frequency center points corresponding to formant frequencies as well as away from formant frequencies.

Figure 4-9a shows results across all frequency center points and widths (*df*) while Figure 4-9b shows results of analysis at select center frequencies. Despite the presence of multiple formants, RMSEs for reconstructions centered at the formant frequencies do not exceed ~0.15 for frequency widths ranging from zero to 0.1π corresponding to ~800 Hz; qualitatively, this corresponds to a reasonable estimates of the original region as shown in Figure 4-8d. RMSE values generally increase up to a local maximum for larger widths followed by a modest decrease. At frequency regions "far away" from formant peaks, e.g., at $\omega = 0.8\pi$, reconstructions also follow this trend though substantially less growth in RMSE beyond frequency widths of 0.1π; this is due to the

absence of interacting formant structure in these frequency regions. Conversely, at $\omega = 0.2\pi$, the slope of the RMSE is sharper than for the individual and $\omega = 0.8\pi$ case, reflecting effects of formant interactions (here, F1 and F2).

4.3 Extensions to Non-stationary Voiced Speech

4.3.1 Dynamic Formants

Modeling: As discussed in Section 4.9, a Fourier transform view of the wideband spectrograms argues for a similar modulation model to that presented in Section 4.2.1 that includes formant dynamics. In the time-frequency space, we view dynamic formant content as a *rotated* rectangle in the time-frequency space such that the 2-D Fourier transform is the rotation of the 2-D Fourier transform of a rectangle from image processing principles (Figure 4-18) (Chapter 3). While the derived model is posed under relatively restrictive conditions in relation to time segments away from excitation impulse onsets, herein we illustrate with a simple example that the model can nonetheless provide a reasonable interpretation of dynamic formants.

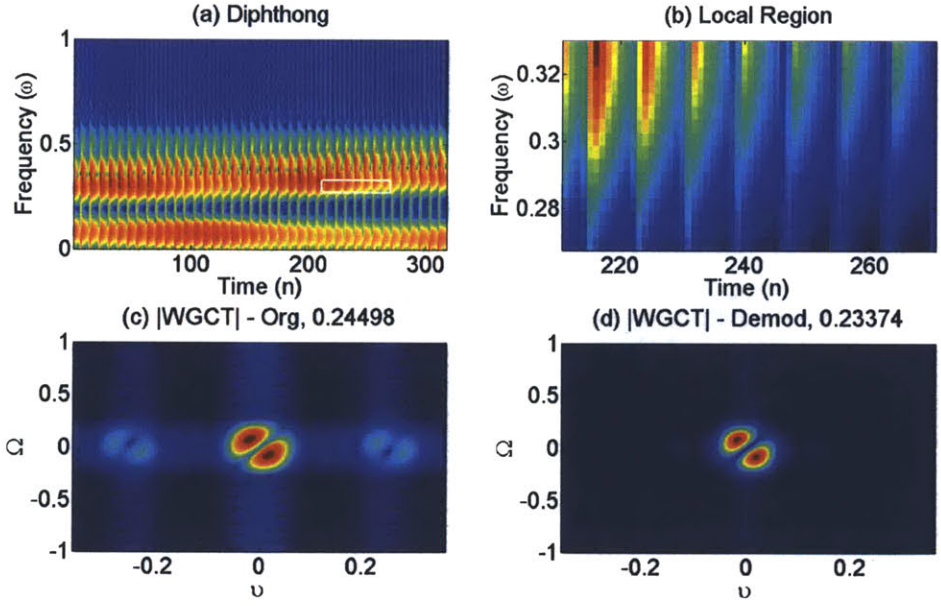

Figure 4-10. (a) Wideband spectrogram of diphthong with local region (white); (b) local region of (a); (c) GCT of (b) with *rotated* (white line) envelope structure near origin; arrows denote demodulation of carrier terms down to DC; (d) WGCT of *demodulated* version of (c) with comparable rotated components to match that in (c). In (c) and (d), DC value is removed for illustrative purposes; in (d), display limited to near-DC region due to presence of cross terms in demodulation.

Dynamic Formant Model Simulations: We synthesize a 200-ms diphthong with start-to-end formant frequencies (bandwidths) of 669, 2349, 2972, 4000 Hz (65, 90, 156, 200 Hz) to 437 2761, 3372, 4000 Hz (38, 66, 171, 200 Hz). The source signal is a pure impulse train with 200-Hz pitch. The wideband spectrogram and WGCTs are computed as in the previous section.

Figure 4-10a-b shows a local region near the increasing second formant. Figure 4-10c shows the corresponding WGCT near the first carrier position; for display purposes, DC values at both the origin and carriers have been removed. At these locations, we observe *rotated* components

corresponding to the local envelope structure present in Figure 4-10b; the rotation of these components can be quantified by measuring the angle of the near-DC peaks relative to the Ω-axis of ~0.24 radians.

As noted in Section 4.2.1, demodulation of envelope content from carrier positions may be used to recover near-DC terms in the WGCT. Figure 4-10d shows an example of demodulating the carrier components in Figure 4-10c to DC. Since in reconstruction we further remove any resulting cross terms by low-pass filtering (see Section 4.6), we restrict our display to the near-DC regions. A set of *rotated* components are obtained at DC with angle ~0.23 radians to match those at DC in Figure 4-10c. These results are consistent with a generalized 2-D envelope $\tilde{H}(n,\omega)$ as argued in Appendix I in relation to the modulation model.

4.3.2 Time-varying Pitch

***Model*:** Time-varying models of pitch have been explored by a number of researchers such as in [41]. In the short-time spectrum, the behavior with time-varying impulse has been described qualitatively as "blurring" (i.e., widening) of harmonic peaks near the "average" pitch; this effect may be interpreted as multiple peaks in the spectrum corresponding to a Bessell function expansion [1]. In our present development, we impose the constraint that local time-frequency regions have time widths such that pitch values are approximately constant. Subsequently, we quantitatively assess the effect this has on a range of pitch variations.

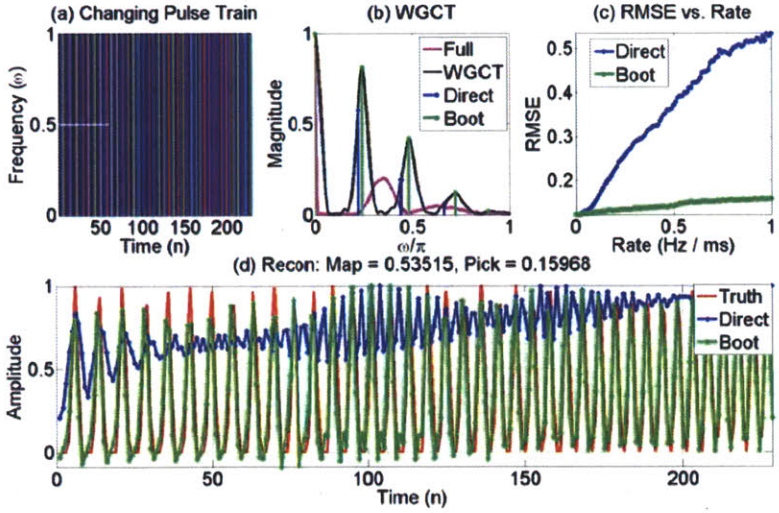

Figure 4-11. (a) Wideband spectrogram of changing pitch with time segment (37.5 ms) denoted (white); (b) WGCT of full time slice (maroon) and time segment of (a) (black); peaks obtained in direct mapping (blue) and bootstrapping (green); (c) RMSE of reconstructions using direct versus bootstrapping methods; (d) reconstruction of 1 Hz/ms case with truth (red), direct (blue), and bootstrapping (green) denoted.

Time-varying Pitch Simulations: We synthesis impulse trains of duration 200 ms with linearly increasing pitch (varied fro m0 to 1 Hz/ms) with starting pitch value of 175 Hz. In analysis, we compute the wideband spectrogram and attempt to resynthesize a full time slice from time segments of size 37.5 ms using 1-D WGCT analysis (Figure 4-11). We extract "peak" locations in the WGCT to resynthesize a sinusoidal series. Peak locations are determined using either 1) the direct pitch information mapping of (4.31) or 2) bootstrapping of the peak locations (Figure 4-11b). In the former method, the pitch value defined at the center of the time segment is used; in

the latter, the mapped locations are reassigned using a 1-D multi-peak picker applied to the WGCT (see Section 4.6 and Chapter 3).

The resulting WGCT shows that the direct mapping can result in "peak" locations that appear harmonically related but deviate from the actual WGCT peaks (Figure 4-11b). As an extremal example of the variation in peak location with time-varying pitch, Figure 4-11b shows a GCT computed for the *full* time slice; we observe two peaks with substantially widened bandwidths consistent with the previously described Bessel-like behavior. We compute the root-mean-squared error (RMSE) between normalized estimates and true time slices. Figure 4-11c shows that RMSE increases dramatically using the direct method for rates > ~0.1 Hz/ms in contrast to the bootstrapping technique. At a rate of 1 Hz/ms, bootstrapping (RMSE = ~0.16) maintains the aperiodicity of the signal while the direct mapping (RMSE = ~0.54) deviates substantially motivates a *bootstrapping* approach to obtain carrier locations that may not correspond exactly to the pitch mapping of (4.31).

4.4 Noise and Onsets/Offsets Models

4.4.1 Noise

Model: We consider now modeling of noisy signals (e.g., fricatives) in the WGCT. The analytical form of the WGCT model of noise is identical to that presented for the Narrowband GCT (NGCT), and we refer the reader to Chapter 3 for more details while focusing on empirical behavior of noise in the WGCT in this section.

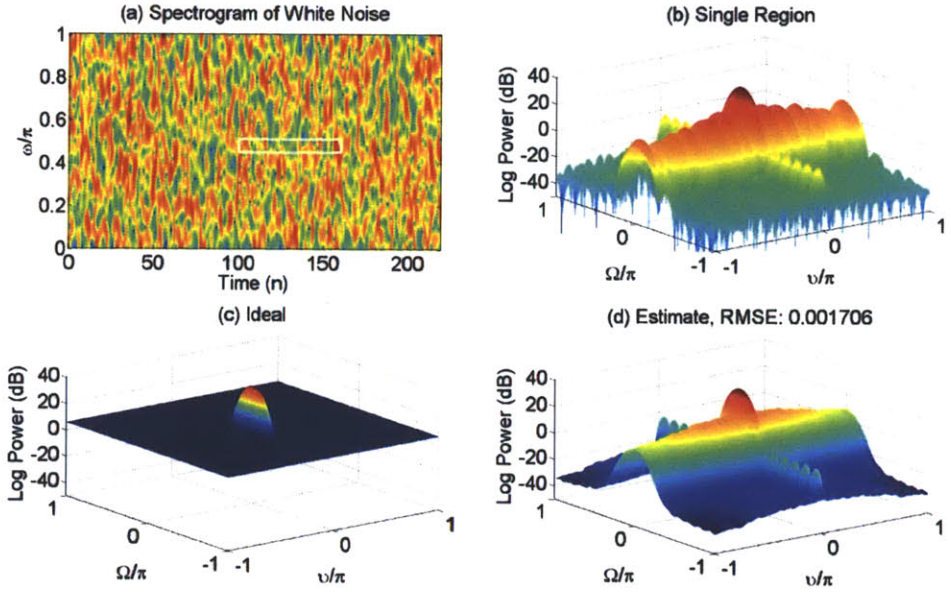

Figure 4-12. (a) Wideband spectrogram of white noise; (b) WGCT of a single region (white); (c) ideal average power spectrum; (d) estimated average power spectrum with RMSE computed for between *normalized* ideal power spectral density and estimate.

A zero-mean independent and identically distributed (i.i.d.) Gaussian process $\boldsymbol{w}[t]$ with standard deviation σ can be analyzed with a wideband short-time Fourier transform magnitude $\boldsymbol{w}[n,\omega]$. We model $\boldsymbol{w}[n,\omega]$ as arising from a 2-D random process under assumptions of i.i.d. time-frequency units with Rayleigh distribution. The (average) 2-D power spectrum of $\boldsymbol{w}[n,\omega]$ is then from Chapter 3

$$S_{ww,GCT}(\upsilon,\Omega) = \left[\frac{4-\pi}{2}\sigma^2 + \frac{\pi}{2}\sigma^2\delta(\upsilon,\Omega)\right] *_{\upsilon,\Omega} |W(\upsilon,\Omega)|^2 \quad (4.36)$$

$$= \frac{\pi}{2}\sigma^2|W(\upsilon,\Omega)|^2 + \frac{4-\pi}{2}\sigma^2\rho$$

$$\rho = \iint_{(-\pi,-\pi)}^{(\pi,\pi)} |W(\upsilon,\Omega)|^2 d\upsilon d\,\Omega \quad (4.37)$$

where $W(\upsilon,\Omega)$ is the 2-D Fourier transform of the 2-D window used to extract localized regions of $\boldsymbol{w}[n,\omega]$.

To obtain an instantaneous model, we invoked in Chapter 3 the Karhunen-Loeve expansion under the assumption of distinct frequency bands of the filterbank view of the 2-D Fourier transform [1]. Specifically, a sum of arbitrary sinusoids on a DC pedestal was viewed as the *carrier* component in the modulation model of (4.29) (and corresponding WGCT):

$$|Y(n,\omega)| = w[n,\omega]\widetilde{H}[n,\omega]E[n,\omega] \quad (4.38)$$

$$E(n,\omega) = K + \sum_{k=1}^{N_c} \alpha_k \cos(\phi_k[n,\omega]) \quad (4.39)$$

$$\phi_k[n,\omega] = \Omega_k(n\cos\theta + \omega\sin\theta) + \varphi_k \quad (4.40)$$

$$Y(\upsilon,\Omega) = W(\upsilon,\Omega) *_{\upsilon,\Omega} \left(\sum_{k=1}^{N_c} K\eta(\upsilon,\Omega) + 0.5\alpha_k\eta(\upsilon \pm \Omega_k\cos\theta, \Omega \pm \Omega_k\sin\,\theta)\right) \quad (4.41)$$

with N_c as the number of carriers, Ω_k is the spatial frequency of the 2-D sinusoid, θ its orientation, φ_k its phase term, and $\eta(\upsilon,\Omega)$ is the 2-D Fourier transform of $\widetilde{H}[n,\omega]$. Here, we have where we have allowed for a 2-D envelope $\widetilde{H}[n,\omega]$ as in the time-varying formant condition. As in the voiced case, this model argues for a *distribution* of envelope content in the WGCT space at carrier locations (see Figure 4-15f).

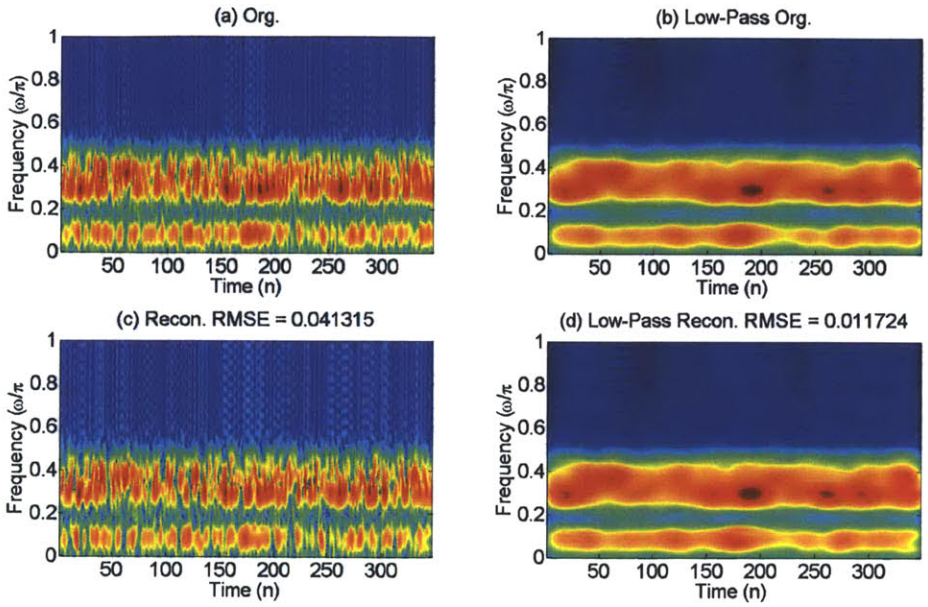

Figure 4-13. (a) Original spectrogram of noise-excited vowel; (b) low-pass filtered version of (a) resulting in envelope term; (c) reconstruction after high-pass filtering and demodulation; RMSE computed between (c) and (a); (d) low-pass filtered version of (c) indicating recovery of low-pass envelope term in (b); RMSE computed between (d) and (b).

Simulations: Herein we compute the empirical 2-D power spectra of white noise in wideband spectrograms for comparison to the proposed model. Figure 4-12a shows a wideband spectrogram computed for $\boldsymbol{w}[t]$ with standard deviation $\sigma = 1$. WGCT analysis was performed using the parameters described previously for vowels. Figure 4-12d illustrates the power spectrum obtained from averaging all regions analyzed. While the model captures the dominance of the near-DC region of the WGCT, it fails to capture substantial 2-D spectral shaping effects. Nonetheless, computing the root-mean-squared error (RMSE) between the normalized (i.e., normalized to have maximum value of 1) ideal power spectral density and estimated power spectral density results in an RMSE of ~0.001, indicating the dominance of the near-DC term. Figure 4-12b shows WGCT analysis results for a single region, consistent with the averaged spectrum in Figure 4-12d. The WGCT has a substantial component along the υ-axis due to correlation across the frequency axis (ω) in the wideband spectrogram such that we observe vertical striations (Figure 4-12a) across time. Specifically, the short-time spectrum is substantially smeared across ω due to the relatively *short* length of the window (and therefore *wide* bandwidth in the spectrum). This behavior is the "dual" of the narrowband GCT that exhibited components along the Ω-axis due to *temporal* correlation effects of processing noise.

In a second set of simulations, we aim to assess the extent to which (4.38) can represent speech-based noise. We compute the wideband spectrogram of a vowel with formant structure as in the previous sections but excited with Gaussian white noise. Next, we adopt the framework of the the simulations for multiple formants in *removing* DC components in the WGCT with the aim of approximately reconstructing them through demodulation (Section 4.7.2). Figure 4-13 shows reconstruction results and low-pass filtering of the original and reconstruction. Observe that the reconstruction results in recovery of the low-pass envelope to match that of the original spectrogram; this is consistent with the demodulation process recovering the near-DC terms of the WGCT from its distributed copies due to the carrier.

4.4.2 Onsets/Offsets

Model: Similar to the noise case, herein we briefly describe onset/offset content observed in wideband spectrograms and similar to that observed for the narrowband case in Chapter 3. An isolated impulse $i[n] = \delta[n - N_0]$ located at N_0 can be modeled as a downsampled short-time analysis window $w[n]$ in the spectrogram domain (denoted as $I[n, \omega]$)

$$I[n, \omega] = w[N_0 - nN] \tag{4.42}$$

where N is the frame rate of the STFT. The GCT is

$$I(\upsilon, \Omega) = W(\upsilon, \Omega) *_{\upsilon} W_n^*\left(\frac{\upsilon}{N}\right) e^{j\upsilon N_0} \tag{4.43}$$

where $*_{\upsilon,\Omega}$ denotes convolution in the GCT domain and $W(\upsilon, \Omega)$ is the 2-D Fourier transform of a 2-D window $w[n, \omega]$ used to extract a localized time-frequency region. As in Chapter 3, we view $I(\upsilon, \Omega)$ as only a portion of an onset/offset *envelope* corresponding to a voicing and/or noise onset/offset through the model of (3.47). As with formant envelopes, we impose a bandlimited constraint on $I(\upsilon, \Omega)$ in the context of modulation (4.30).

Wideband onsets will presumably have a wider bandwidth in the GCT domain than in the narrowband due to the sharper representation of temporally oriented component. As a proxy for bandwidth estimation, we consider that of $I(\upsilon, \Omega)$. Specifically, wideband parameters are a 2.5-ms ($L = 40$) short-time analysis window and frame rate of 0.625 ms. This results in the downsampled Hamming window mainlobe $W_n^*\left(\frac{\upsilon}{N}\right)$ width of $\left(\frac{8\pi}{40}\right)4 = 0.8\pi$ (Chapter 3) in contrast, a 32-ms window and 1-ms frame rate in the narrowband case results in a mainlobe width of $\left(\frac{8\pi}{512}\right)25 = 0.3906\pi$.

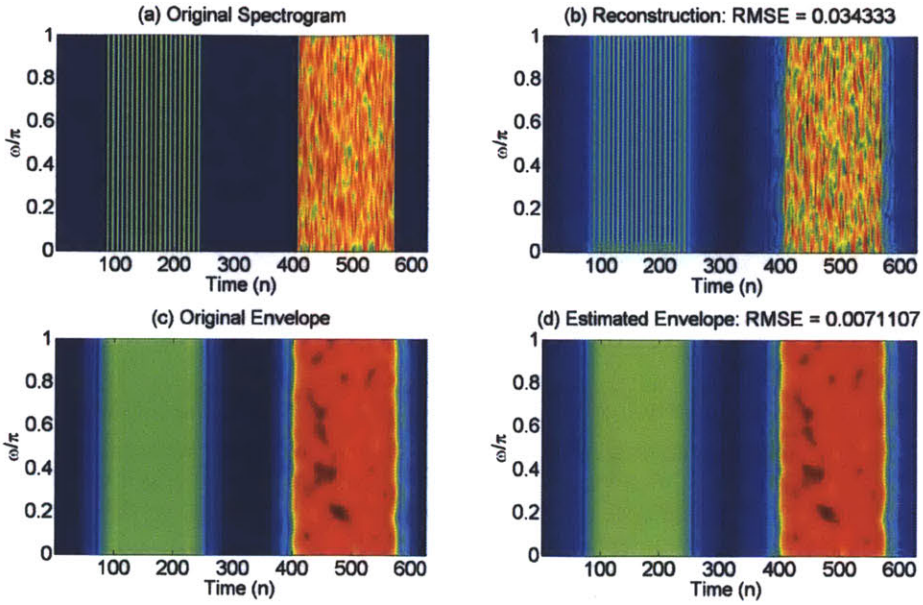

Figure 4-14. (a) Spectrogram of voicing and noise onset; (b) reconstruction of (a); (c) low-pass filtered version of (a) demonstrating onset/offset envelopes; as in (c) but for the reconstruction in (b); associated RMSEs computed after normalization in all cases; log spectrograms plotted to emphasize widening effects.

Simulations: Figure 4-14 shows results of synthesizing and reconstructing voicing and noise onset/offsets; reconstruction was performed using the demodulation technique described in Chapter 3 and in the subsequent Section 4.6.1. The reconstruction in Figure 4-14b exhibits widening of the onsets as may be expected from the bandlimited nature of the analysis/synthesis method. Nonetheless, this widening is consistent with the envelope obtained in low-pass filtering the original signal in Figure 4-14c and as can be shown in filtering the reconstruction in Figure 4-14d.

4.5 A Taxonomy of Speech Signal Behavior in the GCT

Our discussions motivated a *modulation* view of the wideband spectrogram. Specifically, in voiced regions, the wideband spectrogram can be viewed as *summation* of *modulation* components, where each component corresponds to a formant. A carrier $E_c[n,\omega]$ is dependent on source periodicity and (under certain conditions) formant bandwidth and is *modulated* by a smoothed (*single*) formant or envelope $\left|\widetilde{H}_f(n_0,\omega)\right|$. Noise and onsets/offsets are viewed in this framework as carrier and envelope components, respectively.

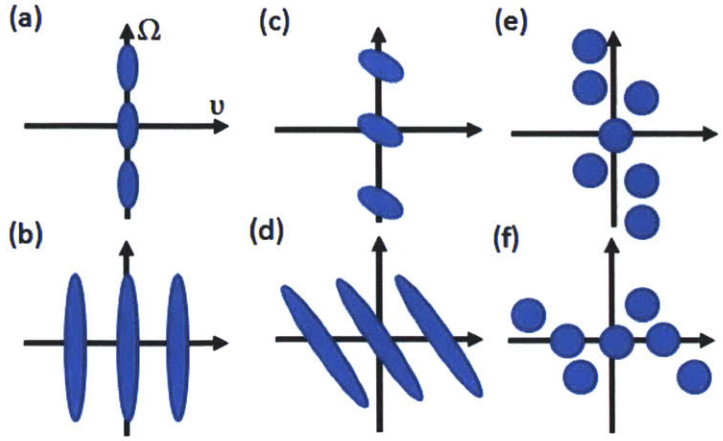

Figure 4-15. Narrow (top) and wideband (bottom) representations of: (a, b) stationary formant and pitch, (c, d) stationary pitch and dynamic formant, and (e, f) noise content.

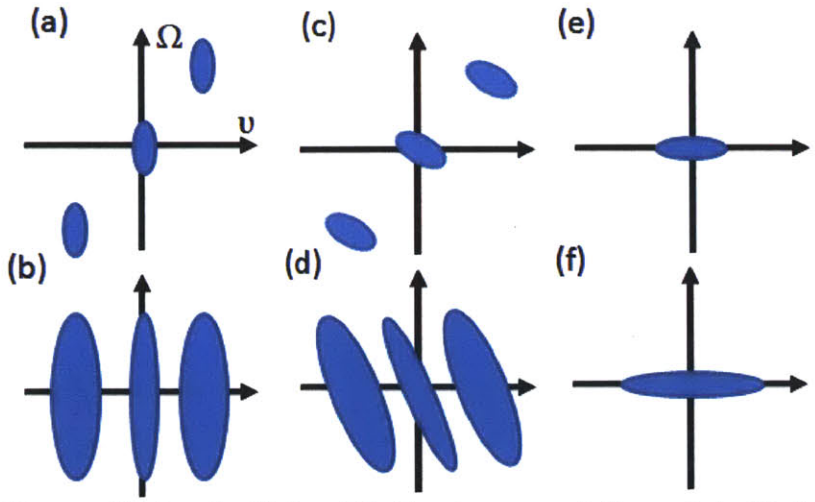

Figure 4-16. Narrow (top) and wideband (bottom) representations of: (a, b) dynamic pitch and stationary formant, (c, d) dynamic pitch and dynamic formant, and (e, f) onset/offset content.

This signal model has some similarity to that proposed for narrowband spectrograms in Chapter 3, and in subsequent sections, we assess its ability to represent speech content using algorithms similar to those used in Chapter 3. Nonetheless, there important distinctions exist in the form and interpretation of the two models, and in Figure 4-15 – Figure 4-16 we compare the mapping of changing/stationary, pitch/formant, and noise and onset/offset content for both representations.

For voiced speech, stationary pitch mappings in the NGCT and WGCT are "duals" of each other along the Ω (narrow) and υ (wide) axes, as schematized in Figure 15a-b. This mapping distinction is preserved even when formant dynamics are introduced (Figure 4-15c-d). In contrast, pitch dynamics invokes a rotation of components in the NGCT while invoking widening of the formant content along the υ-axis in the WGCT due to the presence of widened harmonic content of the carrier as schematized in Figure 16a-d. An additional narrowband/wideband "duality" is observed in mapping noise to the GCT domain with components along the Ω (narrow) and υ (wide) axes (i.e., $\upsilon = 0$ and $\Omega = 0$), respectively (Figure 15e-f. Finally, the

WGCT exhibits greater bandwidth of onset/offset content relative to the NGCT due to differences in short-time analysis resolution (Figure 16e-f).

Table 4-1 presents a taxonomy of speech signal behavior as represented in the narrowband/wideband models. We denote $H(n,\omega)$ as the formant structure, $g(.)$ as a general function, and ω_c as the center frequency of the local region analyzed. Several distinctions include the summation of (WGCT) vs. singular modulation products (NGCT) and single (NGCT) vs. multiple carrier types (WGCT); in addition, carriers have distinct dependencies on source periodicity f_0 (NGCT, WGCT), pitch dynamics $\frac{df_0}{dt}$ (NGCT), formant bandwidth α_f (WGCT), and ω_c. "Dual" behavior exists in pitch mappings between the two GCTs; specifically, *high* pitch values results in *low* (i.e., near GCT origin) frequency components in the NGCT and *high* frequency components in the WGCT. This effect also results in the difference in number N_p of harmonic terms in the GCT as they relate to pitch. While noise is viewed as a carrier term in modulation in both representations, its localization is distinct between the two as previously noted. Finally, onsets/offsets are interpreted as envelope terms in both cases though with differences in bandwidth v_a along the v-axis.

Table 4-1. Comparison of signal model interpretations for narrow- and wideband-based Grating Compression Transforms.

Interpretation/GCT	Narrowband	Wideband
Local Model	$Y(n,\omega) = H(n,\omega)E(n,\omega)$	$Y(n,\omega) = \sum_{f=1}^{N_f} \widetilde{H}_f(n,\omega)E_c(n,\omega)$
Envelope (vowels)	$H(n,\omega)$	$\widetilde{H}_f(\omega,n;\omega_c) \approx \lvert H_f(\omega,n) * W(\omega)\rvert$
Carrier (vowels)	$E(n,\omega;f_0,\frac{df_0}{dt},\omega_c)$	$E_c(n,\omega;f,f_0,\alpha_f,\omega_c) = g(E_d(n),E_w(n),R(\omega))$
f_0 mapping	$\Omega_0 \propto \frac{1}{f_0}$; $v_0 \propto \frac{df_0}{dt}$; ω_c	$v_0 \propto \frac{1}{f_0}$
$f_0 \rightarrow N_P$	$N_P \propto f_0$	$N_P \propto \frac{1}{f_0}$
Noise	Along $v = 0$; carrier	Along $\Omega = 0$; carrier
Onsets/Offsets	$v_a = 0.39\pi$	$v_a = 0.8\pi$

4.6 Spectrogram Analysis/Synthesis and Co-channel Speaker Separation

Herein we describe approaches to test the proposed model's ability to represent speech content through spectrogram analysis/synthesis and co-channel speaker separation. As these methods are generally the same algorithmically to those in Chapter 3 and [21], we refer the reader to those works for details and focus here on the general framework and distinctions.

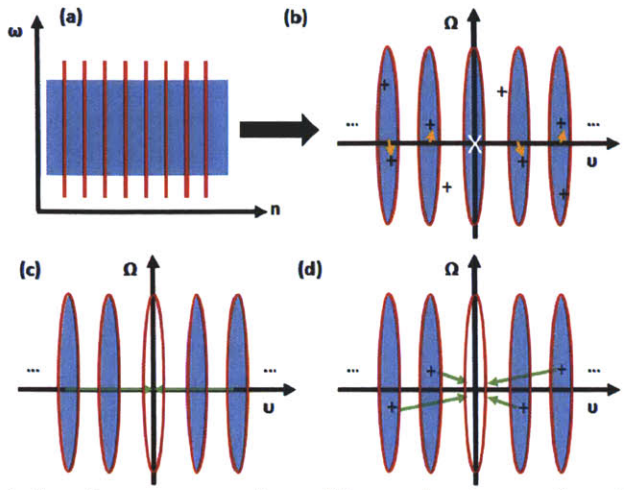

Figure 4-17. (a) Local time-frequency region with carrier (orange) and envelope (shaded) components; (b) corresponding WGCT with candidate peaks from peak-picking ('+') and *reassignment* of directly mapped carrier locations to candidate peaks; 'x' denotes removal of near-DC term; (c) demodulation of components located at carrier locations obtained from *direct* mapping for reconstruction; (d) as in (c) but using *reassigned* carrier locations (bootstrapping).

4.6.1 Analysis/Synthesis

In the proposed signal model, the WGCT domain consists of envelope content near the origin and at carrier locations (Figure 3-12, Figure 4-17) due to sinusoidal-series-based modulation. As a framework for reconstruction, we aim to approximately *recover* the near-DC terms in the GCT using their modulated version at carrier locations using sinusoidal *demodulation* [21] (Chapter 3) (Figure 4-17). Synthesized carriers are multiplied by the local region followed by low-pass filtering to invoke the bandlimited constraint along the υ-axis of the envelope terms in the GCT domain (4.30). Demodulation is done locally across time-frequency regions via a least-squared-error fitting method. The reconstructed spectrogram is combined with the phase of the original signal to estimate a waveform using overlap-add. This waveform estimate represents an "upper limit" of reconstruction due to inclusion of the phase of the original signal. Finally, as a reference approach, we apply fitting with a sinusoidal series (i.e., without the envelope) as was done in Section 3.3 using known carrier locations.

To obtain carrier parameters for voiced speech, we use the pitch mapping (4.31) in conjunction with prior pitch information. In contrast to the narrowband model, a *direct* mapping *forces all carriers to be located on the υ-axis*. For unvoiced speech and in the *bootstrapping* method to be subsequently discussed, peak-picking is done using a multi-peak picker similar to that of Chapter 3. The GCT magnitude is analyzed by a series of binary masks to extract peak locations based on a point's neighbors and amplitude thresholding.

As described in Chapter 3, carrier assignments for demodulation are made for voiced speech using a *direct* method with mapped locations (for voiced speech). In the *bootstrapping* method, directly mapped carrier locations are reassigned to those obtained from peak-picking using a minimal distance criterion in an iterative algorithm (see Section 3.2). Noise carriers are assigned based on peak-picking in both direct and bootstrapping approaches.

4.6.2 Co-channel Speaker Separation

As mentioned in the previous section, one motivation for analysis/synthesis with recovering the near-DC terms from their modulated versions is the separation (or removal) of interfering speakers. Specifically, we assume according to our model that near-DC terms of multiple speakers overlap, while carrier terms often do not, and that recovery of the (uncorrupted) DC region must be consistent with modulation of the carriers.

WGCT Approach: In our WGCT-based approach, similar but not identical to the narrowband GCT (NGCT) approach, we apply least-squared-error demodulation using the *sum* of two modulation models to fit local time-frequency regions of the mixture spectrogram. As in Chapter 3 and [21], this framework utilizes a sum-of-magnitudes approximation to the mixture spectrogram. Diagonal loading of the resulting least-squares matrix was performing using a threshold value obtained from a held-out development set. Carrier parameters are obtained as in the single-speaker case using direct mapping and peak-picking. Permutations of mixture voicing conditions are used to assign carriers to distinct speakers for demodulation (Chapter 3). In the *voiced-on-voiced* case, the pitch mapping of (4.31) is used to obtain carrier positions that are used *directly* or as reference values for *bootstrapping*/reassignment as in the single-speaker case using candidates from the peak-picker. In the *voiced-on-unvoiced* case, the *direct* pitch mapping is used to obtain the voiced speaker's carriers while the unvoiced speaker is assigned to carrier locations from peak-picking. In *bootstrapping*, the voiced speaker's carriers are first reassigned while the remaining candidate carrier locations from peak-picking are assigned to the unvoiced speaker. In the *unvoiced-on-unvoiced* case, carrier positions from peak-picking are used to fit the local region; the resulting estimate is halved and assigned to both speakers. A distinction of the WGCT approach from the NGCT is that we apply *bootstrapping* of the carrier positions as an alternative method to the *direct* approach instead of the exclusion/re-estimation method described in Chapter 3. Finally, as in Chapter 3, we use for a reference a sinusoidal-based separation system that operates on a frame-by-frame basis as a reference.

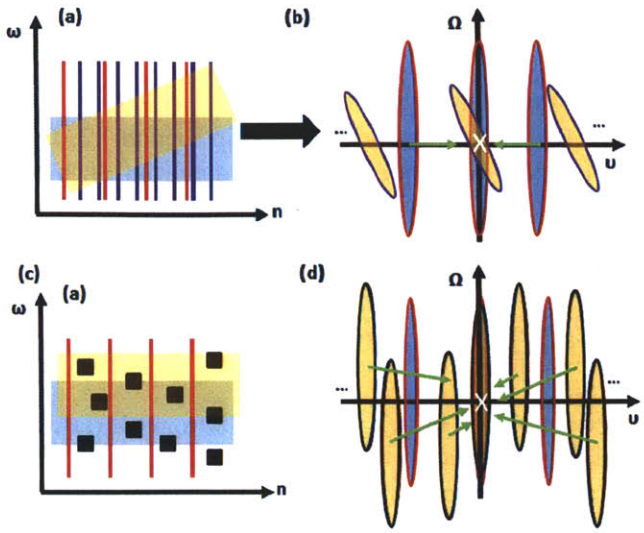

Figure 4-18. (a) Local region of wideband spectrogram for voiced speaker1 (red lines, shaded blue) and voiced speaker2 (purple lines, shaded yellow) mixture; (b) corresponding WGCT with removal of near-DC terms and demodulation to extract speaker1; (c) voiced speaker1 (red lines, blue shaded) and unvoiced speaker2 (black squares, yellow shaded) mixture; (d) WGCT of (c) indicating removal of near-DC terms and demodulation to recover speaker2. Demodulation in (b) and (d) are illustrated for the *direct* approach though this is done similarly in bootstrapping with reassigned carrier positions.

Fusion Methods: From Section 4.5, recall that the *number* of harmonic components in the GCT depends on the short-time analysis window. For instance, male speakers with low pitch exhibit fewer terms in the NGCT than females while the opposite is true for WGCT; this effect was suggested in Chapter 3 as contributing to differences in performance in both analysis/synthesis and separation. It is conceivable that a *fusion* of separation estimates from both the NGCT and WGCT can lead to better overall estimates. We consider a simple fusion method using a weighted sum (here, $0 \leq \alpha \leq 1$)

$$\hat{x}_{fused}[n] = \alpha \hat{x}_{wide}[n] + (1-\alpha)\hat{x}_{narrow}[n]. \tag{4.44}$$

Finally, as was done in Chapter 3, we fuse both the narrow and wideband methods with the baseline sinusoidal-based separation system with two weighting parameters α and β ($0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$)

$$\hat{x}_{fused}[n] = \alpha \hat{x}_{wide}[n] + \beta \hat{x}_{narrow}[n] + (1-\alpha-\beta)\hat{x}_{SB}[n]. \tag{4.45}$$

4.7 Evaluation

4.7.1 Data Set

In spectrogram analysis/synthesis and speaker separation, we use data from TIMIT identical to that in Chapter 3. For analysis/synthesis, 10 males and 10 females speaking 2 distinct utterances are used for a total of 40 examples. In separation, the development set consists of 5 male-male (MM), female-male (FM), and female-female (FF) mixtures while the test set consists of 24 male-male (MM), 24 female-female (FF), and 64 female-male (FM) mixtures, all mixed at 0 dB overall signal-to-signal ratio. The selected pairs cover a large range of overlapping voiced and unvoiced conditions, including crossing pitch tracks. Pitch tracks for individual utterances are obtained

using a correlation-based analysis followed by dynamic programming from the Wavesurfer package [38].

4.7.2 Spectrogram Analysis/Synthesis

Wideband spectrograms ($s_{full}[n,\omega]$) are computed as in Section 4.2. GCT analysis is done using a 2-D 512-point DFT on local time-frequency regions of size 500 Hz by 37.5 ms extracted using a 2-D Hamming (overlap factor 4). A high-pass (low-pass) 1-D filter $h_{hp}[n]$ ($h_{lp}[n]$) is designed using the frequency sampling method [29] $h_{hp}[n]$ ($h_{lp}[n]$) of order 80 with pass-band (stop-band) beginning at

$$0.5v_b = 60\frac{2\pi(0.625\times10^{-3}\times16000)}{16000} = 0.075\pi \tag{4.46}$$

corresponding to an extremal low-pitch case of 60 Hz from (4.31) with stop-band (pass-band) roll-off to v_b. $h_{hp}[n]$ is applied to $s_{full}[n,\omega]$ to obtain $s_{full,hp}[n,\omega]$. $s_{full,hp}[n,\omega]$ is multiplied by a set of sinusoidal carriers followed by low-pass filtering by $h_{lp}[n]$ to obtain envelope estimates which are used to fit gain parameters in a least squares formulation. In demodulation, we used $s_{full,hp}[n,\omega]$ instead of $s_{full}[n,\omega]$; this was observed in preliminary experiments to reduce the influence of cross terms near WGCT origin after demodulation such as in the case of low-pitch values (e.g., for males).

As a goodness metric, we compute root-mean-squared errors (RMSE)

$$RMSE = \sqrt{\frac{1}{N\omega_N}\sum_{n=1}^{N}\sum_{\omega=1}^{\omega_N}\left[s'_{full}[n,\omega] - \hat{s}'_{full}[n,\omega]\right]^2} \tag{4.47}$$

where ω_N denotes the total number of DFT frequency bins in the spectrogram and $s'_{full}[n,\omega]$ and $\hat{s}_{full}[n,\omega]$ are the original and reconstructions, respectively, normalized to have maximum value of unity. In addition, we compute the signal-to-noise ratio (SNR)

$$SNR = 10log\left(\frac{\sum_n x^2_{single}[n]}{\sum_n\left[x_{single}[n]-\hat{x}_{single}[n]\right]^2}\right) \tag{4.48}$$

where $\hat{x}_{single}[t]$ is waveform estimated obtained in combining $\hat{s}_{full}[n,\omega]$ with the phase of the original signal [1] (Chapter 3).

Figure 4-19 shows results of a single female utterance. For reference, we show also an error spectrogram computed as the absolute difference between the bootstrap and true spectrograms after normalization. One limitation of the demodulation approach (in both bootstrapping and direct methods) is a "smoothing" effect on onset/offset structure, presumably due to bandlimiting of the envelope term in the proposed modulation mode (Figure 4-19, time 750). In addition, both methods fail to capture aperiodic content such as at time 500 as may be due to glottalization [1]. For voiced speech, the "enforcement" of periodic carriers and their use as guides for reassignment in bootstrapping are evidently insufficient to fully address these effects.

We show in Figure 4-20 the full spectrograms and Figure 4-21 and Figure 4-22 individual spectral and time slices comparing the sinusoidal fit and demodulation methods. Here, we refer to "sinusoidal fit" as the two-dimensional sinusoidal-series fit for analysis/synthesis described in Section 3.3. Observe that the sinusoidal fit is unable to capture the sharpness of the initial formant peak in Figure 4-21 similar to that shown in Chapter 3. We can expect that due to the

smoothing of the formant envelope from the window that this effect would be less pronounced in the wideband reconstructions relative to the narrowband reconstructions. Indeed, in the extremal case of a *flat* envelope, demodulation is equivalent to fitting with a sum of sinusoids. Nonetheless, the present results demonstrate that a more general *envelope* component as described in the modulation model is necessary for representing spectral shaping effects in the wideband spectrogram. In addition to spectral shaping effects, observe in Figure 4-22 that the sharpness of the onset is better modeled through demodulation than through the sinusoidal fit. In the waveform (Figure 4-23), the latter effect results in generation of noisy content prior to onsets using the sinusoidal method.

Overall, reconstruction results demonstrate that speech content is generally well-represented by the modulation model with errors values ranging from 7e-3 to 4e-2 on a scale of unity as the maximum value. Quantitatively, bootstrapping appears to modestly outperform the direct method (Table 4-2) using the RMSE metric. Nonetheless, this is not reflected in the resulting waveforms in SNR, presumably due to phase effects in reconstruction. Quantitatively, bootstrapped demodulation outperforms the sinusoidal fits in all cases. In informal listening, (non-author) subjects did not distinguish waveform reconstructions between demodulation methods employing the direct mapping, bootstrapping and the original. Onsets were observed to be sharper using the demodulation method reflecting the original waveform than in the sinusoidal fit method, consistent with the the higher SNR values using both demodulation methods and the poorer reconstruction of the onsets in the waveforms shown in Figure 4-23.

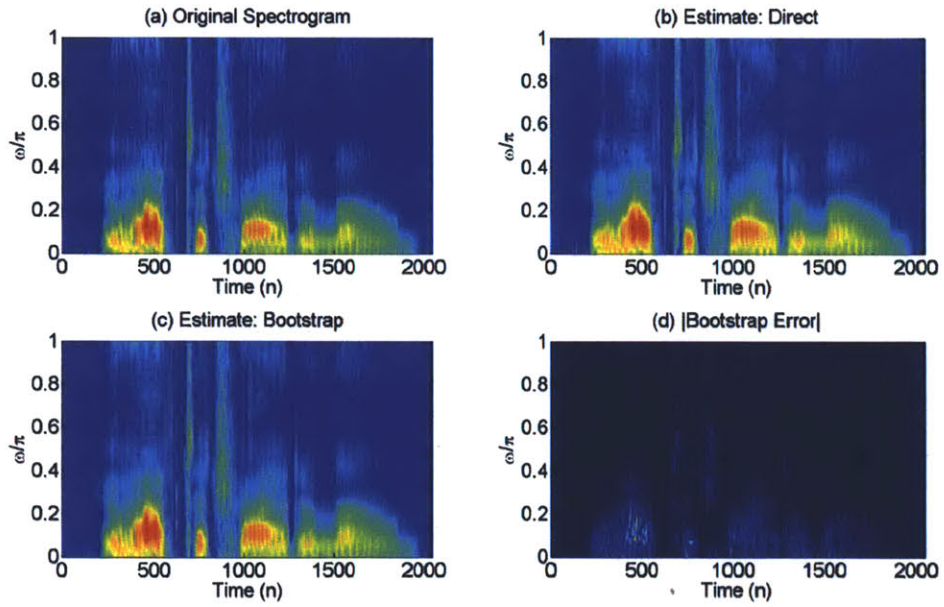

Figure 4-19. (a) Original spectrogram of female utterance "You'll have to try it alone."; (b) reconstruction using direct method; (c) reconstruction using bootstrapping method; (d) Absolute error spectrogram computed as the absolute value of the difference between (b) and (a).

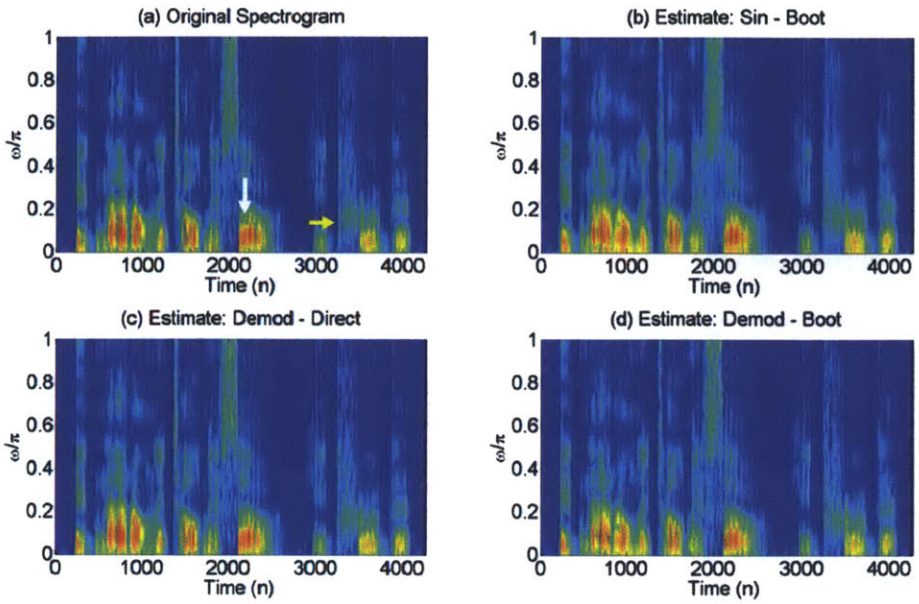

Figure 4-20. (a) Original spectrogram of male utterance "He'd not only told me so, he'd proved it."; (b) reconstruction using a 2-D sinusoidal-series fitting method (see Section 3.3) with bootstrapping (c) reconstruction using demodulation (direct) and (d) bootstrapping. Extraction of spectral slice (white arrow, a) and time slice (yellow arrow, a) for Figure 4-21 and Figure 4-22, respectively.

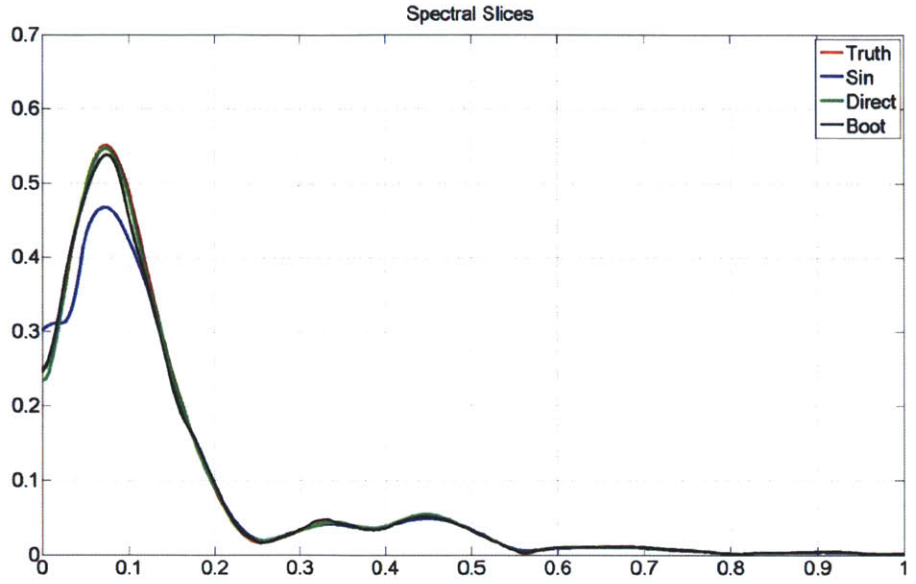

Figure 4-21. Spectral slice extracted from Figure 4-20 with 2-D sinusoidal-series fitting with reference spectrum (red), sinusoidal fitting with bootstrapping (blue), and demodulation with direct mapping (green) and bootsrapping (black).

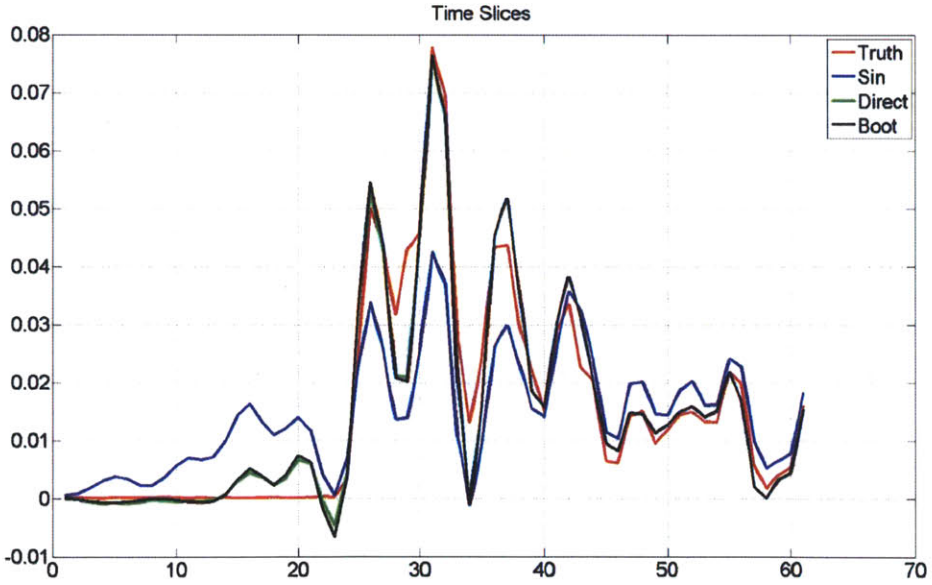

Figure 4-22. Time slice extracted from Figure 4-20.

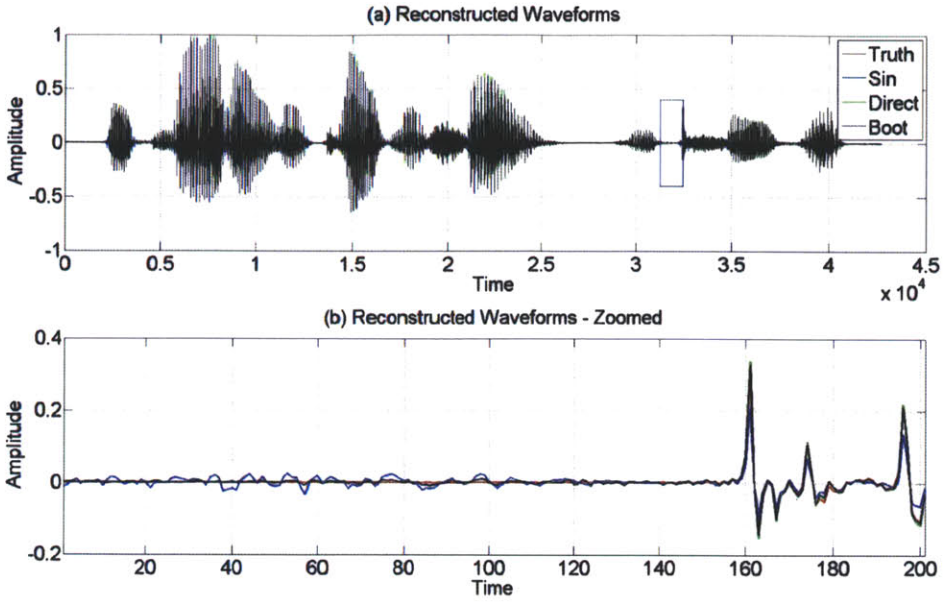

Figure 4-23. Reconstructed waveforms from Figure 4-20 showing additive noise content from sinusoidal fitting (a, b).

Table 4-2. Average RMSE and SNRs for analysis/synthesis of spectrograms and standard errors.

	Direct – Demod.	Boot – Demod.	Direct - Sin	Boot – Sin
RMSE (Males)	4.35e-2 [9.54e-3]	6.80e-3 [5.01e-3]	9.09e-3 [5.86e-4]	5.05e-2 [1.11e-2]
RMSE (Females)	3.32e-2 [6.09e-3]	7.92e-3 [5.44e-4]	1.27e-2 [9.64e-4]	3.58 [6.4e-3]
SNR (dB) (Males)	24.61 [0.46]	22.15 [0.18]	19.78 [0.37]	20.59 [0.33]
SNR (dB) (Females)	21.89 [0.39]	23.04 [0.43]	18.94 [0.40]	19.81 [0.41]

4.7.3 Co-channel Speaker Separation

In speaker separation, the mixed signal $x_{mix}[n]$ is analyzed with short-time and GCT parameters identical to those in analysis/synthesis. We compute RMSE errors in the spectrogram estimate as in analysis/synthesis for applications such as pre-processing for speech recognition. For human listening, reconstructed spectrograms are combined with the phase of the mixed signal to obtain a waveform estimate. We denote waveform estimates as $\hat{x}_i[n]$, and we compute the signal-to-interferer ratio (Chapter 3)

$$SNR_i = 10log\left(\frac{\sum_n x^2{}_i[n]}{\sum_n [x_i[n]-\hat{x}_i[n]]^2}\right) \tag{4.49}$$

where $x_i[n]$ is the original (unmixed) utterance. In fusion, the α parameter was swept on the development set from 0 through 1 with a step size of 0.01; we used the exclusion method from Chapter 3 and the bootstrap method the narrow and wideband estimates. The α value corresponding to the highest average SNR across all waveforms was used in testing. We obtained a "best" value of $\alpha = 0.41$ to be applied in testing. In fusing with the sinusoidal-based system we obtained a "best" weighting of $\alpha = 0.23$, $\beta = 0.38$. The relative contribution to the final

waveform is therefore ranked in the following order based on this weighting: sinusoidal-based, narrowband estimate, wideband estimate.

Figure 4-24 and Figure 4-25 show the results of wideband-based speaker separation. Demodulation is capable of suppressing harmonic content from an interferer, thereby leading to separation of speakers; observe for instance at time 700 in Figure 4-24a harmonic content from the interfering speaking suppressed in Figure 4-24c and d. A limitation in separation can be observed in Figure 4-25 naer time 1700 where the onset of the target is poorly replicated in the estimate. As in analysis/synthesis, this is likely due to bandlimiting of the envelope term in demodulation. Quantitatively, separation can result in RMSEs on the order of 3e-2 (on a scale of unity as the maximum value) and 4~6 dB global SNR gains across all permutations of mixtures (Table 4-3, Table. 4-4). In general, bootstrapping appears to provide modest gains over the direct method. In informal listening, (non-author) subjects reported good reconstruction of the target speaker with suppression (but not complete removal) of interfering speakers using both the bootstrapping and direct methods.

In Figure 4-26, we consider our fusion results between the narrowband and wideband estimates. Observe in this example that although the narrowband estimate provides better estimates overall, the wideband estimate provides complementary information in better suppressing content from an interferer at time ~2.6e4. Fused between the narrowband and wideband estimates were reported in informal listening (non-authors) to exhibit less "abrupt" insertions of interfering speakers, consistent with the overall gains observed in average SNR values up to ~ 1 dB relative to either estimate alone.

Finally, in Figure 4-27, we show the results of fusing the narrowband, wideband, and the reference sinusoidal-based (i.e., frame-based) method for a mixture of two male speakers. Observe that near time 6000 that the sinusoidal-based method exhibits residual periodic structure from an interfering speaker; this is not the case in the narrowband and wideband estimates. Fusion of the three waveform results in suppression of this periodic component and an SNR gain of ~1 dB. In informal listening, the non-author listeners reported a reduction of the periodic component of the interfering in the fused waveform relative to the sinusoidal-based estimate, consistent with these observations. This gain is consistent with the overall gains obtained across mixture components as summarized in Table. 4-4.

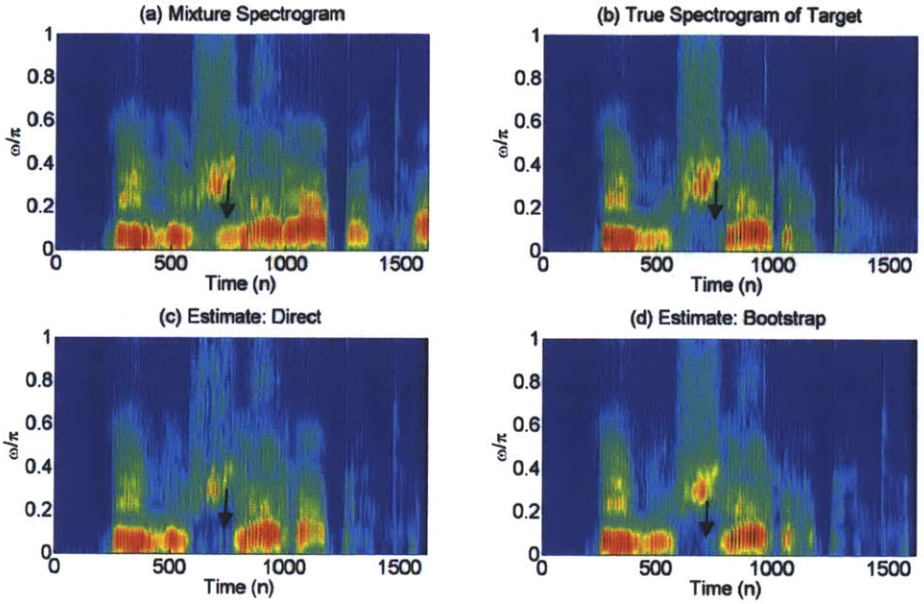

Figure 4-24. (a) Mixture spectrogram of a male ("They were shattered.") and female ("Neither his appetite"); (b) true male target; (c) male estimate using direct method; (d) male estimate using bootstrap method. Observe suppression of harmonic content near time 800 due to demodulation in (c) and (d) relative to (a) (arrows).

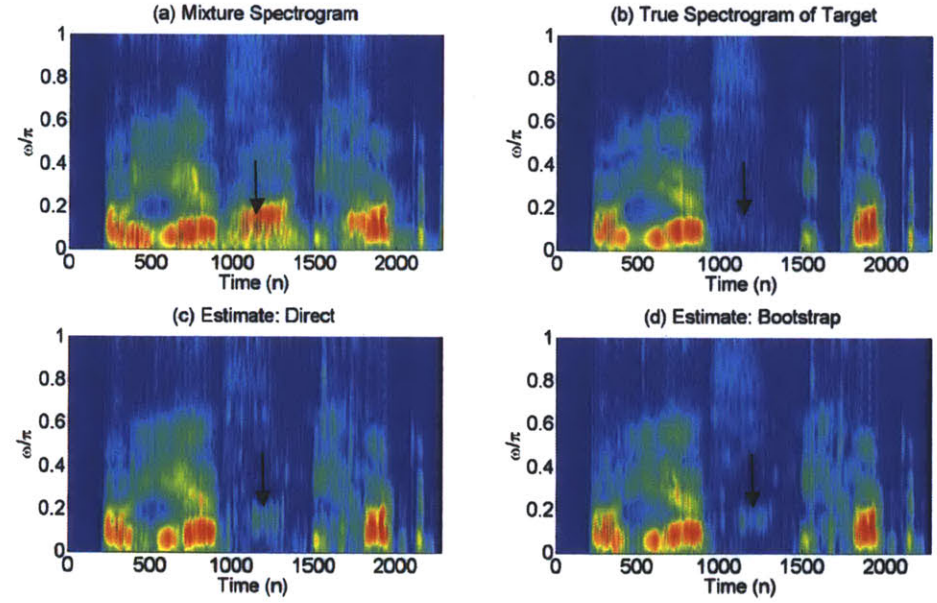

Figure 4-25. As in Figure 4-24 but for two female mixtures ("Oh yes, he talked", "Anything wrong captain?") with "talked" target utterance. Observe suppression of harmonic content near time 1200 due to demodulation in (c) and (d) relative to (a) (arrows).

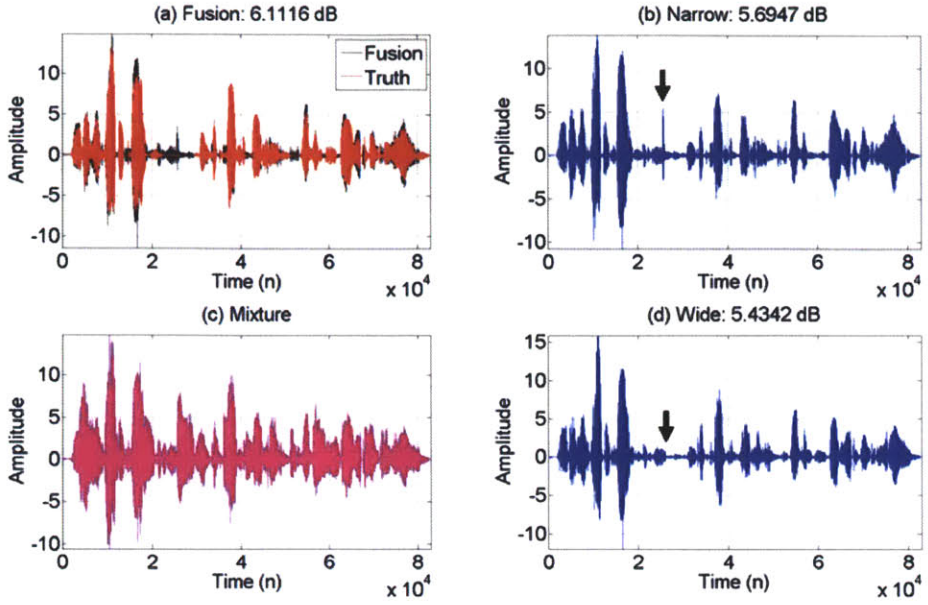

Figure 4-26. (a) Fusion estimate between narrowband and wideband estimates and truth target utterance "appetite"; (b) narrowband estimate of target; (c) mixture waveform of two females ("Neither his appetite, his exacerbations, nor his despair were akin to yours." + "Forty-seven states assign or provide vehicles for employees and state business.") (d) wideband estimate of target; note suppression in (d) of outstanding interferer in (b) (arrows).

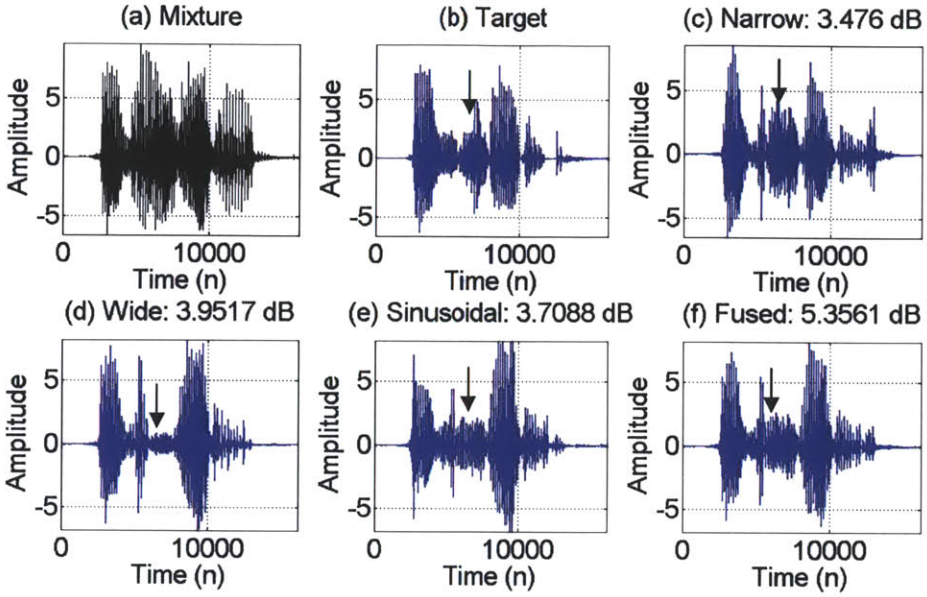

Figure 4-27. (a) Waveform of mixture of two males ("They were shattered" + "He merely said"); (b) target waveform "They were shattered."; (c) narrowband estimate; (d) wideband estimate; (e) sinusoidal-based estimate; (f) fused estimate of (c), (d), and (e); (c-f) list SNR gains. Observe near time 6000, the frame-based sinusoidal separation method exhibits periodicity from the interfering speaker that is suppressed in the narrowband and wideband estimates (arrows).

Table 4-3. Average RMSEs for speaker separation and standard errors [] on test set; "Fusion (N+W)" refers narrowband and wideband fusion; here, "Sinusoidal" refers to the frame-based sinusoidal separation method; "All Fusion" refers to narrowband, wideband, and sinusoidal fusion.

	Direct	Bootstrap	Fusion (N+W)	Sinusoidal	All Fusion
MM	3.38e-2 [1.4e-3]	3.28e-2 [1.40e-3]	2.82e-2 [1.40e-3]	2.95e-2 [1.10e-3]	2.59e-2 [1.10e-3]
FF	3.22e-2 [8.96e-4]	3.19e-2 [8.81e-4]	2.56e-2 [6.10e-4]	2.01e-2 [5.24e-4]	2.14e-2 [4.78e-4]
FM(M)	2.77e-2 [8.61e-4]	2.82e-2 [9.29e-4]	2.49e-2 [9.42e-4]	2.23e-2 [5.84e-4]	2.09e-2 [6.72e-4]
FM(F)	3.52e-2 [1.00e-3]	3.64e-2 [1.00e-3]	2.79e-2 [7.08e-4]	2.77e-2 [5.81e-4]	2.55e-2 [5.43e-4]

Table. 4-4 Average SNRs (dB) for speaker separation (dB), standard errors [] on test set; "Fusion (N+W)" denotes fusion of narrowband and wideband estimates; "All Fusion" denotes fusion of narrowband, wideband, and sinusoidal-based estimates.

	Direct	Bootstrap	Narrow	Fusion (N+W)	Sinusoidal	All Fusion
MM	4.42 [0.12]	4.86 [0.15]	3.80 [0.15]	5.21 [0.16]	4.73 [0.13]	5.97 [0.13]
FF	5.63 [0.18]	6.02 [0.19]	6.31 [0.14]	6.75 [0.16]	9.18 [0.18]	8.77 [0.15]
FM – Male	5.46 [0.13]	5.92 [0.14]	4.91 [0.11]	6.35 [0.10]	6.81 [0.13]	7.68 [0.10]
FM – Female	5.66 [0.14]	5.54 [0.15]	5.83 [0.11]	6.55 [0.14]	6.32 [0.10]	7.59 [0.14]

4.8 Conclusions

This work has proposed a model of speech-signal content as represented in 2-D analysis of wideband spectrograms. We have validated the utility of this model for representing speech

content in both analysis/synthesis and co-channel speaker separation experiments. In conjunction with our previous work, the model motivates a novel *taxonomy* of speech-signal behavior in the 2-D Grating Compression Transform (GCT) that exhibits important distinctions in interpretation, particularly in relation to "dual" behavior.

One implication of the proposed taxonomy is its potential for interpreting *other* time-frequency distributions. For instance, the auditory spectrogram of [42] is generally viewed as being "narrowband"/"wideband" in its low/high-frequency regions. The periodicity- and formant-dependent carrier derived in the current GCT framework may be applicable to high-frequency regions, thereby providing an explicit interpretation for modulation components observed in the auditory spectrogram in relation to speech parameters. As suggested by our results in speaker separation, the GCT may have additional applications due to its representation of speech parameters. For instance, modifying carrier components in the WGCT may be used for pitch and/or formant bandwidth modification in voice transformation. As suggested in Chapter 3, the mapping of noise and speech content in distinct regions of the GCT space also motivates applicability to speech enhancement. These new directions will be further discussed in Chapter 7. Finally, the present speaker separation framework may be combined with existing multi-pitch tracking methods towards a full separation system; efforts towards developing such a system are described in Chapter 6.

4.9 Appendix I

Consider a *time-varying* decaying sinusoid represented by Green's function $g[n,m]$, where m is the time of excitation, and n is the time axis along which we observe the resulting response [1], i.e.,

$$g[n,m] = \xi e^{-\int_m^n \dot{\alpha}(z)dz} \cos\left(\int_m^n \dot{\phi}(z)dz\right) u[n-m]. \tag{4.50}$$

$\dot{\alpha}(z)$ and $\dot{\phi}(z)$ are integrable functions corresponding to the *instantaneous* decay rate and center frequency of the formant, respectively, and ξ is the initial amplitude of the response. The output $y[n]$ of $g[n,m]$ excited by $p[n]$ (4.4) is a superposition sum [1]

$$y[n] = \sum_{m=-\infty}^{\infty} g[n,m]p[m]. \tag{4.51}$$

Substituting (4.4) and (4.50) into (4.51), we obtain

$$y[n] = \sum_{k=0}^{N_k} \xi e^{-\int_{kP}^n \dot{\alpha}(z)dz} \cos\left(\int_{kP}^n \dot{\phi}(z)dz\right) u[n-kP]. \tag{4.52}$$

Let n_0 denote the time at which the window is shifted to extract a segment $y_{n_0}[n]$ of $y[n]$, i.e.,

$$y_{n_0}[n] = w[n-n_0] \sum_{k=0}^{N_k} \xi e^{-\int_{kP}^n \dot{\alpha}(z)dz} \cos\left(\int_{kP}^n \dot{\phi}(z)dz\right) u[n-kP]. \tag{4.53}$$

Consider n_0 in (4.53) such that the entirety of the window is located between impulses at N_kP and $(N_k+1)P$. Within $y_{n_0}[n]$, we assume that *the decay rate and frequency of the sinusoid are constant and a function of the time of analysis* n_0

$$y_{n_0}[n] \approx w[n-n_0] \sum_{k=0}^{N_k} \xi e^{-\left(\dot{\alpha}(n_0)n + \int_{kP}^{n_0} \dot{\alpha}(z)dz\right)} \cos\left(\dot{\phi}(n_0)n + \int_{kP}^{n_0} \dot{\phi}(z)dz\right) u[n-kP]. \tag{4.54}$$

This "frozen time" approximation is similar to that assumed in typical short-time analysis methods (e.g., linear prediction [3]) to invoke stationarity of speech parameters. The contribution of the k^{th} component in (4.54), although time-varying, appears to come from a decaying sine with constant decay and frequency. Nonetheless, its starting amplitude (of the decay) and phase (of the sinusoid) will differ as a function of the distance between n_0 and point of excitation kP.

We make a further approximation by assuming that each contribution to the summation across k in (4.54) is *aligned at the window onset such that it may be viewed as a scaled and shifted decaying sinusoid*, thereby ignoring effects of the phase terms $\int_{kP}^{n} \phi(z)dz$ and temporal overlap, i.e.,

$$y_{n_0}[n] \approx w[n-n_0] \sum_{k=0}^{N_k} e^{-\int_{kP}^{n_0} \dot{\alpha}(z)dz} h[n-n_0; n_0]. \tag{4.55}$$

$$h[n; n_0] = \xi e^{-\dot{\alpha}(n_0)n} \cos\left(\dot{\phi}(n_0)n\right) u[n] \tag{4.56}$$

The Fourier transform of (4.55) and its magnitude are

$$Y(n_0, \omega) \approx \sum_{k=0}^{N_k} e^{-\int_{kP}^{n_0} \dot{\alpha}(z)dz} [W(\omega) *_{\omega} H(\omega, n_0)] e^{-j\omega n_0} \tag{4.57}$$

$$H(\omega, n_0) = \frac{0.5\xi}{\dot{\alpha}(n_0) + e^{j(\omega - \dot{\phi}(n_0))}} + \frac{0.5\xi}{\dot{\alpha}(n_0) + e^{j\left(\omega + \dot{\phi}(n_0)\right)}} \tag{4.58}$$

$$|Y(n_0, \omega)|_{local} \approx w[n_0, \omega] E_d[n_0] \widetilde{H}(n_0, \omega) \tag{4.59}$$

$$E_d[n_0] = \sum_{k=0}^{N_k} e^{-\int_{kP}^{n_0} \alpha(z)dz} \tag{4.60}$$

$$\widetilde{H}(n_0, \omega) = |W(\omega) *_{\omega} H(\omega, n_0)|, \tag{4.61}$$

where $\left|\widetilde{H}(n_0, \omega)\right|$ is a smoothed version of the formant and $E_d[n_0]$ is a time-dependent amplitude term. In (4.59), we have added a 2-D window term $w[n, \omega]$ to emphasize analysis in a local time-frequency region.

While $E_d[n_0]$ is not a periodic function in general, it can be made periodic in P under certain constraints such as $\dot{\alpha}(z) = \alpha_0$ or $\dot{\alpha}(z) = \cos\left(\frac{2\pi}{P} z\right)$ corresponding to constant or sinusoidally-varying decay rates. These conditions therefore allow for time-varying formants to be represented as a general time-dependent envelope term in conjunction with a periodic carrier. For instance, $\dot{\alpha}(z) = \alpha_0$ reflects a condition of constant decay but potentially changing formant frequency. For periodic $E_d[n_0]$, it can be shown that the 2-D Fourier transform of (4.59) (i.e., the WGCT) is

$$Y(\upsilon, \Omega) = W(\upsilon, \Omega) *_{\upsilon,\Omega} \left[\eta(\upsilon, \Omega) *_{\upsilon} \left(K\delta(\upsilon) + \sum_{l}^{N_l} 0.5\beta_l \delta\left(\upsilon \pm \frac{2\pi}{P}\right)\right)\right] \tag{4.62}$$

where $\eta(\upsilon, \Omega)$ is the 2-D Fourier transform of $\widetilde{H}(n_0, \omega)$ and K, β_l, and N_l are parameters of a sinusoidal series. Our discussion motivates a *modulation* view of the wideband spectrogram to include time-varying formant structure. Nonetheless, this view holds only approximately in time regions away from excitation impulse onsets due to the choice of the window position.

Chapter 5

Multi-Pitch Estimation

In this chapter[6], we describe approaches to multi-pitch estimation using the proposed two-dimensional (2-D) speech processing framework for the narrowband spectrogram representation. In Section 5.1, we briefly review the narrowband signal model for the (narrowband) Grating Compression Transform (GCT) and highlight some of its analytical properties. In Section 5.2, we demonstrate the ability of the GCT-based representation to analyze pitch and pitch-dynamic information of synthetic signals consisting of single and multiple periodic sources. In Section 5.3, we describe a multi-pitch algorithm for all-voiced speech using a simple variation of the GCT model; we evaluate this algorithm on a collected data set to assess the utility of the GCT in addressing an outstanding problem in existing multi-pitch estimation methods in handling pitch trajectory mixtures that have pitch values that are "close" to each other, thereby exhibiting *crossings* and/or *mergings*. We conclude the chapter in Section 5.4.

5.1 Signal Model for Pitch and Pitch-dynamic Information

Recall from Chapter 3 that a localized region of a narrowband spectrogram computed for voiced speech can be modeled using a sinusoidal series-based amplitude modulation formulation, i.e.,

$$s_w[n,\omega] = w[n,\omega]s[n+n_c,\omega+\omega_c] \tag{5.1}$$

$$s_w[n,\omega] \approx a_w[n,\omega][\mathrm{K} + \textstyle\sum_{\mathrm{k}=1}^{\infty} \alpha_{\mathrm{k}} \cos(\phi_k[n,\omega])] \tag{5.2}$$

$$\phi_k[n,\omega] = k\Omega_s(n\cos\theta + \omega\sin\theta) + \psi_k. \tag{5.3}$$

To analyze pitch, recall that the θ and Ω_s terms of the sinusoidal series carrier term $K + \sum_{k=1}^{\infty} \alpha_k \cos(\phi_k[n,\omega])$ can be shown to relate to pitch and pitch-dynamic information. From Chapter 3, recall also for the case of two concurrent speakers that we invoke a linearity assumption of the model such that

$$s_{mix}[n,\omega] \approx \textstyle\sum_{i=1}^{2} s_i[n,\omega] \tag{5.4}$$

$$s_i[n,\omega] = a_i[n,\omega]\left(K_i + \textstyle\sum_{k=1}^{N_i} \alpha_{i,k} \cos\phi_{i,k}[n,\omega]\right) \tag{5.5}$$

[6] Substantial portions of this chapter were obtained from two publications: [53] [20]

with resulting GCT mapping

$$S_{mix}(\upsilon,\Omega) \approx \sum_{i=1}^{2} A_i(\upsilon,\Omega) + \sum_{i=1}^{2}\sum_{k=1}^{N_i}\begin{bmatrix} e^{j\psi_{i,k}}A_{k,i}(\upsilon - k\Omega_s\cos\theta_k, \Omega + k\Omega_s\sin\theta_k) \\ +e^{-j\psi_{i,k}}A_{k,i}(\upsilon + k\Omega_s\cos\theta_k, \Omega - k\Omega_s\sin\theta_k)\end{bmatrix} \quad (5.6)$$

In multi-pitch analysis, existing methods typically rely on differences in either 1) pitch values and/or 2) energy content within a frequency region to extract pitch information for a distinct speaker. For instance, if two speakers exhibit different pitch values, they can be separated based on this distinction alone; in addition, if one speaker exhibits substantially greater energy within a frequency band, a reliable estimate of the dominant speaker can also be obtained while ignoring the weaker speaker (i.e., as in binary masking [43]). The present GCT mapping can similarly exploit such conditions. Nonetheless, the GCT's explicit representation of *pitch dynamics* can invoke additional separability that builds on that provided by conditions 1) and 2).

5.2 GCT Analysis of Synthetic Pitch Signals

Previous work in [7] has shown that the GCT can accurately estimate pitch using a single low-frequency region of the spectrogram when applied to single speakers. In the subsequent sections, we explore pitch and pitch and multi-pitch analysis with the GCT using *multiple* frequency regions to assess its properties and limitations. First, in Section 5.2.1, we demonstrate empirically the limits of analysis fidelity in extracting pitch and pitch-dynamic information across multiple frequency regions for a *single* source. Next in Section 5.2.2 and 5.2.3, we demonstrate the GCT's ability to provide separation of pitch information for *multi*-pitch signals with *identical* spectral shaping. Finally, in Section 5.2.4, we demonstrate the GCT's ability to analyze more general multi-pitch signals with different spectral shaping. Our discussion in this section focuses on pitch *analysis,* which we delineate from pitch *tracking*. Specifically, we refer to pitch analysis as extracting pitch values of signals without making assignments of those values to distinct speakers (in multi-pitch signals). We refer to pitch tracking (to be subsequently performed in Section 5.3) as both pitch analysis and making assignments of pitch values to distinct speakers for multi-pitch signals

5.2.1 Multi-Region Analysis of a Synthetic Vowel

Synthetic Data: A synthetic impulse training with rising pitch of 150 Hz to 250 Hz across a 500 ms duration was generated using linearly spaced impulses with a glottal flow component [19]. The signal was generated at an oversampled 80 kHz but downsampled to 8 kHz for processing. This synthesis method results in a spectrogram with distinct harmonic structure with rising pitch, consistent with that typically observed in real speech [19], and avoids step-like behavior in the harmonic structure as observed in [16]. An all-pole filter with stationary formant frequencies of 500, 1500, and 2500 Hz and bandwidths of 80 Hz was used to filter the source signal to generate the vowel.

Analysis: The waveform is pre-processed using a pre-emphasis filter $H(z) = 1 - 0.97z^{-1}$. The short-time Fourier transform (STFT) was computed using a 25-ms Hamming window, 1-ms frame rate, and a 512-point discrete Fourier transform (DFT). The magnitude of the STFT (STFTM) was taken followed by the log operation to generate the log-STFTM. A 2-D gradient operator was applied to the log-STFTM as a preprocessing step to remove near-DC terms in the Grating Compression Transform [7]. Region sizes of 100 ms by 700 Hz were extracted from the

middle of vowel in time across multiple frequency regions; step sizes along frequency were set to $1/16^{th}$ the size of the region corresponding to ~22 Hz. Finally, the GCT was computed using a 2-D DFT of size 512 by 512. The position of the maximum value of the GCT magnitude was extracted and used in the pitch and pitch-dynamic mappings (3.24) and (3.26) to obtain estimates (see Chapter 3).

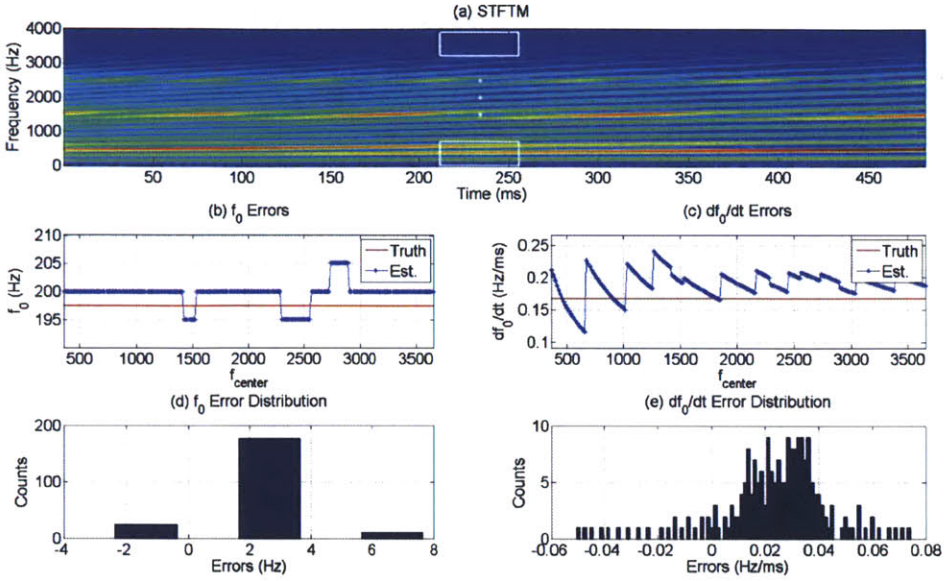

Figure 5-1. (a) STFTM of synthetic vowel and regions across frequency (solid white); (b) Pitch estimates (blue '*-') and true pitch (solid, red); (c) as in (b) for pitch derivative; (d) histogram of errors from (b); (e) histogram of errors from (e).

In this example, we observe that absolute pitch estimate errors range from 0 to 6 Hz. Errors can be attributed to two factors. First, since the local formant structure is not strictly bandlimited, the $A_l(\upsilon, \Omega)$ term can interact with the carrier components when mapped to the GCT domain; this can be particularly deleterious to estimates if the peak value of $A_l(\upsilon, \Omega)$ is not located at the GCT origin. This is consistent with the observation that pitch errors peak formant frequencies of 500, 1500, and 2500 Hz (Figure 5-1b). In addition, as observed in [19] and Chapter 3, changing pitch invokes fanned harmonic line effects in the overall spectrogram in contrast to the approximation of parallel lines within a local time-frequency region. Pitch derivative estimates exhibit an oscillatory behavior and are generally within ~0.05 Hz/ms (relative to the underlying rate of 0.2 Hz/ms). These errors are presumably due to a combination of errors in the pitch itself (as this is used in computing the pitch derivative) as well as fanned harmonic line effects. The present results quantitatively demonstrate that the GCT framework can analyze pitch and pitch-dynamic information by performing simple peak-picking in the GCT domain (while ignoring the GCT's near-DC terms) albeit with limitations. As will subsequently be demonstrated, however, these errors in analysis can be at least be partially overcome by *combining* pitch (and/or pitch-derivative) estimates across frequency regions for a single point in time.

5.2.2 GCT-based Separability of Pitch Information for Concurrent Vowels

This section investigates properties of multi-pitch analysis using the GCT. Traditional approaches to multi-pitch analysis obtain pitch candidates from autocorrelation estimates of band-pass filtered versions of the waveform on a *frame-by-frame* basis (e.g., [33]). This approach

provides distinct pitch candidates for a single point in time but does not represent the pitch dynamics of multi-pitch signals. Here, we show by example that the GCT's representation of pitch dynamics within a local time segment invokes *separability* of pitch information distinct from that obtained in short-time autocorrelation analysis.

Consider for instance a condition in both speakers have the *same* pitch value *and* similar energy within a frequency region. Figure 5-2 illustrates two scenarios of pitch content under this condition. Observe that while separability in the GCT can occur based on pitch alone, it can *also* occur if pitch values are the same but exhibit distinct dynamic information. The GCT representation can therefore provide extraction of pitch content and separability in three ways: 1) energy dominance, 2) pitch values, and 3) pitch dynamic information.

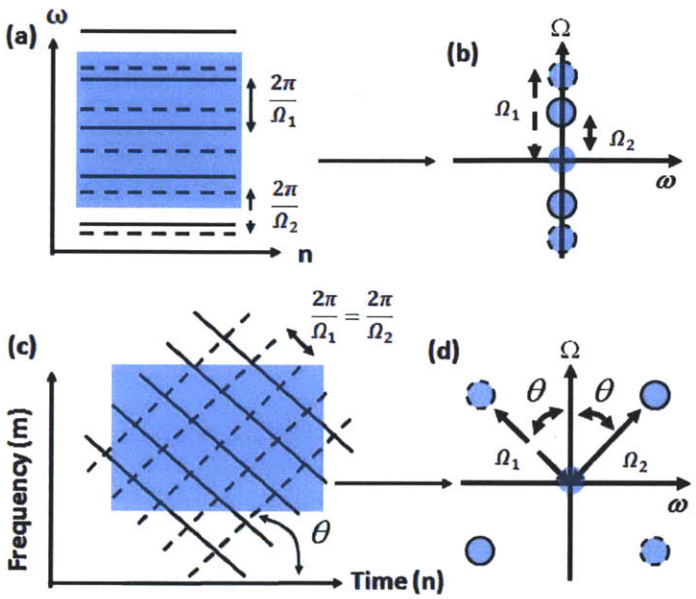

Figure 5-2. (a) Schematized localized region with local formant structure (shaded) equivalent in mixture of two speakers (solid, dashed lines) and distinct pitch values; (b) GCT of (a) indicating separability of pitch information from pitch value; additional carrier terms omitted for simplicity; (c) As in (a) but with *equal* pitch values but trajectories moving in opposite directions; (d) GCT of (c) showing separability of pitch information based on pitch dynamic information.

Synthetic Data: Two vowels with distinct formant structures (Figure 5-3) were excited using source signals with linearly rising and falling pitch values of 150 to 200 Hz; source signals were generated as in Section 5.2.1. Along the frequency axis (Figure 5-3), the vowel structures were chosen to be distinct below ~1500 Hz but equal above ~1500 Hz) in terms of their magnitudes. This is done to investigate properties of analysis techniques under differing/equal energy conditions within distinct frequency regions. The vowels were sampled at 8 kHz.

Analysis: In GCT analysis, we again apply the 2-D gradient operator to the entire spectrogram to remove effects of the near-DC components in the GCT. Localized regions centered at both 2050 Hz (denoted as R2), where the formant structure of the two vowels is nearly identical, and at 1200 Hz (denoted as R1), where they differ, were extracted using 2-D Hamming windows of size 50 ms by 700 Hz. The GCT was computed using a 512 by 512-point 2-D discrete Fourier transform.

As a representative method of existing short-time analysis techniques, two linear-phase band-pass filters centered at 1200 Hz and 2050 Hz were applied to the waveform. To obtain an envelope, filtered waveforms were half-wave rectified and low-pass filtered (cutoff = 800 Hz). The normalized autocorrelation (denoted as $r_{xx}[n]$) was computed for a 30-ms duration of the envelopes.

Figure 5-4 shows results of these analyses for R1 and R2. Observe in short-time analysis for R1 that a single distinct pitch estimate and its sub-harmonics are present; however, R2 reflects the interaction of closely-spaced periodicities and appears "noisy". These observations are similar to those observed in which these "noisy" bands were discarded in favor of those exhibiting a dominant pitch to compute a summary correlogram at a single point in time as described in [33]. Figure 5-4c-d show GCTs computed over localized time-frequency regions at R1 and R2. A single dominant set of impulses, corresponding to a single pitch value, is present in the GCT for R1, similar to $r_{xx}[n]$ for R1. This is to be expected given that the formant envelope magnitudes are different, thereby allowing for the log-max approximation. In contrast, observe that two distinct sets of peaks can also be seen for R2 corresponding to two similar pitch values *and* energy values of the formant envelope. The GCT can therefore separate pitch information of two speakers with similar energies and pitch values in a localized set of frequency bands by explicitly representing the temporal dynamics of distinct speakers in contrast to traditional short-time methods. This separability generalizes to the case where envelope structures exhibit similar energies as well as different pitch values/temporal dynamics as shown in Figure 5-2.

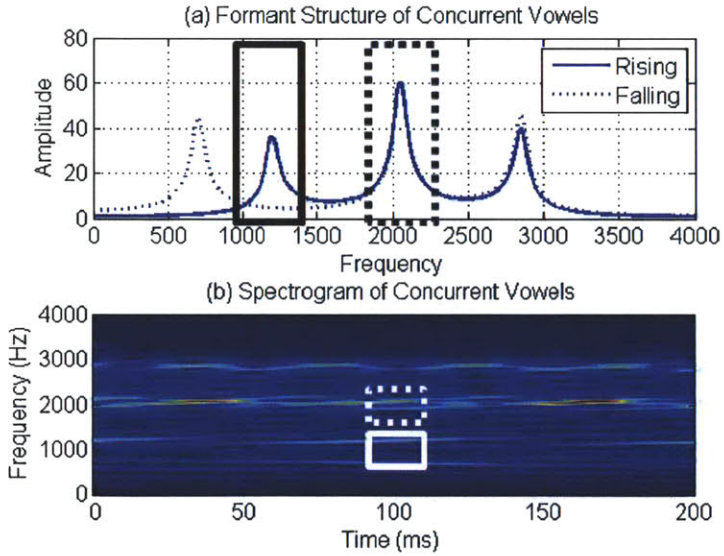

Figure 5-3. (a) Spectrum showing formant structures of rising and falling vowels. Regions 1 (solid) and 2 (dotted) for analysis using short-time autocorrelation analysis; (b) spectrogram of concurrent vowels (plotted on linear amplitude scale) with corresponding Region 1 (dotted) and Region 2 (solid) shown for GCT analysis.

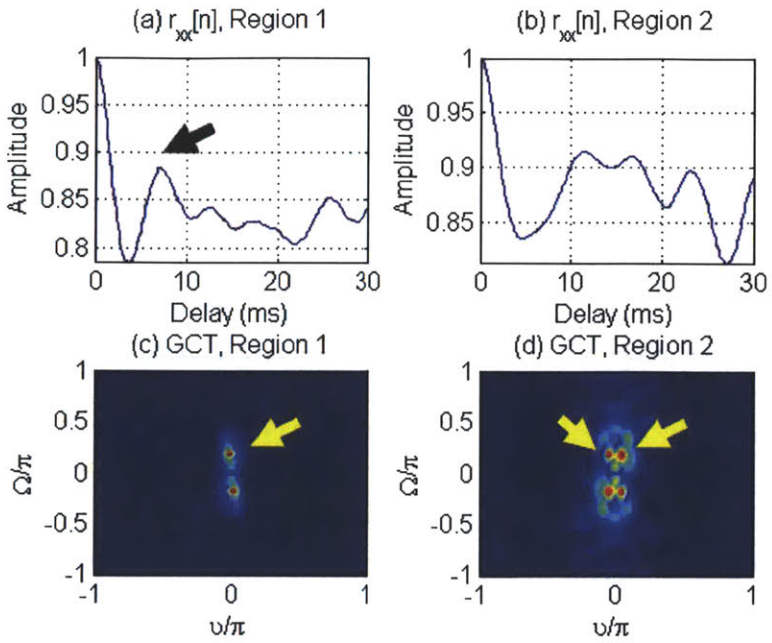

Figure 5-4. (a) Region 1 analysis showing distinct peak (arrow) corresponding to pitch value; (b) as in (a) but for Region 2; observe that no distinguishable peak at the pitch value is present; (b) GCT analysis of Region 1 showing one set of dominant peaks corresponding to dominant local formant structure; (d) as in (c) but for Region 2; observe that two distinct pairs of peaks corresponding to both vowels can be obtained (yellow arrows).

5.2.3 Multi-Pitch Analysis of Concurrent Vowels – Identical Formants

Building on results from the previous section, we demonstrate the utility of the described GCT framework in multi-pitch analysis for a condition of synthetic vowels with *identical* formant structure.

Synthetic Data: We use for the analysis signal a concurrent vowel from Section 5.2.2 excited with five distinct pairs of pitch trajectories to account for a variety of conditions:

1. Stationary pitch of 100 Hz + Stationary pitch of 300 Hz
2. Rising pitch from 100- to 250-Hz + Stationary pitch of 175 Hz
3. Rising pitch from 125- to 150-Hz + Falling pitch from 250- to 200-Hz
4. Rising pitch from 100- to 150-Hz + Rising pitch from 175- to 225-Hz
5. Rising pitch from 150- to 200-Hz + Falling pitch from 200-Hz to 150-Hz

We denote these conditions subsequently as "Condition1", "Condition2", etc. Pitch trajectories were synthesized using pure impulse trains and with sampling rate of 8 kHz and used to excite the vowel structure of the *rising vowel* in Figure 5-3. All signals had duration of 0.5 seconds. Observe that conditions 2 and 5 result in pitch trajectories that exhibit crossings at the center of the signal. The conditions synthesized reflect permutations of decreasing/increasing tracks in conjunction with same/different rates of pitch change.

Analysis Methods: A narrowband spectrogram spectrogram was computed using a 25-ms Hamming window, 1-ms frame rate, and 512-point discrete-Fourier transform (DFT). Local region sizes were extracted for GCT analysis of size 700 Hz by 150 ms; in discrete samples, this corresponds to in 45 frequency and 150 in time. In our preliminary analyses, we observed that the choice of region size could be varied in 500 Hz ~ 1000 Hz and 100 ~ 200 ms with similar results. The mean value of each local region is removed prior to windowing with a 2-D Hamming window; a 2-D DFT of size 512 by 512 is then used to compute the GCT and avoids aliasing since the DFT length is larger than the region size. To extract pitch candidates, a 2-D multi-peak picker is applied to the resulting GCT magnitude. Details of this peak-picker are described in Chapter 3. For each local region analyzed, the two peaks with largest magnitudes are kept. Note that we extract *two* peaks in this case *due to the formant structure of the concurrent vowel being the same* for both signals in the mixture. The pitch mapping of (3.24) is then used to obtain an estimate of the pitch.

The resulting pitch candidates are used in three distinct ways to assign pitch values to points in time: 1) single-region, 2) clustering, and 3) an oracle assignment (Figure 5-5). In the single-region method, pitch candidates obtained using only a single low-frequency region are used and assigned to each point in time. This method was used in [7] in previous efforts in pitch analysis of a single speaker and serves as a reference method for comparison. In clustering, we use a simple median-based clustering method to obtain the two pitch estimates. Specifically, for each point in time, the collection of candidates obtained across frequency regions are used to compute a median value. Subsequently, we compute the median of the set of candidate with higher and lower pitch values with respect to the median and assign the results as the pitch estimates. Finally, the *oracle* method utilizes all pitch candidates across frequency regions for a specific point in time; subsequently, the two pitch values closest in absolute frequency value to the *true* pitch values are used as the estimates. The final method assesses the ability of the multi-region based GCT analysis for extracting accurate pitch estimates, independent of how the pitch candidates are assigned to distinct time points.

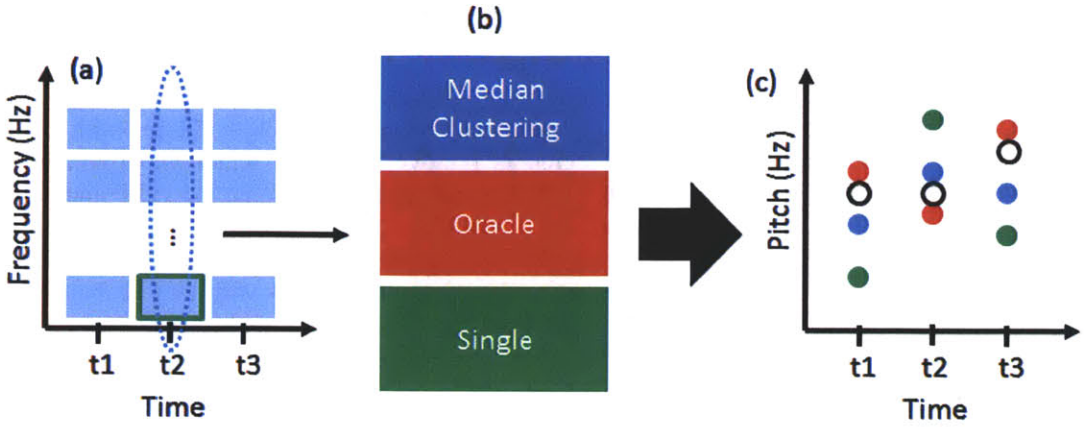

Figure 5-5. Schematic illustrating multi-pitch analysis method; (a) local regions (shaded blue) consisting of pitch candidates; extraction of all pitch candidates at time point t2 across all frequencies (dashed blue); (b) distinct analysis methods; (c) resulting pitch values assigned for each point in time with true pitch value (black hollow), median clustering (blue), single-region (green), and oracle (red) denoted.

As an additional reference, we evaluate the effect of assuming "dominance" of a single target signal in local time-frequency regions, a property typically used in multi-pitch analysis approaches as previously noted (e.g., Figure 5-4). Note that because the mixed signal analyzed in our experiments is a concurrent vowel with *identical* formant structure for both underlying pitch tracks, the frequency sparsity assumption for distinct signals in the mixture is violated. We apply the oracle and clustering approaches to a *subset* of the pitch candidates. Specifically, we remove from the full set of candidates those that corresponds to the 2nd-largest peak in the GCT domain for each local region analyzed. By doing so, we effectively extract pitch information *using only the dominant peak* in the GCT of each local time-frequency region. We refer to this set of candidates as the "dominance" set.

We use two methods to characterize the performance of our analyses methods. First, we use the collection of pitch candidates at each point in time to compute a *histogram* per time point of all possible pitch values (ranging from 80 to 350 Hz); note that this is done using the full set of candidates as well as the *subset* of candidates used in assessing the effects of the "dominance" assumption. These histograms qualitatively assess the distribution of resulting pitch candidates obtained in analysis. Second, as a quantitative metric for performance, we compute root-mean-squared-errors (RMSE) defined as

$$RMSE = \sqrt{\frac{1}{2L}\sum_{i=1}^{2}\sum_{t=1}^{L}\left(\hat{x}_{t,i} - x_{t,i}\right)^2} \tag{5.7}$$

where L is the length of the mixture, $x_{t,i}$ and $\hat{x}_{t,i}$ are the reference and estimated pitch values, respectively; here, the estimated pitch values correspond to the "single", "clustering", and "oracle" methods. Since we do not perform explicit assignment of the pitch values to distinct signals in the mixture, we use the "best" assignment at each time point of the resulting estimates relative to the true pitch values, where "best" corresponds to the minimal absolute difference in frequency.

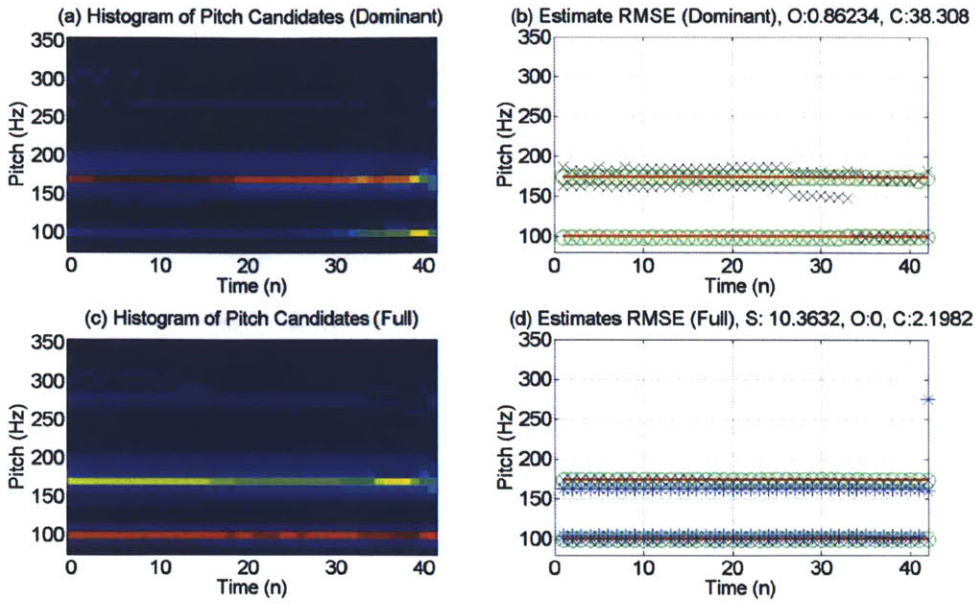

Figure 5-6. (a) Histogram of pitch estimates for Condition1 obtained from GCT analysis assuming a single *dominant* signal in a local time-frequency region; (b) Multi-pitch analysis results from the candidates in (a) with true (solid, red) pitch tracks, clustering ('x') and oracle ('o') denoted. RMSE values (Hz) denoted for oracle (O), and clustering (C); (c) as in (a) but for the *full* set of candidates (i.e., two peaks per region); (d) as in (b) but now also including single-region analysis as in [7].

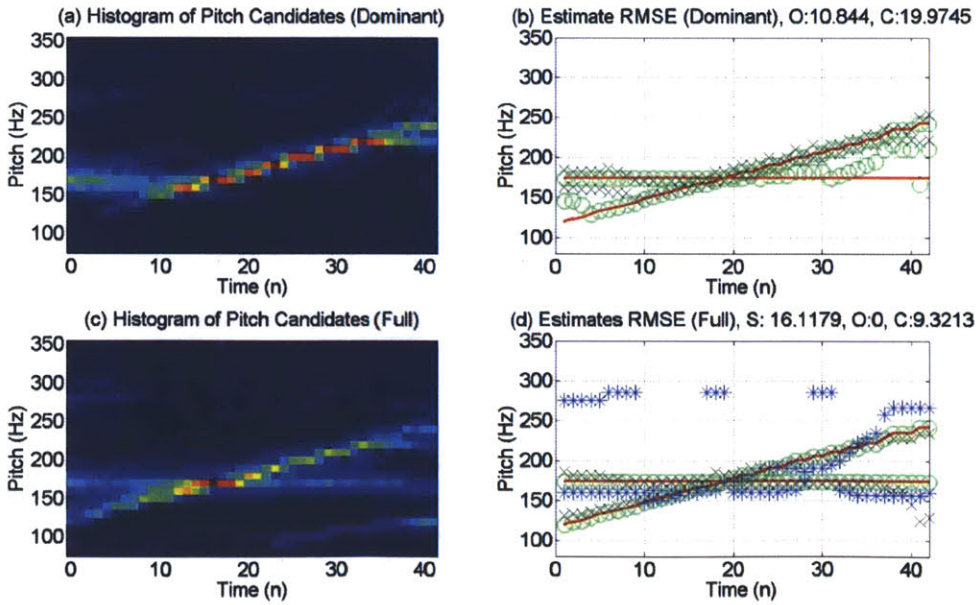

Figure 5-7. As in Figure 5-6 but for Condition2.

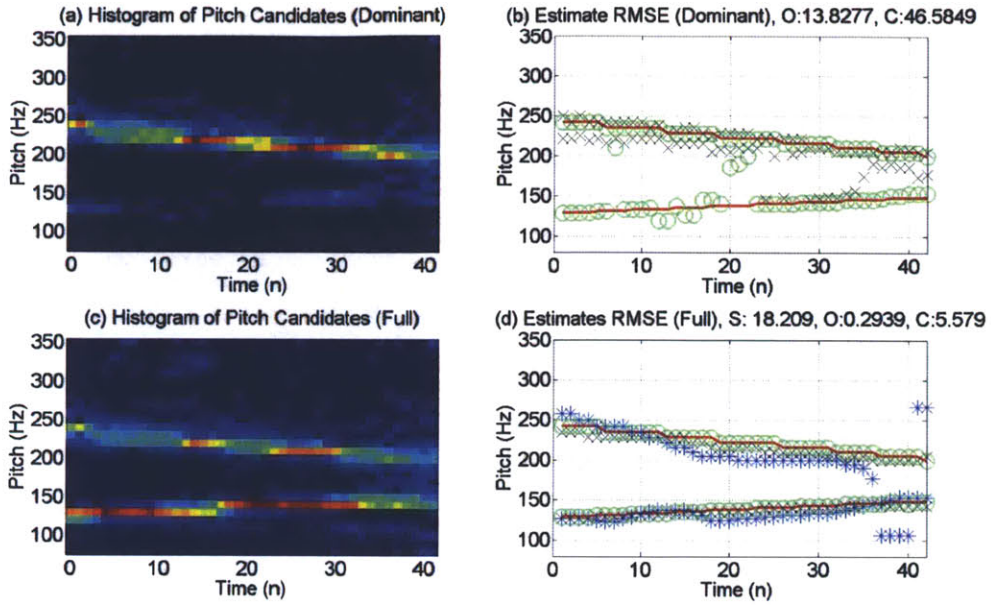

Figure 5-8. As in Figure 5-6 but for Condition3.

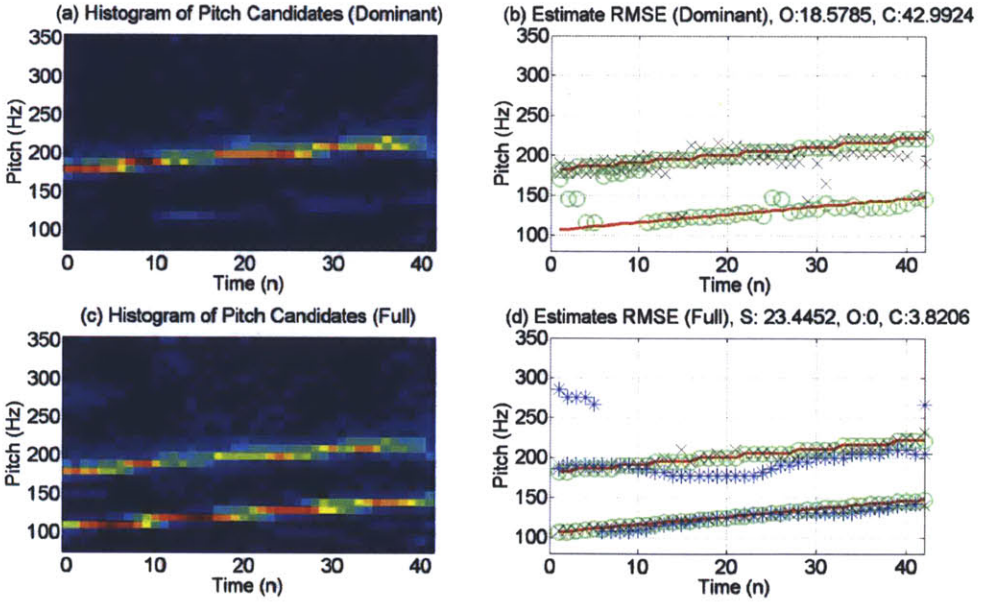

Figure 5-9. As in Figure 5-6 but for Condition4.

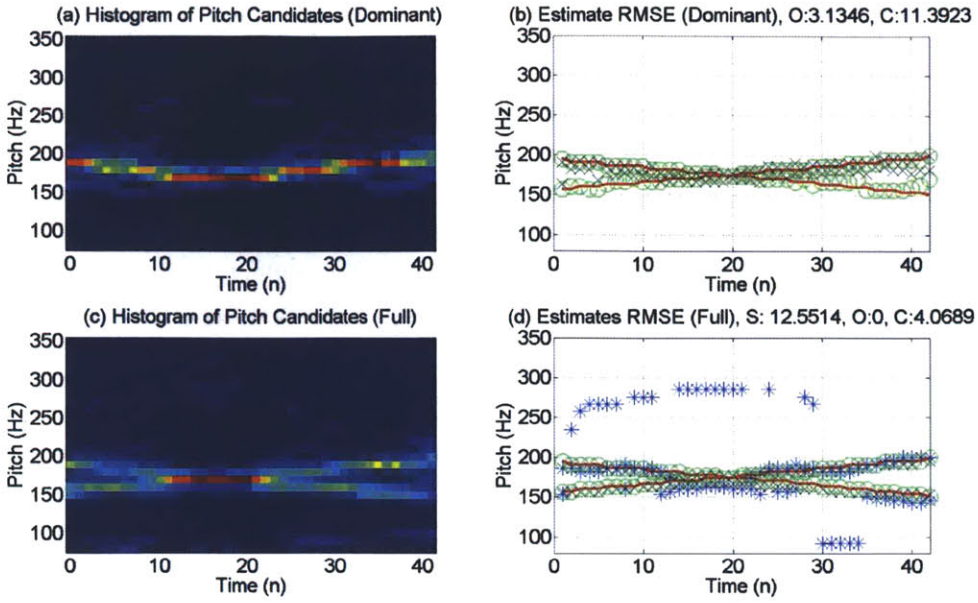

Figure 5-10. As in Figure 5-6 but for Condition5.

Results: Figure 5-6 through Figure 5-10 show the results of our analyses for each pitch trajectory pair. Comparing histograms between the full pitch candidate set and those of the "dominance" subset, observe that the full set exhibits high counts at pitch values close to *both* underlying pitch trajectories across time. In contrast, candidates in the "dominance" histograms appear to "jump" between pitch tracks across time (Figure 5-7a). For instance, observe in Figure 5-7 that for time 0~10, candidates primarily correspond to the stationary pitch track while from time 10~40, they correspond to the rising pitch track. This effect is not observed using the full set, in which candidates distinctly highlight both underlying pitch tracks. One disadvantage of the full set, however, appears to be potential pitch "halving" (e.g., Figure 5-7c, time 30~40) trajectory, indicating that the amplitude of the peaks corresponding to the stationary pitch are dominated by the 2nd harmonics of the moving pitch. In addition, high pitch values, unrelated to the underlying true pitch values (or their harmonics in the GCT space), can also be are observed in the full set histograms. These peaks are presumably due to the near-DC terms in the GCT corresponding to the envelope in the signal model.

Table 5-1. RMSE (Hz) across pitch trajectory conditions and methods for mixtures with identical formant structure; oracle (O), clustering (C), and single (S), for dominant-peak only and full set of peaks.

RMSE (Hz)/Condition	Condition1	Condition2	Condition3	Condition4	Condition5
O/C (Dominant)	0.86/38.31	10.84/19.97	13.83/46.58	18.58/42.99	3.13/11.39
O/C (Full)	0/2.20	0/9.32	0.29/5.58	0/3.82	0/4.07
S (Full)	10.36	16.12	18.21	23.45	12.55

In the pitch analysis results, observe that, consistent with the discrepancy between the "dominance" set and the full set, the *oracle* assignment is consistently worse in the former when compared to the latter. Most notably, at certain time points (e.g., ~35 in Figure 5-7b), two candidates with similar values corresponding to one underlying pitch track are obtained, thereby implying the absence of the other concurrent signal. This reflects the fact that pitch candidates

from the 2nd speaker are completely absent in the "dominant-peak only" subset. In contrast, when using the full set, the oracle assignments result in a nearly-exact estimate in all conditions with a maximum RMSE of 0.26 Hz in Figure 5-8. Quantitatively, Table 5-1 summarizes the RMSE values across pitch conditions and pitch analysis methods. When using the full set of pitch candidates, clustering methods result in lower RMSE values overall relative to use of a analysis from a single region, thereby demonstrating the ability of using *multiple* regions across frequency to obtain accurate multi-pitch estimates. In addition, the full candidate set generally results in lower RMSE in the oracle and clustering methods relative to using the dominant-peak only subset of candidates. This is consistent with the fact that separation of pitch information due to "dominance" cannot occur since the formant structures of the two vowels are identical.

5.2.4 Multi-Pitch Analysis of Concurrent Vowels – Distinct Formants

The previous section applied the GCT for multi-pitch analysis of mixtures of vowels with *identical* formant structure, thereby highlighting the GCT's unique ability in this condition to provide separability of pitch information despite the lack of frequency-region based "dominance" typically used in short-time/frame-based analysis methods. In this section, we explore the effects of mixtures of vowels having *distinct* formant structure. Specifically, we synthesize a set of vowel mixtures with pitch tracks identical to those in Section 5.2.3 but formant structures shown in Figure 5-11. The formant structures are designed to account for conditions in which spectral shaping is equal (i.e., at Hz), partially overlapping (at Hz), and generally non-overlapping (at Hz). All GCT-based analysis steps are identical to those used in Section 5.2.3.

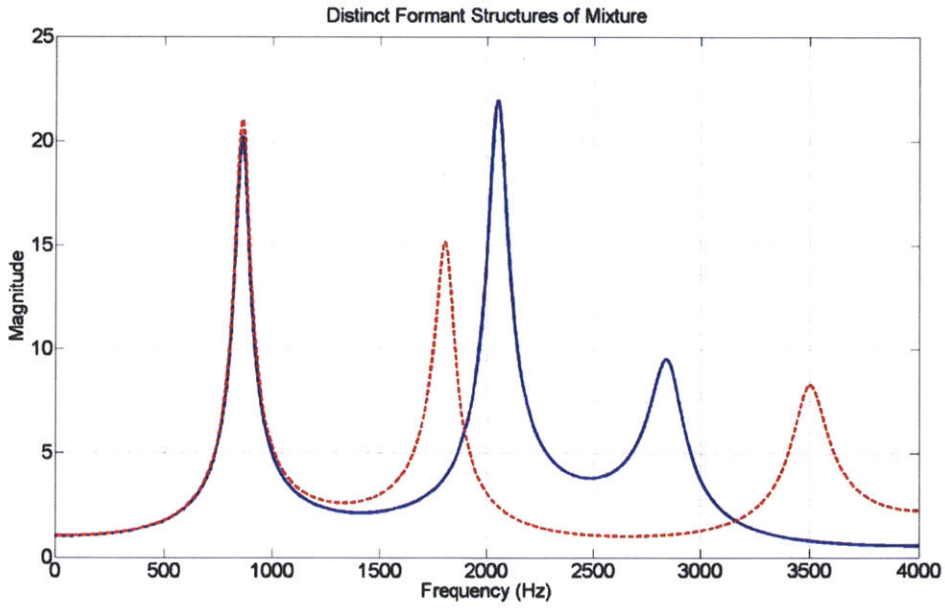

Figure 5-11. Frequency response magnitudes of two distinct formant structures (red dotted and blue solid) used in synthesizing vowel mixtures; observe that the spectral shaping is identical ~800 Hz corresponding to the first formant, partially overlapping near ~1900 Hz corresponding to the second formant, and virtually "separate" at the third formant at ~2700 and ~3500.

Figure 5-12 through Figure 5-16 show results of this analysis while Table 4-3 summarizes our results. Observe that in contrast to the identical-formants condition, use of the full set of pitch candidates results in substantial degradations in pitch analysis accuracy in contrast to use of only the subset of dominant-peak only pitch candidates. This reflects the fact that the primary peak in the GCT in regions of *differing* spectral shaping (e.g., at the second and third formants in Figure 5-11) will correspond to the dominating speaker while a *secondary* peak in the GCT does *not* correspond to the pitch value of the weaker speaker. Qualitatively, we see that effects of this behavior in comparing histograms of the pitch values (e.g., compare Figure 5-12a versus Figure 5-12c). Inclusion of the secondary peak results in a multitude of addition pitch values that are substantially far away from the true pitch values of either speaker. While these "spurious" peaks are also present when using only a dominant peak, this effect is notably greater when using the full set of peaks. Quantitatively, RMSE values are consistent with these observations in showing that use of the dominant-peak only set of candidates results in significantly lower RMSE values than when incorporating the *full* set of candidates; this occurs for both the oracle and clustering methods. Clustering nonetheless results in lower RMSEs than use of a single region, further motivating the use of multiple regions in pitch analysis via the GCT.

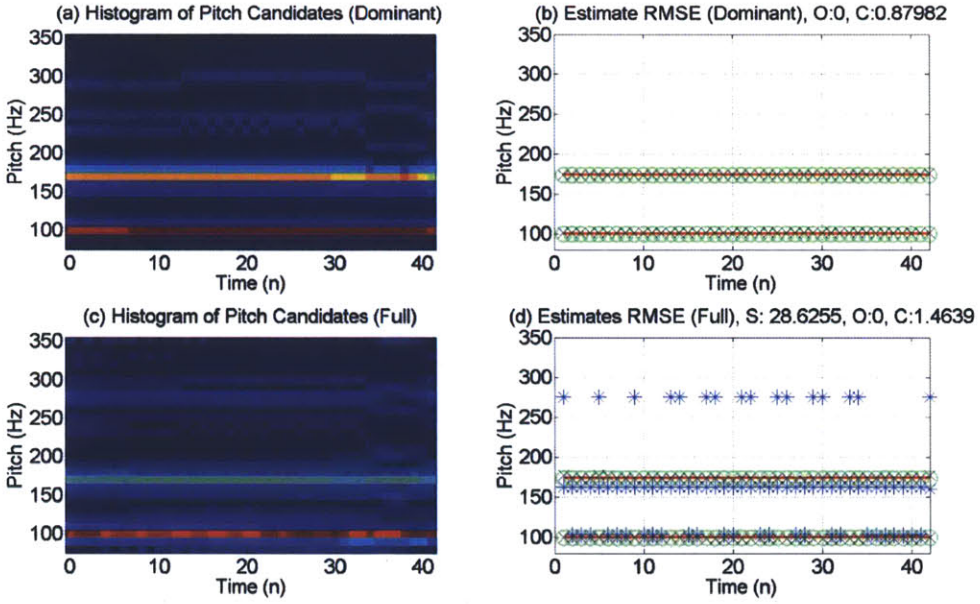

Figure 5-12. As in Figure 5-6 for Condition1 but for distinct formant structure. Legend for (b) and (d): true pitch values (red), oracle (green 'o'), clustering (black 'x'), single (blue '*').

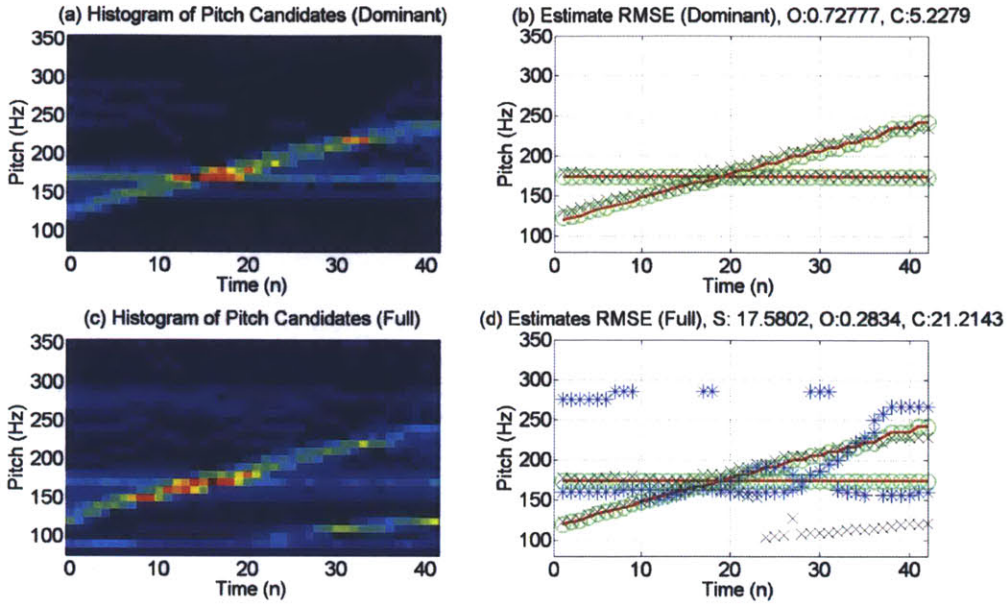

Figure 5-13. As in Figure 5-6 for Condition2 but for distinct formant structure.

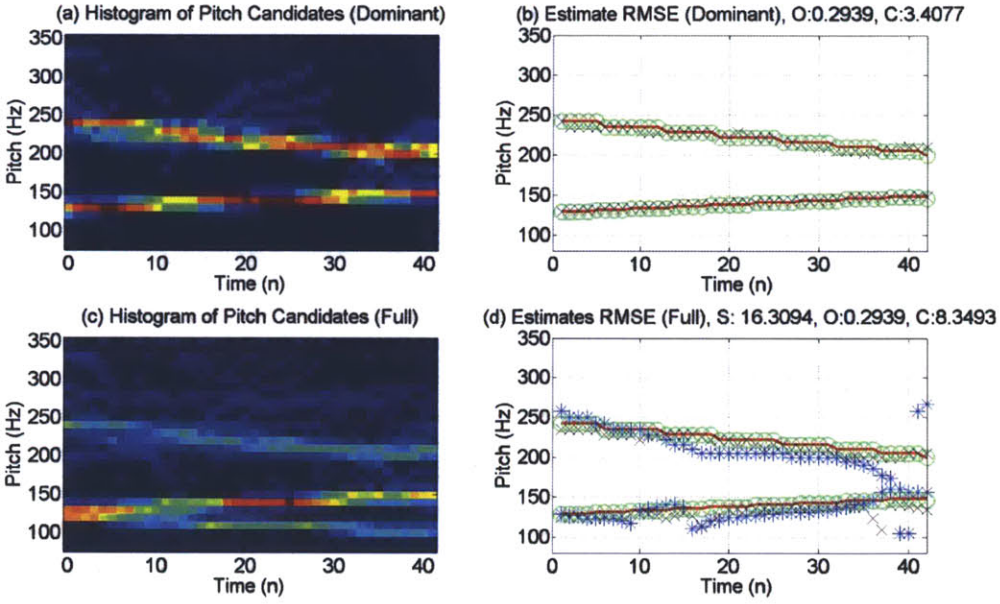

Figure 5-14. As in Figure 5-6 for Condition3 but for distinct formant structure.

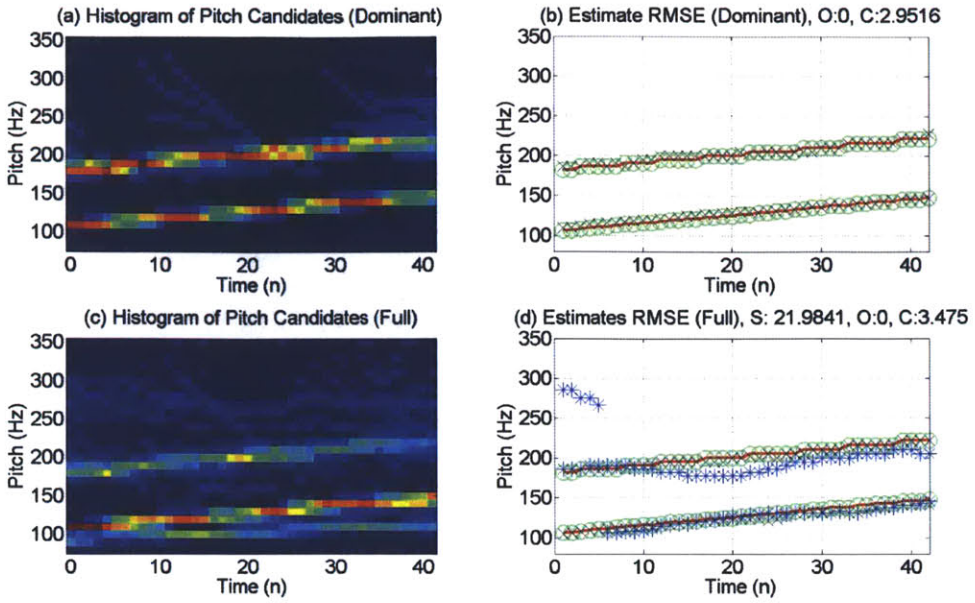

Figure 5-15. As in Figure 5-6 for Condition4 but for distinct formant structure.

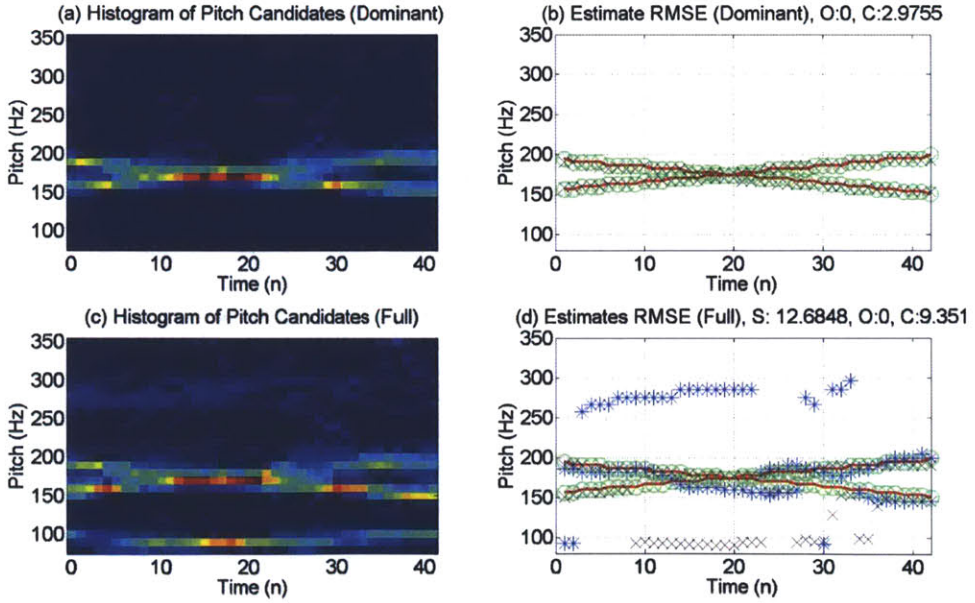

Figure 5-16. As in Figure 5-6 for Condition5 but for distinct formant structure.

Table 5-2. RMSE (Hz) across pitch trajectory conditions and methods for mixtures with identical formant structure; oracle (O), clustering (C), and single (S), for dominant-peak only and full set of peaks.

RMSE (Hz)/Condition	Condition1	Condition2	Condition3	Condition4	Condition5
O/C (Dominant)	0/0.88	0.73/5.23	0.29/3.41	0/2.95	0/2.98
O/C (Full)	0/1.46	0.28/21.21	0.29/8.35	0/3.48	0/9.35
S (Full)	28.63	17.58	16.31	21.98	12.68

The results of Sections 5.2.2 through 5.2.4 have demonstrated the ability of the GCT to analyze multi-pitch signals under conditions of identical and distinct formant structures. We have shown that under the identical-formants condition, the GCT's explicit representation of pitch dynamics can invoke a unique form of separability of pitch information that cannot be obtained using traditional short-time/frame-based analysis methods. In this condition, extracting *two* pitch candidates in each GCT computed for a local time-frequency region resulted in better pitch analysis results than use of a single candidate. This is due to the fact that the dominance of individual speakers in distinct frequency regions cannot be used to obtain pitch information from distinct speakers (due to identical formant structure). Nonetheless, in the more general mixture condition in which the formant structures of the mixture are *distinct*, the extraction of two pitch candidates results in relatively *poorer* pitch analysis results than use of a single dominant peak. This effect is due to the dominance of individual speakers' formant structures within localized regions. Nonetheless, use of a single peak can still obtain accurate pitch analysis results in a variety of pitch trajectory mixtures.

5.3 Multi-Pitch Estimation/Tracking of All-voiced Speech

5.3.1 Framework

Our previous discussion has highlighted properties and limitations of the GCT-based pitch and multi-pitch analysis framework. We explored properties of GCT analysis of multi-pitch signals under conditions of distinct and identical formant structure. In general, the occurrence of *identical* formant structure within multi-pitch signal mixtures occurs less frequently than distinct formant structure. This effect been demonstrated empirically through simulations of real speech mixtures in [26], thereby motivating the exploitation of "dominance" of individual speakers across frequency bands in speaker separation and multi-pitch analysis methods (e.g., [33]). Simulations in that work were performed on the *log* spectrograms computed for mixtures of speech. We analyze now the effect of the log operation on the proposed 2-D signal model of speech. Specifically, taking the log of (3.22), we obtain

$$\log s_w[n,\omega] = \log a_w[n,\omega] + \log(\mathrm{D} + \textstyle\sum_{\mathrm{k}=1}^{\infty} \alpha_{\mathrm{k}} \cos(\phi_k[n,\omega])). \qquad (5.8)$$

for $a_w[n,\omega]$ and D constrained such that the log arguments are positive. Observe that as a result of this nonlinear transformation, the carrier terms containing pitch information are separated in summation from the envelope term. In addition, since the log operation maintains the periodicity of its argument, $\log[\mathrm{K} + \sum_{\mathrm{k}=1}^{\infty} \alpha_{\mathrm{k}} \cos(\phi_k[n,\omega])]$ is again a periodic signal (in n and ω) in Ω_s and θ such that pitch and pitch-dynamic information can be obtained from this term (see Chapter 3). The corresponding Grating Compression Transform (GCT) representation is then

$$S_{log}(v,\Omega) = A_l(v,\Omega) + \textstyle\sum_{\mathrm{k}=1}^{\infty} \delta'_k(v,\Omega)^+ + \delta'_k(v,\Omega)^- \qquad (5.9)$$

$$\delta'_k(v,\Omega)^{\pm} = 0.5\alpha_k e^{\pm j\psi_{i,k}} \delta(v \mp k\Omega_s \cos\theta_k, \Omega \pm k\Omega_s \sin\theta_k) \qquad (5.10)$$

where we have used the fact that $\log(D + \sum_{k=1}^{\infty} \alpha_k \cos(\phi_k[n,\omega]))$ may again be expanded into a sinusoidal series with fundamental 2-D spatial frequency Ω_s as in Chapter 3. From Chapter 3, recall also for the case of two concurrent speakers that we invoke a linearity assumption of the model such that

$$s_{mix}[n,\omega] \approx \sum_{i=1}^{2}\left[a_i[n,\omega]\left(D_i + \sum_{k=1}^{N_i} \alpha_{i,k} \cos \phi_{i,k}[n,\omega]\right)\right]. \tag{5.11}$$

The log of this mixture is then

$$\log s_{mix}[n,\omega] = \log \sum_{i=1}^{2}\left[a_i[n,\omega]\left(D_i + \sum_{k=1}^{N_i} \alpha_{i,k} \cos \phi_{i,k}[n,\omega]\right)\right] \tag{5.12}$$

We view this representation in the context of the *log-max* representation that has been demonstrated through empirical observations to approximately match the true log magnitude [26] of spectrograms computed on mixtures of speakers. Specifically, $\log s_{mix}[n,\omega]$ becomes (in the case of speaker 1 being "dominant")

$$\log s_{mix}[n,\omega] \approx \max(\log s_1[n,\omega], \log s_1[n,\omega]) \tag{5.13}$$

$$\log s_{mix}[n,\omega] \approx \log a_1[n,\omega] + \log\left(D_1 + \sum_{k=1}^{N_1} \alpha_{1,k} \cos \phi_{1,k}[n,\omega]\right). \tag{5.14}$$

where the max operation is extended for the GCT analysis to operate across the local time-frequency region for each speaker. We apply this "dominance" approximation with the underlying assumption that speakers will seldom exhibit *equal* energy in frequency regions due to the diversity and sparsity of *different* formant structure. Furthermore, in multi-pitch estimation, we do not require reconstruction of the spectrogram and/or waveform itself, such that the term $\log\left(D_1 + \sum_{k=1}^{N_1} \alpha_{1,k} \cos \phi_{1,k}[n,\omega]\right)$ is sufficient for representing the pitch and pitch dynamic information (Chapter 3). In preliminary efforts, we observed that performance of the system degrades when we attempted to utilize 1) a "full set" of two peaks within a time-frequency region to obtain multi-pitch estimates and 2) the linear spectrogram instead of the log spectrogram. In relation to the former condition, the 2nd dominant peak in the GCT was virtually always "spurious" (see Section 5.3.2). In relation to the latter condition, improvements using the log operation are presumably due to the reduction of effects from the underlying formant structure in the GCT domain such that pitch and pitch dynamic information is "separated" in the GCT domain from the envelope term (Figure 5-17). Despite this assumption of dominance, the current pitch analysis framework nonetheless allows for representation of pitch and pitch *dynamic* information for individual speakers. As will subsequently be discussed, inclusion of this information is exploited to address the challenging condition of pitch trajectories that exhibit "close" values in frequency.

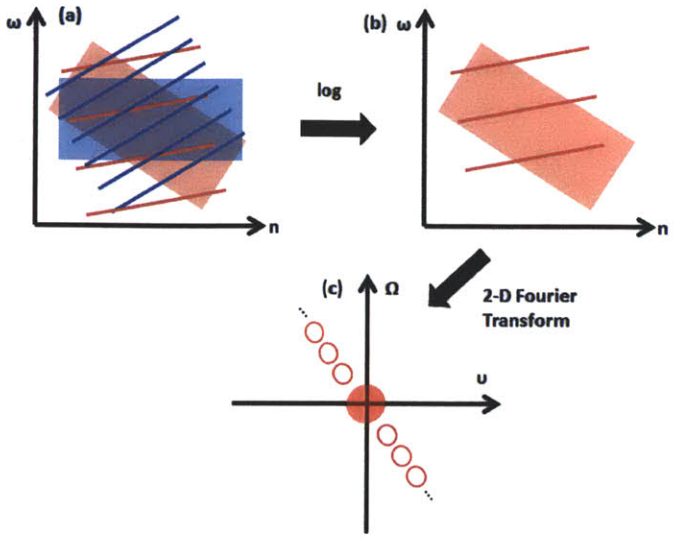

Figure 5-17. Schematic of (a) linear spectrogram showing sum of two speakers (red and blue); (b) dominance assumption of red speaker dominating after applying log operation; (c) GCT of (b) showing distinction of near-origin term (shaded) in GCT reflecting envelope and harmonic components reflecting sinusoids only (hollow) from (5.9).

5.3.2 Estimation/Tracking Algorithm

This section describes an algorithm for performing multi-pitch estimation/tracking on mixtures of all-voiced speech using this analysis framework. We refer to "estimation/tracking" as both extracting pitch candidates from a multi-pitch signal (e.g., a mixture of two speakers) as well as *assigning* pitch candidates to distinct speakers to generate pitch tracks. Our multi-pitch estimation algorithm consists of short-time analysis, GCT analysis, discriminant-based pitch candidate pruning, clustering, and a Kalman filtering framework as outlined in Figure 5-18.

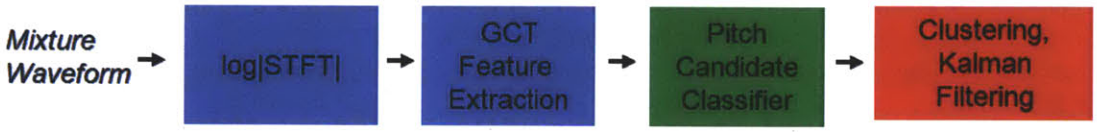

Figure 5-18. GCT-based multi-pitch estimation algorithm.

Short-time Analysis: Mixture waveforms were analyzed using the short-Time Fourier transform (STFT) to form the log spectrogram. A 32-ms Hamming window, 1-ms frame interval, and 512-point discrete Fourier transform (DFT) was used to compute the logarithm of the STFT magnitude, denoted as log-STFTM. A representative log-STFTM computed for a mixture of the "Walla Walla" and "Lawyer" sentences spoken by two female speakers is shown in (Figure 5-19).

GCT Analysis: The log-STFTM is subsequently used for GCT analysis. A 2-D high-pass filter was applied to log-STFTM to reduce the effects of the DC components in the GCT representation and is denoted as log-STFTM_{HP} [7]. Localized regions of size 800 Hz by 100 ms were extracted using a 2-D Hamming window from the magnitude of both log-STFTM and log-STFTM_{HP}. Overlap factors of 10 and 4 were used along the time and frequency dimensions and result in a set of center frequencies for GCT analysis along the frequency axis and overlapped regions for

analysis in time. A 2-D DFT of size 512 by 512 is used to compute two GCT's: GCTM (from log-STFTM) and GCT_{HP} (from log-$STFTM_{HP}$). Seven features were then extracted:

(1) Pitch estimate $\hat{f}_0$ from peak-picking the dominant peak in the magnitude of GCT_{HP} ($|GCT_{HP}|$).

(2) Pitch-derivative estimate $\frac{\partial \hat{f}_0}{\partial t}$ from the peak of (1)

(3) Amplitude of the dominant peak in $|GCT_{HP}|$.

(4) Normalized value of 3 relative to the DC value (6).

(5) "Harmonic to noise ratio" of dominant peak in $|GCT_{HP}|$

(6) DC value of GCTM

(7) Overall energy of GCTM

(1) and (2) are obtained using the pitch and pitch-derivative mappings of (3.24) and (3.26). Feature4 is computed as

$$Feature4 = \frac{Feature3}{\iint_{\omega,\Omega}|GCT_{HP}(\omega,\Omega)|}. \tag{5.15}$$

Feature 5 is computed as a "harmonic to noise ratio" by

$$Feature5 = \frac{\Sigma_n \Sigma_m |\alpha \cos \phi[n,m]|^2}{\Sigma_n \Sigma_m |s[n,m]|^2 - \Sigma_n \Sigma_m |\alpha \cos \phi[n,m]|^2}. \tag{5.16}$$

where $\alpha = Feature3$ and $s[n,m]$ to the localized region of log-STFTM; this metric assess the relative energy present in a sinusoid located at the dominant peak position relative to the energy present in its removal from the original time-frequency region. Features 3 - 7 relate to properties of the GCT not captured by the pitch and pitch-derivative and were motivated from the subsequent aim of pitch candidate pruning. The present analysis extracts a single peak to estimate the pitch value in 1) and contrasts the approach in 5.2.3 in extracting two peaks. This approach is favored for real speech mixtures due to the mixture of different formant structure between speakers (in contrast to identical formant structure in the simple multi-pitch analysis conditions presented previously).

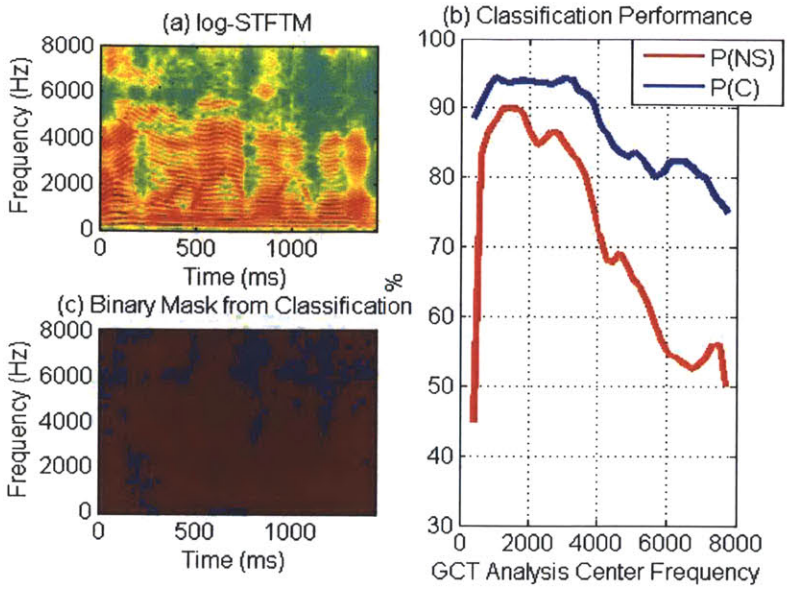

Figure 5-19. (a) log-STFTM of mixture of "Walla Walla" and "Lawyer" sentences spoken by a male and female speaker; (b) Band-wise classification performance of LDA on test data; P(C) = percent correct classification, P(NS) = prior probability of a "not spurious" candidate (c) Resulting binary mask of pruning with 1's (red) and 0's (blue).

Linear Discriminant Analysis-based Pruning: From Section 5.2.2, we observed that GCT analysis of multi-pitch signals can result in $\hat{f}_0$ far removed from the true pitch value (denoted as f_0). One cause for this is that some regions exhibit low amplitudes of the harmonic structure due to the formant structure. To account for these "spurious" candidates, linear discriminant analysis (LDA) [44] was applied to the previously described set of 7 features to prune the candidates. In training, we define a "spurious" candidate as one in which $|f_0 - \hat{f}_0| > \gamma$. γ is set to 3σ, where σ is the standard-deviation of the one-step differences in the pitch values of the training data. Specifically, given a pitch track $f_0[n]$, we compute the standard deviation σ of all the differences $|f_0[n-1] - f_0[n]|$ across all pitch tracks and utterances; in this work, $\sigma = 4.85$ Hz. A discriminant function is trained for each center frequency in GCT analysis and applied in a band-wise fashion along time to prune the candidates. Figure 5-19b shows classification performance on the testing data (Section 4) per band. Figure 5-19c shows a typical binary mask generated from pruning across time and center frequency for the mixture in Figure 5-19a. 1's denote regions in which candidates are kept while 0's denote regions in which they are discarded. Observe that the pruned regions in Figure 5-19c roughly corresponding to regions in which there is minimal harmonic structure in the spectrogram in Figure 5-19a.

Clustering and Kalman Filtering Framework: Given the pruned candidates across time-frequency regions, k-means clustering [44] was used to obtain local estimates in time. As our mixtures contained all-voiced speech from two speakers, two centroids were extracted from pruned candidates across all frequency bands at each time point. In contrast to clustering in Section 5.2.3, the present work performs clustering along *both* the pitch and pitch-derivative dimensions, where the pitch-derivative estimate is that tied to the pruned pitch value. *This method therefore accounts for conditions where pitch values may be identical but pitch-derivatives may differ for two speakers.*

To generate the pitch track for each speaker, each pair of centroids at a point in time were used as observations to a pair of Kalman filters (KF) [45]. For each speaker i we adopt a state-space model

$$\begin{aligned} \underline{x}_{t+1,i} &= A\underline{x}_{t,i} + \underline{v}_t \\ \underline{y}_{t,i} &= \underline{x}_{t+1,i} + \underline{w}_t \end{aligned} \tag{5.17}$$

where $A = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}$, $\underline{x}_{t,i} = \begin{bmatrix} f_0(t,i) \\ \partial f_0/\partial t\,(t,i) \end{bmatrix}$ is the true state, $\underline{y}_{t,i} = \begin{bmatrix} \hat{f}_0(t,i) \\ \partial \hat{f}_0/\partial t\,(t,i) \end{bmatrix}$ is the centroid, $\underline{v}_t$ and $\underline{w}_t$ are correspond to the model and observation noise terms of the state-space model, respectively, and are assumed to Gaussian. Given the assignment of a centroid to a speaker, the standard KF equations are used to generate the pitch track. In training, the covariances of $\underline{v}_t$ and $\underline{w}_t$ are obtained using the optimal assignment of centroids on a training set of mixtures. We define optimal as when the observation is closest in normalized distance to the true state. We normalize each observation by the means and standard deviations of the set of pruned observations at each time point and compute the geometric distance between the observation and the true state. In testing, estimation from observations is done using two Kalman filters that utilize the same parameters $\underline{v}_t$ and $\underline{w}_t$ obtained from the training data.

To perform assignment of the centroids to each speaker/pitch track in testing, we compute distances between the *predicted states* of the two pitch tracks (corresponding to speaker 1 and speaker 2) denoted as $\underline{\hat{x}}_{t,1|t-1}$ and $\underline{\hat{x}}_{t,2|t-1}$ and the two *observations* $\underline{y}_{t,a}$ and $\underline{y}_{t,b}$, all at time t. Specifically, we define $\chi_{1,a}$ as

$$\chi_{1,a} = \left(\underline{y}_{t,a} - \underline{\hat{x}}_{t,1|t-1}\right)^T \Lambda^{-1}{}_{t|t-1} \left(\underline{y}_{t,a} - \underline{\hat{x}}_{t,1|t-1}\right) \tag{5.18}$$

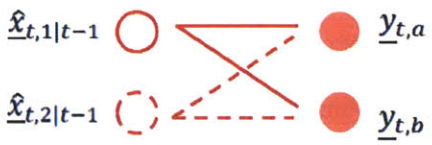

Figure 5-20. Assignment method with solid and dashed circles corresponding to distinct pitch tracks and lines corresponding to distance metrics between observations and the track.

where $\Lambda^{-1}{}_{t|t-1}$ is the covariance associated with the prediction at time t. The value of $\chi_{1,a}$ is the Mahalanobis distance [44] between the observation $\underline{y}_{t,a}$ and the predicted state from the Kalman filter of speaker 1 $\underline{\hat{x}}_{t,1|t-1}$. This metric represents how "likely" $\underline{y}_{t,a}$ was generated from the observation process of speaker 1 based on its predicted state. $\chi_{1,b}$, $\chi_{2,a}$, and $\chi_{2,b}$ are similary defined.

To make the assignments, the minimum of $\chi_{1,a}$ and $\chi_{1,b}$ is used to make the assignment to $\underline{\hat{x}}_{t,1|t-1}$; the same rule is applied for $\underline{\hat{x}}_{t,2|t-1}$ but with $\chi_{2,a}$ and $\chi_{2,b}$. If $\underline{\hat{x}}_{t,1|t-1}$ and $\underline{\hat{x}}_{t,2|t-1}$ acquire the *same* observation (e.g. if they both acquire $\underline{y}_{t,a}$), the assignments are changed based on the following criterion:

$$If\left(\chi_{1,b} + \chi_{2,a} > \chi_{2,b} + \chi_{1,a}\right), assign\ \underline{y}_{t,b}\ to\ \underline{\hat{x}}_{t,2|t-1}$$
$$Otherwise, assign\ \underline{y}_{t,b}\ to\ \underline{\hat{x}}_{t,1|t-1} \quad (5.19)$$

The same rule is applied if $\underline{\hat{x}}_{t,1|t-1}$ and $\underline{\hat{x}}_{t,2|t-1}$ both acquire $\underline{y}_{t,b}$ but with $\underline{y}_{t,a}$ replacing $\underline{y}_{t,b}$. This assignment method uses individual uncertainties of predicted observations and the combined uncertainty of both assignments to prevent pitch tracks from merging. Fixed-interval smoothing (across the entire duration of the pitch track) is applied to the filtered estimates[45]. We refer to the previously described method utilizing both pitch and pitch-derivative information in multi-pitch estimation as method "$f_0\_df_0/dt$".

Reference Approach for Comparison: To assess the utility of the GCT's joint representation of pitch and pitch-dynamics, we use a reference system that does *not* utilize $\partial f_0/\partial t$ explicitly. As in the "f0_df0/dt" method, candidates are pruned based on the $|f_0 - \hat{f}_0| > \gamma$ criterion, but k-means clustering is done using *only* the pitch values. In tracking, the state-space model of is modified such that $A = [1]$, $\underline{x}_{t,i} = [f_0(t,i)]$ is the true state, and $\underline{y}_{t,i} = [\widehat{f_0}(t,i)]$ is the centroid. All other steps are identical to the proposed method. We refer to this reference method as "f_0_only".

5.3.3 Data Set and Evaluation

For evaluation, a data set was collected consisting of 8 males (m1 - m8) and 8 females (f1 - f8) speaking 8 all-voiced utterances (Table 5-3); data was sampled at 16 kHz. Speakers were instructed to maintain voicing throughout each utterance. Reference (or "true") pitch values of the sentences were obtained using Wavesurfer prior to mixing [38]. Speech files were pre-emphasized at a 0-dB overall signal-to-signal ratio. To train the LDA-based pruning and Kalman filters, mixtures generated from m1 - m4 and f1 - f4 speaking s1 - s4 were used. In testing, mixtures generated from m5 - m8 and f5 - f8 speaking s5 - s8 were used. Distinct speakers and sentences were used in each mixture such that train and test sets consisted of 336 total mixtures each. We further divided the test data into mixtures of separate and close pitch track conditions. Close refers to mixtures where at least one time point contains a pair of pitch values within 10 Hz of each other. This accounts for 136 mixtures, the majority of which contained either crossings, or both crossings and mergings. The remaining 200 mixtures are considered separate. Representative mixtures are shown in Figure 5-22 through Figure 5-27. As a quantitative metric for performance, we use the root-mean-squared-errors metric defined in (5.7). Standard errors of the RMSE values were also computed.

5.3.4 Results and Discussion

Figure 5-21 shows average RMSEs in both the separate and close datasets for the two described estimation methods along with standard errors. In Figure 5-22 and Figure 5-23, we show the results of separate cases comparing the two methods. Consistent with the quantitative results in

Figure 5-21 we see that "$f_0\_df_0/dt$" exhibits similar performance to "f_0_only" and obtains reasonable estimates of pitch values. In contrast, observe that "$f_0\_df_0/dt$" outperforms "f_0_only" in the close conditions with crossing and/or merging pitch tracks (Figure 5-24 through Figure 5-27) due to both improved pitch candidate clustering and assignments. Nevertheless, an outstanding limitation of "$f_0\_df_0/dt$" is when pitch tracks exhibit similar pitch values and pitch-derivatives. As an example, observe that erroneous estimates are made by "$f_0\_df_0/dt$" after 800 ms in Figure 5-27, where the two pitch tracks are close in absolute frequency *and* have similar slopes.

Table 5-3. Table of all-voiced sentences for evaluating multi-pitch estimation.

s1 - "May we all learn a yellow lion roar."
s2 - "Why were you away a year, Roy?"
s3 - "Nanny may know my meaning."
s4 - "I'll willingly marry Marilyn."
s5 - "Our lawyer will allow your rule."
s6 - "We were away in Walla Walla."
s7- "When we mow our lawn all year."
s8 - "Were you weary all along?"

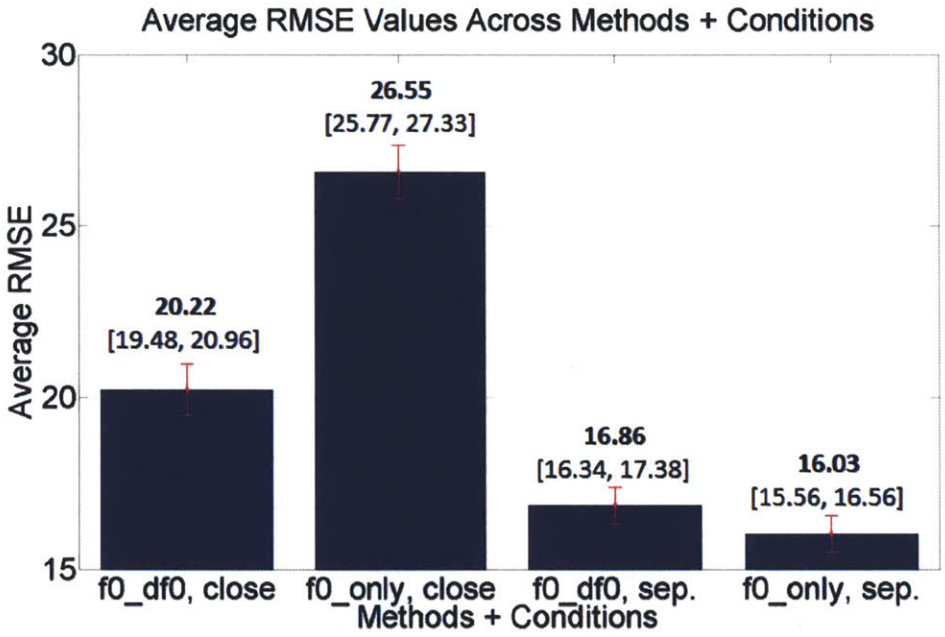

Figure 5-21. Average RMSE (Hz) (bars, bold) and standard errors [] across conditions and methods.

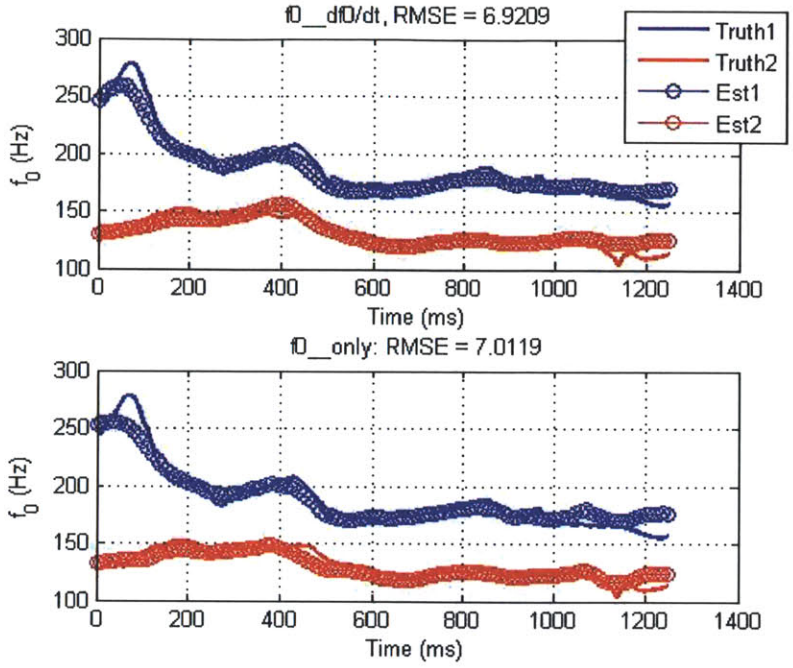

Figure 5-22. Estimation results from (top) "f0_df0/dt" and (bottom) "f0_only" for separate pitch tracks.

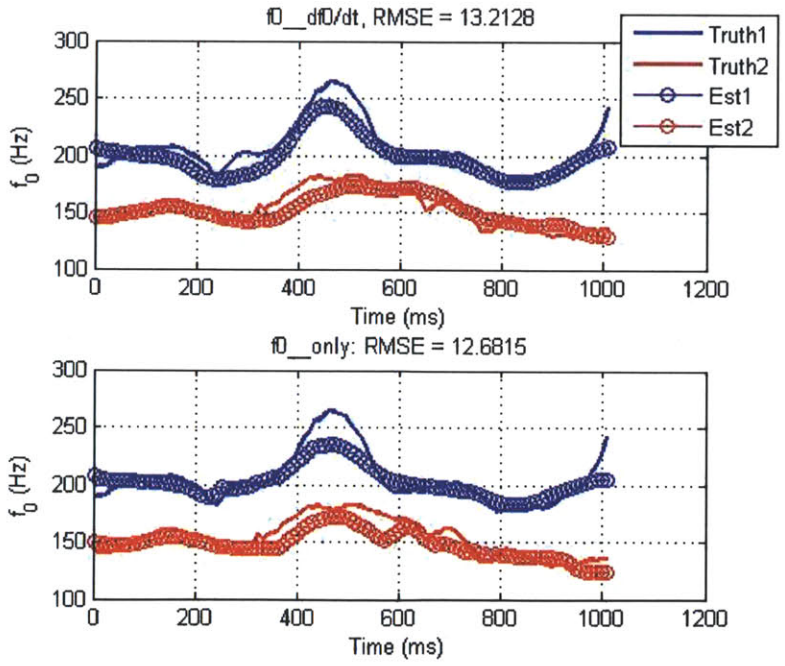

Figure 5-23. As in Figure 5-22 but for a distinct separate case.

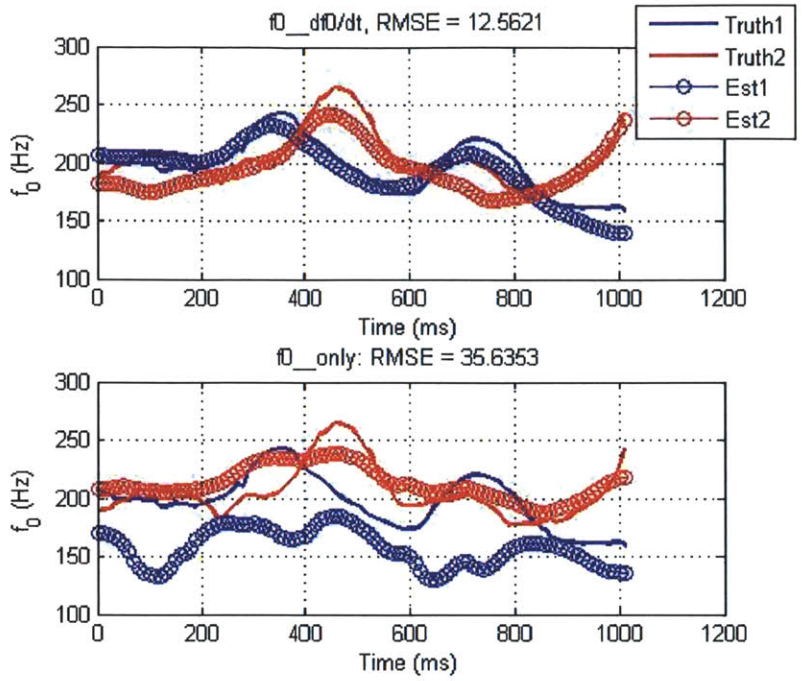

Figure 5-24. As in Figure 5-22 but for a close case for which *crossings* occur in pitch trajectories occur

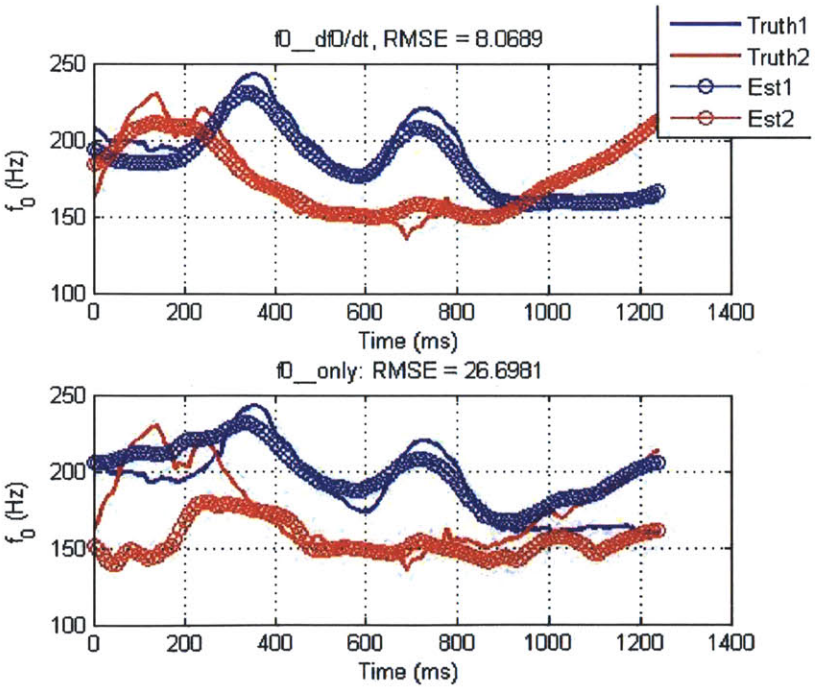

Figure 5-25. As in Figure 5-22 but for a close case with crossings.

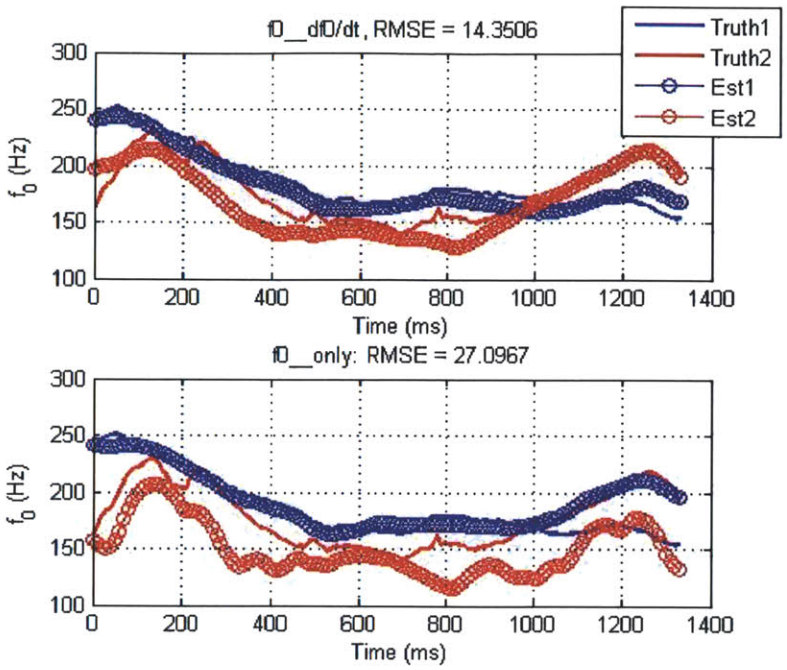

Figure 5-26. As in Figure 5-22 but for a close case with crossings *and* some merging (e.g., at 200 ms).

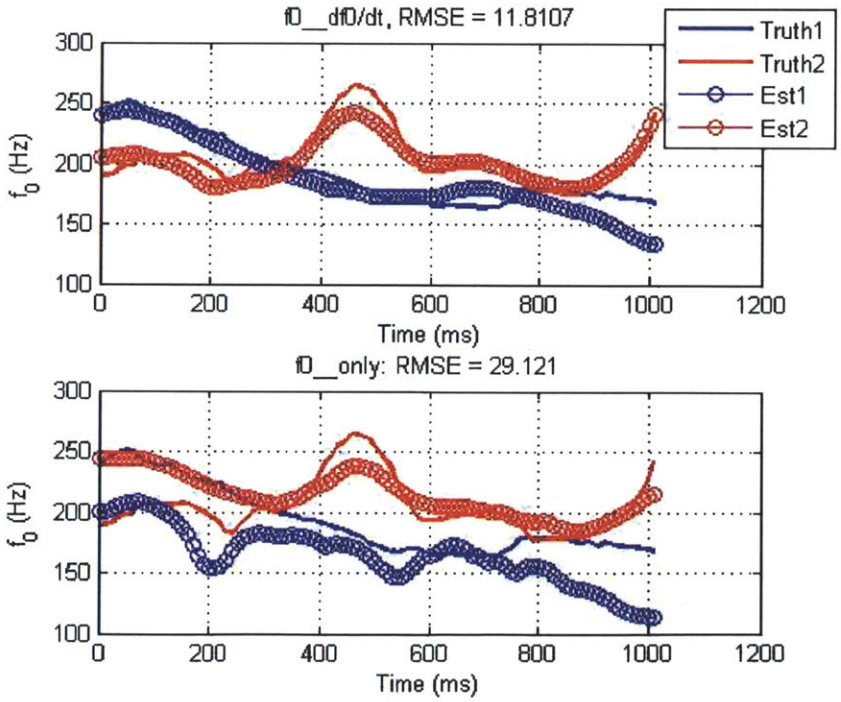

Figure 5-27. As in Figure 5-22 but for a close case with crossings *and* some merging (e.g., at 800 ms).

5.4 Conclusions

Results in this chapter on synthetic and real speech have demonstrated the utility of the GCT framework for multi-pitch analysis and estimation. By jointly representing pitch and pitch-derivative information from distinct speakers, the GCT can provide separability of pitch information from distinct speakers in analysis. Specifically, the GCT can be viewed as a more general signal representation for providing such separability with respect to traditional short-time analysis methods. We have further shown that an algorithm explicitly exploiting this joint representation can afford benefits in addressing the problem of close pitch trajectories in multi-pitch estimation at crossings and mergings. Limitations of the current algorithm are in its in ability to accurately estimate pitch values in conditions in which both the pitch value and pitch dynamics are the same between the two speakers. Furthermore, to apply the present framework to the general multi-pitch estimation problem, a method must be developed to account for mixtures in which there is both voiced and *unvoiced* speech as will be discussed in Chapter 7.

Chapter 6

Toward a Co-channel Speaker Separation System Using 2-D Processing of Speech

In this chapter, we describe efforts toward developing a complete system for the example application of co-channel speaker separation using the proposed 2-D signal models of Chapter 3 and Chapter 4. As in the previous chapters, our emphasis is on assessing the utility of the *models* themselves for this task rather than developing a state-of-the-art speaker separation system. We therefore highlight key difficulties encountered in the present development that limit the system for application to speech mixtures satisfying certain constraints to be subsequently discussed. Despite this limitation, our efforts nonetheless demonstrate that the GCT framework is a promising one for this task and motivates several important future directions of research.

This chapter is organized as follows. In Section 6.1, we briefly describe the overall framework of the system including a multi-pitch estimation and signal separation methods. Sections 6.2 and 6.3 describe these components in detail and highlight their limitations. Section 6.4 then describes our evaluation of the method and presents our results. We conclude in Section 6.5 and briefly discuss future directions.

6.1 Framework

A co-channel speaker separation system can be developed in the context of the GCT by utilizing elements of the multi-pitch estimation procedure in Chapter 5 with the signal separation methods of Chapter 3 and Chapter 4 (Figure 6-1). Overall, the system first obtains a set of multi-pitch estimates and voicing decisions and then utilizes these in signal separation according to the 2-D signal models proposed in this thesis. Subsequently, we discuss the development of both the multi-pitch estimation and signal separation methods and highlight their limitations.

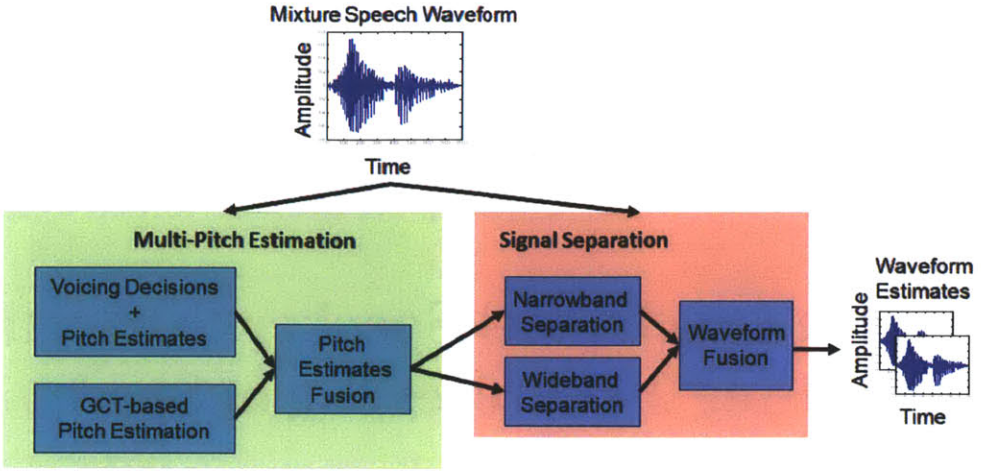

Figure 6-1. Framework for GCT-based speaker separation system including multi-pitch estimation (green) and signal separation (red) components.

6.2 Multi-Pitch Estimation

The general speech mixture case exhibits mixtures of voiced and unvoiced speech; in addition, pitch trajectory crossings and/or merging may occur for voiced speech (Chapter 5). As discussed in Section 3.5.1 this results in three *mixture voicing conditions*: voiced-on-voiced, voiced-on-unvoiced, and unvoiced-on-unvoiced. A complete multi-pitch tracking system would 1) accurately identify all three conditions, 2) estimate pitch values in the mixture when they are present, and 3) assign pitch values to their corresponding speakers.

In this section, we first describe two important difficulties encountered in our current development that led to the final multi-pitch estimation algorithm. Subsequently, we describe the multi-pitch estimation algorithm itself. As a consequence of these limitations, this algorithm is constrained to operating on speech mixtures in which 1) only *two* mixture voicing states (unvoiced-on-unvoiced and voiced-on-voiced) are available at the output and 2) the underlying pitch trajectories are assumed to be well-separated by pitch values alone.

6.2.1 Limitations of Existing Framework

Detection of Co-channel Mixture Voicing States: In relation to the multi-pitch estimation algorithm described in Chapter 5, recall that the method can only be applied to speech mixtures that are known a priori to be voiced-on-voiced. Detection of voicing either in the voiced-on-voiced or voiced-on-unvoiced mixture condition is therefore required as either a preliminary or integral component of a multi-pitch estimator for the general mixture condition containing voiced and unvoiced speech. As will subsequently be demonstrated, detection of the presence of voicing (i.e., *either* voiced-on-voiced or voiced-on-unvoiced) can be reasonably performed using standard methods in voicing detection used for analysis of a *single* speaker. Nonetheless, detecting the distinction between voiced-on-voiced and voiced-on-unvoiced is an outstanding challenge in the co-channel speaker separation task [46][47] (Figure 6-2).

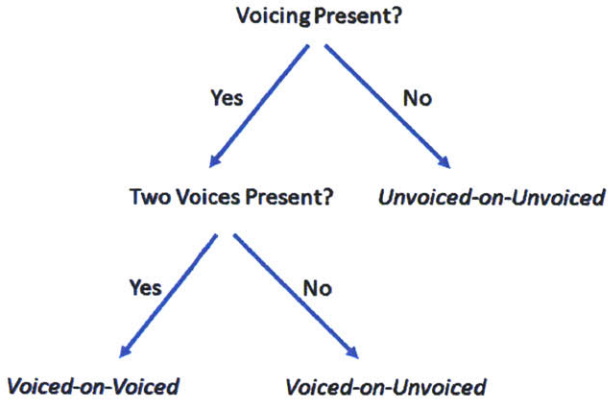

Figure 6-2. Decision-tree representation of mixture voicing conditions [46].

In our present development, we utilize a voicing detector (to detect presence of *any* voicing) as a preliminary step followed by a simplifying assumption that any region determined to be voiced is assumed to correspond to the voiced-on-voiced condition. Consequently, preliminary analyses demonstrated that applying the Kalman filtering framework of Chapter 5 results in very poor pitch estimates, presumably due to inaccurate tracking of pitch and pitch-dynamic values in voiced-on-unvoiced conditions. As will subsequently be discussed, an alternative estimation method using a simple k-means clustering method and pitch candidate pruning was applied. In future work, we may incorporate existing methods such as nonlinear state-space modeling, Bayesian classification, or GCT-based approaches to address this limitation[46][47].

Ambiguity of Assignment Using Pitch and Pitch Dynamic Information: As a second limitation of the current framework, we consider the mixture voicing condition of voiced-on-voiced. Assuming for now that detection of this condition is either known/detected a priori, we demonstrate by "counterexample" that *pitch and pitch dynamic information* alone are *insufficient* to address the general multi-pitch problem in the presence of unvoiced speech regions.

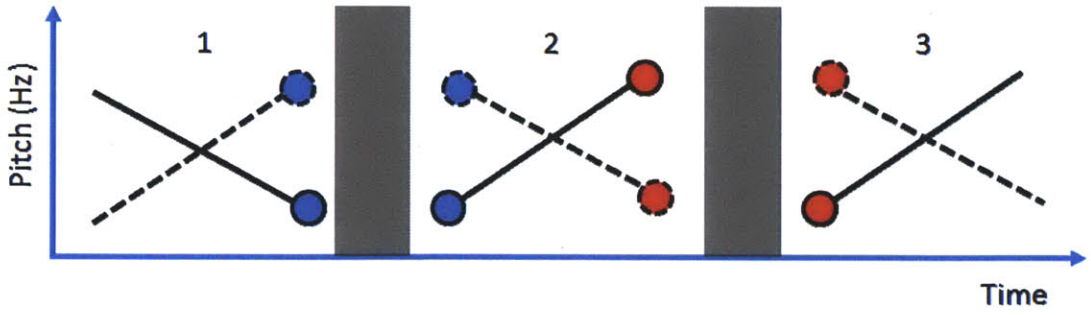

Figure 6-3. Schematic of pitch trajectories of two speakers (dashed and solid lines) in the presence of unvoiced regions (shaded grey) across three regions (1-3); optimal assignments based on either pitch value alone (blue) or pitch dynamic information (red).

In Figure 6-3, we show schematized pitch tracks of two speakers in the presence of unvoiced regions with crossing pitch trajectories. In Region 1 note that pitch and pitch dynamics combined can be used to group pitch values to distinct pitch tracks (as in Chapter 4). Upon reaching the subsequent unvoiced region, there will exist ambiguity in relating pitch tracks in Region 1 with those in Region 2. The optimal assignment shown in this transition is based on the proximity to

the *pitch values* themselves. Note that the *incorrect* assignment would be made if we grouped speakers based on pitch dynamics in Region 1. Continuing to Region 2, observe that pitch and pitch dynamic again may be used to track the two pitch tracks accurately. However, following the unvoiced region between Region 2 and 3, the optimal assignment in transitioning from Region 2 to Region 3 is based on using only *pitch dynamics*; using pitch information alone would lead to the *incorrect* assignment. The present example demonstrates therefore that pitch and pitch dynamic information alone (e.g., as represented in the GCT) are *insufficient* features in addressing the general multi-pitch estimation problem. This example further generalizes to conditions in which the pitch *and* pitch dynamics of a speaker are similar within a time span as highlighted by Figure 5-27 in Chapter 5.

As a consequence of this limitation, the multi-pitch estimation algorithm in our present development is therefore constrained to operate on speech mixtures that have pitch trajectories well-separated by the pitch values themselves. To address this limitation, one possible area of future work is to combine pitch and pitch dynamic content with estimated spectral characteristics estimated for individual speakers (e.g., formant structure) for use in grouping. For instance, while pitch and pitch dynamic information leads ambiguity in assignment in the example presented, the general spectral envelope (e.g., an average) of the distinct speakers may remain constant across all three regions in Figure 6-3.

6.2.2 Multi-Pitch Estimation Algorithm

In our previous discussion, we have highlighted two important difficulties in multi-pitch estimation in relation to the existing Grating Compression Transform-based processing framework. As a consequence of these limitations, we describe in this section a simple multi-pitch estimation algorithm that is restricted in two ways. First, its outputs in terms of mixture voicing conditions consist only of the unvoiced-on-unvoiced and voiced-on-voiced cases. In addition, the algorithm can be readily applied only to speech constrained to have pitch values that are well-separated based on pitch values alone (i.e., without crossings and/or merging of the pitch trajectories).

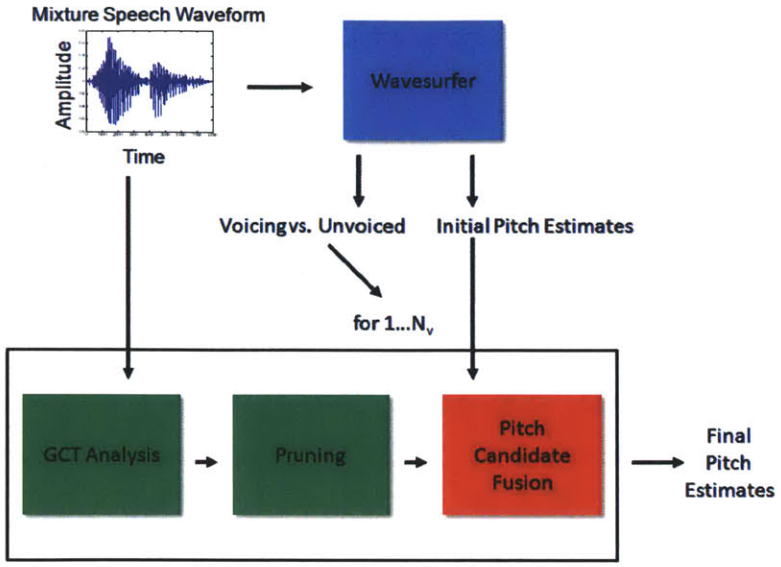

Figure 6-4. Schematic illustrating steps in multi-pitch estimation algorithm.

In Figure 6-4, we show a schematic of the multi-pitch algorithm developed starting with voicing detection of the mixture signal as well as establishing an *initial* set of pitch estimates using the Wavesurfer software package [38]; this first step further provides a segmentation in time of the waveform between regions of voiced speech and unvoiced speech. Denoting N_v as the number of *voiced* regions, we perform pitch candidate extraction and pruning from the GCT and *fuse* pitch estimates from the GCT and the *initial* pitch estimates through a pitch-based grouping to the *final* pitch estimates for each individual speaker. We describe these steps subsequently in detail.

Wavesurfer-based Voicing Detection and Initial Pitch Estimates: In the first step, we apply a *single* pitch tracker to obtain an estimate of regions of voicing versus unvoiced regions as well as an initial set of pitch estimates. Specifically, we apply the Wavesurfer pitch estimation procedure to the mixture waveforms. Wavesurfer obtains estimates of pitch and voicing using correlation-based analysis followed by dynamic programming as described in [48]. A representative estimate arising from this on a female-male mixture is shown in Figure 6-5a; we denote the true female and male pitch tracks as $f_{0,f}[n]$ and $f_{0,m}[n]$ and the estimate as $f_{0,e}[n]$. Observe that the pitch estimator obtains accurate pitch values of individual speakers within distinct time regions but exhibits "jumps" between speakers; this is effect is due to the dominance of individual voiced speakers in distinct time regions. In addition, observe that despite the presence of two speakers, the pitch estimator nonetheless accurately detects the presence of voicing; here, voicing refers to when either the voiced-on-voiced *or* voiced-on-unvoiced mixture condition occurs. To illustrate this voicing detection, we show in Figure 6-5b a plot of

$$v_e[n] = 1, f_{0,e}[n] > 0; 0, otherwise \quad (6.1)$$

$$v[n] = 1, f_{0,f}[n] > 0 \; or \; f_{0,m}[n] > 0; 0, otherwise \quad (6.2)$$

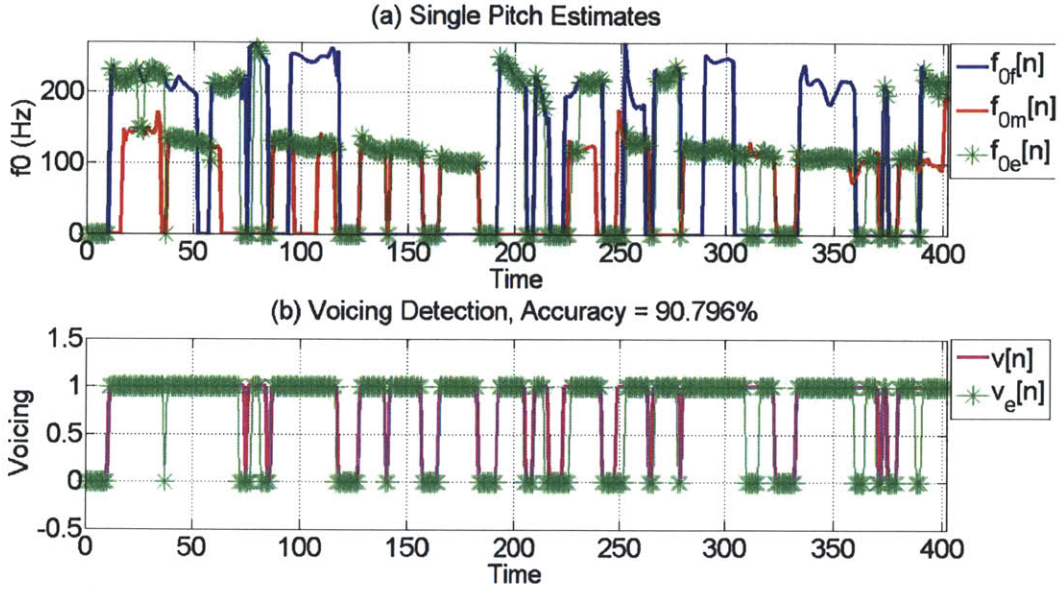

Figure 6-5. (a) Pitch estimates (green, '*') from a single pitch estimator (Wavesurfer) on a mixture signal of two speakers (red and blue); (b) Voicing detection as determined by presence of pitch value (green, '*') and true voicing of mixture (maroon).

Quantitatively, $v_e[n]$ provides a good match to the true voicing state using an accuracy metric

$$Accuracy = 100\frac{N_{match}}{N} \tag{6.3}$$

where N is the length of $v[n]$ and N_{match} is the count of $v[n] = v_e[n]$. Similar observations were observed on other mixtures. More specifically, an average accuracy across all the data

Fusion to Obtain Final Pitch Estimates: In the next step of multi-pitch estimation, GCT analysis is performed on the narrowband spectrogram of the mixture to obtain the refined pitch estimates. Specifically, we define a *voiced* region in time from $v_e[n]$ in the previous section as any subset of time points where $v_e[n] = 1$. This rule results in a set of N_v voiced regions.

We perform GCT analysis in each voiced region in time by extracting pitch values and features identical to those described in Section 5.3.2 using a narrowband spectrogram. Similarly, the set of pitch values are *pruned* as in Section 5.3.2 using a bandwise linear discriminant classifier applied across time. Finally, using the set of pruned pitch candidates denoted as $\boldsymbol{f}_{0,cands}[n]$, we obtain pitch tracks within each voiced region using a simple assignment method for each point in time. Specifically, k-means clustering is first performed on the non-zero pitch values of the *initial pitch estimates* from Wavesurfer (i.e., $f_{0,e}[n]$) to obtain two centroids c_{low} and c_{high} corresponding to the estimated pitch tracks $f_{high,e}[n]$ and $f_{low,e}[n]$. Subsequently, for each point in time, we assign the value of $f_{0,e}[n]$ to either $f_{high,e}[n]$ or $f_{low,e}[n]$ based on

$$min\left(|f_{0,e}[n] - c_{low}|, |f_{0,e}[n] - c_{high}|\right) \tag{6.4}$$

where *min* is performed with respect to either the *low* or *high* pitch centroids (corresponding to the pitch tracks $f_{high,e}[n]$ and $f_{low,e}[n]$). For illustrative purposes, assume that $f_{0,e}[n]$ is assigned to $f_{low,e}[n]$. Then, we set $f_{high,e}[n]$ equal to

$$f_{high,e}[n] = mean(\boldsymbol{f}_{0,cands-high}[n]) \tag{6.5}$$

where the set $\boldsymbol{f}_{0,cands-high}[n]$ is the set of pruned pitch values closer in absolute distance to c_{high} than c_{low}.

The described method is performed across all voiced regions to obtain pitch estimates of the individual speakers. All unvoiced regions (i.e., when $v_e[n] = 0$) are set to zero. In Figure 6-6a, we show pitch estimates obtained at each point in time based on the described multi-pitch estimation method. Observe that these tracks correspond reasonably well to the true pitch tracks in voiced regions. Nonetheless, the current method has two primary limitations. First, due to the voicing decision reference accounting for both the presence of either two voices or a single voice, estimates can often be erroneous in obtaining pitch estimates when the true pitch track is unvoiced (e.g., compare high pitch track at time ~150 with the estimate). Furthermore, the current method cannot account for pitch trajectories that are merged/crossing due to its reliance on pitch values in assignment.

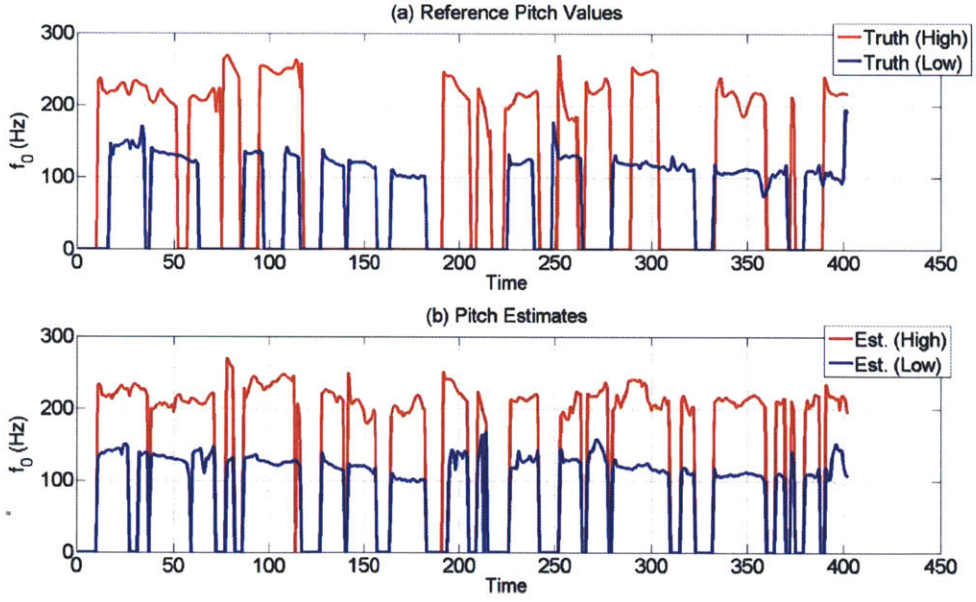

Figure 6-6. (a) True pitch tracks of high- (red) and low-pitched (blue) speaker; (b) Estimates of high (red) and low (blue) pitch tracks.

6.3 Signal Separation

As previously discussed, the developed multi-pitch tracker classifies voicing mixtures as either *voiced-on-voiced* or *unvoiced-on-unvoiced* for mixtures along with pitch estimates. The voiced-on-unvoiced condition is omitted. To utilize these estimates in signal separation, we apply the demodulation techniques described in Chapter 3 and Chapter 4 for the narrowband and wideband

GCT representations. Specifically, pitch estimates of individual speakers are mapped to the GCT domain for use in demodulation and least-squared-error (LSE) fitting to the mixture spectrogram, thereby resulting in spectrogram estimates of the individual speakers. We refer the reader to details of these methods in previous chapters and briefly summarize here the overall approach for the full separation system as well as motivate modifications to the algorithms. In addition to the individual narrowband and wideband estimates, we also perform *fusion* of the waveforms according to the simple linear weighting as in Equation (4.44).

Narrowband Algorithm: Due to the multi-pitch estimates representing only the voiced-on-voiced and voiced-on-unvoiced voicing mixture conditions, the case of *voiced-on-unvoiced* is in fact "hidden" within the time points labeled as voiced-on-voiced. The direct and exclusion/re-estimation methods described in narrowband-based separation (Section 3.4) can therefore be expected to "enforce" harmonic structure in the spectrogram estimate for the unvoiced speaker if the underlying mixture voicing condition is voiced-on-unvoiced.

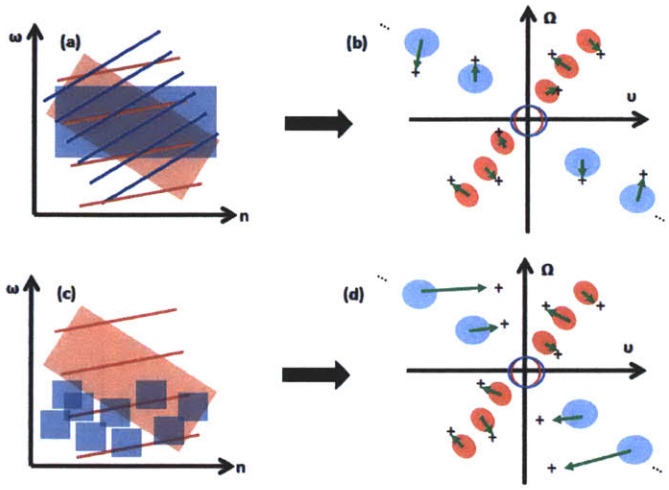

Figure 6-7. (a) Schematic illustrating local time-frequency region containing two voiced speakers (red and blue); (b) GCT representation of (a) with mapped carrier positions (shaded) and candidate positions ('+') from peak-picking; mapped positions are reassigned (green arrows) to candidate positions based on distance; terms near the GCT origin (hollow red and blue) are overlapped according to the signal model (see Section 3.4); (c) local time-frequency region with voiced (red) and *unvoiced* (blue) speaker; (d) GCT of (c) indicating mapped harmonic locations for the red speaker and blue speaker; the blue speaker mappings are incorrect due limitations in multi-pitch estimation.; reassignment results in the blue speaker carrier positions resembling a noise carrier.

In an attempt to avoid this effect, for narrowband-separation, we perform *bootstrapping* for the nominally labeled voiced-on-voiced cases as was done for the wideband-based separation method (Section 4.6.2). Figure 6-7a and b schematizes this approach. A set of directly mapped carrier positions obtained from pitch and pitch dynamic mappings are initially obtained. In addition, an alternative set of carrier positions obtained from peak-picking are also made available. Subsequently, each directly mapped position is *reassigned* to a position obtained from peak-picking through an iterative method that minimizes the distance between the mapped and peak-picked positions (Section 3.2.2). By reassigning carrier positions to those obtained from the GCT itself (rather than the directly mapped positions), a set of carrier positions for an individual speaker can be *reassigned* to correspond to those of a noise carrier in the voiced-on-unvoiced

case as shown in Figure 6-7c and d. Consider for instance Figure 6-7a and b, in which two voiced speakers (red and blue) are indeed present in the local time-frequency region as indicated by two sets of harmonic line structure. Bootstrapping in this case will result in reassignment of the mapped carrier positions to those that are located near the mapped positions. In contrast, in Figure 6-7c and d, the local time-frequency region corresponds to the voiced-on-*unvoiced* case with the red speaker being voiced while the blue speaker is unvoiced. Due to limitations of the multi-pitch tracker, however, a set of harmonically related positions will be generated that reflects "voicing" by the blue speaker. Nonetheless, in *reassignment* of the carrier positions, the blue speaker can be appropriately assigned to carrier positions reflecting the speaker's underlying unvoiced component as a noise carrier.

Wideband Algorithm: In the *wideband* representation, we similarly apply the identical steps as in Section 4.7.3 in using the *bootstrapping* method for assignment and estimation of carrier positions. In the voiced-on-voiced condition, pitch values are first mapped to the GCT space followed by an iterative algorithm to reassign carrier locations from peak-picking to the mapped locations (Section 4.7.3).

In both the narrowband and wideband separation approaches, in the unvoiced-on-unvoiced condition, we perform an least-squared error (LSE) fit to the data and assign half the amplitude of the fit to each speaker as was done in both Section 3.4 and Section 4.6.2. Finally, in fusion, we use the simple linear weighting method with $\alpha = 0.41$ as in Section 4.6.2.

6.4 Evaluation

Data and Evaluation Criteria: As previously discussed, limitations of the multi-pitch algorithm constrain its applicability to speech mixtures in which the underlying pitch tracks are well separated based on the pitch values alone. We use for our data only the male and female mixtures described in Chapter 3 to better satisfy this constraint. Our results are evaluated quantitatively using the global SNR values relative to 0 dB signal-to-signal ratio as in Section 3.5.3.

Results: In Figure 6-8, we show spectrograms of an example mixture and the two true targets. In Figure 6-10 and Figure 6-11, we show spectrograms of the wideband, narrowband, and fusion estimates for the target male and female speakers, respectively. From these results, we observe that despite incomplete and potentially erroneous pitch estimates, the bootstrapping signal separation method is capable of suppressing interfering speakers for both the male and female targets. Presumably, bootstrapping allows for carrier positions to be more accurately determined despite pitch estimation errors. Quantitatively, Table 6-1 shows average SNR gains of these methods for male and female targets 3~4 dB for both methods alone; in addition, we list the values obtained in separation using prior pitch information as in Table 3-2 and Table. 4-4. Consistent with our previous results in Chapter 4, fusion of the methods results in gains in SNR above 4 dB. In informal listening, estimated waveforms exhibited similar reconstruction quality of the target speaker, though with less suppression of interfering speakers than in waveforms from separation using prior information consistent with the ~2 dB reduction in SNR.

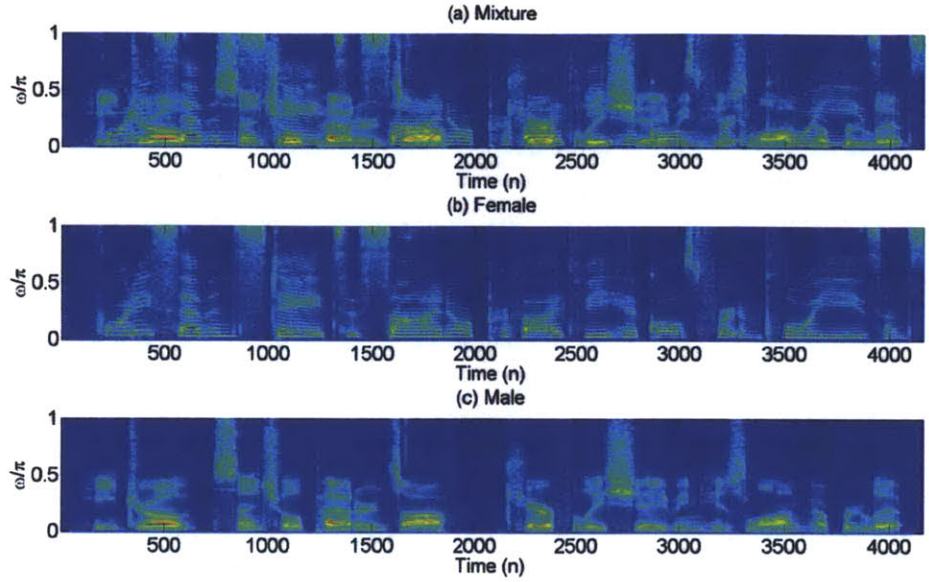

Figure 6-8. (a) Mixture of female ("Forty-seven states assign or provide vehicles for employees…") and male utterances ("He drove essential patterns off, carefully shaving his long upper lip."); (b) original female target; (c) original male target;

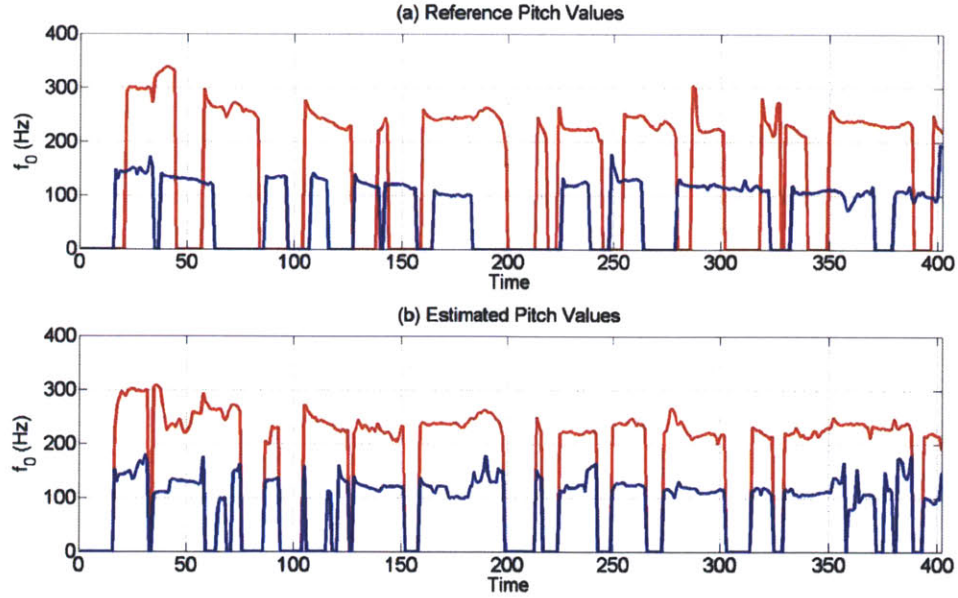

Figure 6-9. (a) Reference and (b) estimated pitch values for female (red), male (blue) mixture.

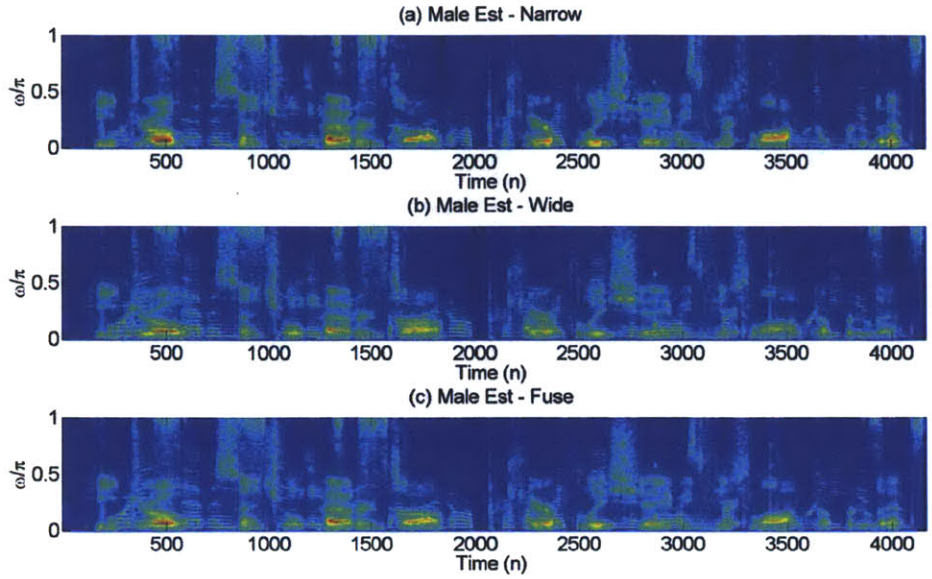

Figure 6-10. Estimates of male target obtained from (a) narrowband, (b) wideband, and (c) fusion.

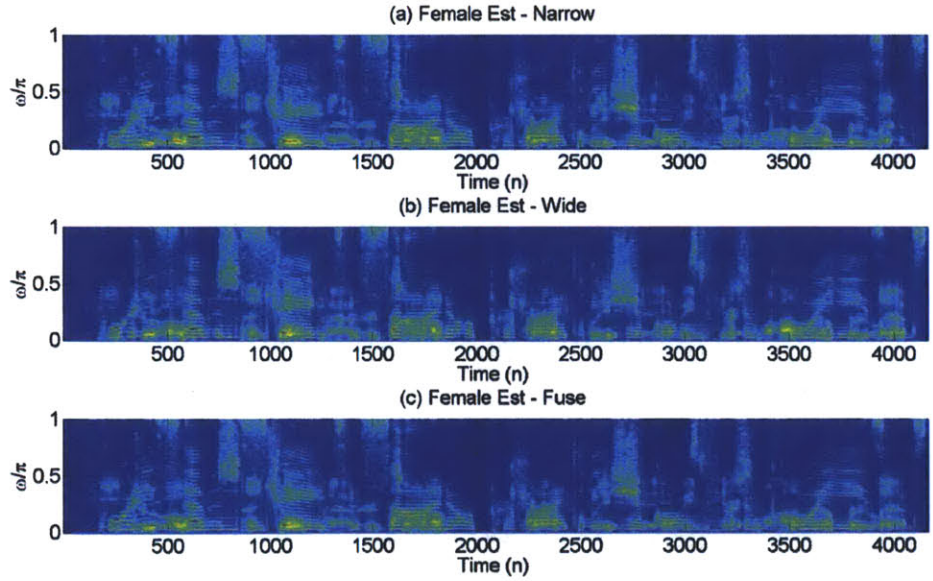

Figure 6-11. Estimates of female target obtained from (a) narrowband, (b) wideband, and (c) fusion.

Table 6-1. Average SNR gains (relative to 0 dB) and associated errors of full separation system.

	Narrowband	Wideband	Fusion	Prior Pitch (Fusion)
Female	4.28 [0.13]	2.79 [0.15]	4.42 [0.14]	6.55 [0.14]
Male	3.62 [0.14]	3.55 [0.12]	4.46 [0.10]	6.35 [0.10]

6.5 Conclusions

In this chapter, we have discussed development toward a complete system for co-channel speaker separation. A multi-pitch estimation algorithm was developed by fusing an initial set of pitch estimates obtained from a standard speech analysis tool in conjunction with GCT-based pitch estimates. The pitch estimates are used in signal separation by applying demodulation techniques in the GCT domain to account for inaccuracies of the pitch estimates. Though application of our multi-pitch estimation method is restricted, the present results further demonstrate the GCT's ability to represent speech content as well as its promise as a signal representation for the co-channel speaker separation problem. As previously discussed, limitations of the current approach can be explored in future work to further improve the system. Specifically, a more sophisticated voicing mixture condition detector can be incorporated into the existing framework; in addition, alternative grouping mechanisms for speakers in pitch tracking (in addition to pitch and pitch dynamics) may be used to address ambiguity of pitch value assignment in multi-pitch tracking.

Chapter 7

Conclusions and Future Directions

7.1 Contributions

Speech-signal modeling and processing has traditionally been performed using short-time analysis methodologies that analyze temporally local spectral content of a speech waveform. In this thesis, we have proposed a class of signal models in an alternative framework using two-dimensional processing (2-D) of speech. Analytically, our models arise out of interpreting fundamental speech production properties such as pitch, pitch dynamics, and formant structure (both stationary and dynamic), and onset/offset content (e.g., voicing onset) in local time-frequency regions of the canonical wide and narrowband spectrograms. The resulting models for both cases are viewed analytically through sinusoidal-series amplitude modulation. In developing the model, we have therefore provided an *explicit* interpretation of the concept of "modulation" as represented in spectrograms. This is in contrast to existing approaches that have interpreted "modulation" either qualitatively through phenomenological analyses or data-driven machine learning methods (Chapter 2).

As an interpretive framework, the modulation model views production properties such as pitch, pitch dynamics, noise source, and (under certain conditions) formant bandwidth in a sinusoidal series *carrier*. In modulation, this carrier is modulated (i.e., multiplied) by a local envelope representing formant structure, formant dynamics, and onsets/offsets. Here, our interpretation of "local" refers to analysis within small time-frequency regions of the spectrogram. In the corresponding GCT space, modulation results in *distribution of copies* of envelope structure in either an ordered fashion (i.e., in voiced speech) or without ordering (i.e., in noise.). In developing this model, we have highlighted its properties and limitations through simulations on synthetic speech and real speech examples. Furthermore, we have demonstrated its ability to represent a variety of speech content through spectrogram reconstructions. Additional evidence for this was presented in demonstrating its applicability to the co-channel speaker separation problem in multi-pitch analysis/synthesis, signal separation and reconstruction, and a simple preliminary separation system.

7.2 Research Issues

A number of outstanding research issues remain in the context of this thesis. In this section, we describe several important issues that may be addressed in future work in relation to modeling and representation (Section 7.2.1) and the co-channel speaker separation task (Section 7.2.2).

7.2.1 Modeling and Representation

Modeling Limitations: This thesis has derived a set of models for local time-frequency regions of the canonical narrowband and wideband spectrograms. While empirical results have demonstrated the ability of these models to accurately represent speech, we have also highlighted through simulations their limitations. As one example of our models' limitations, models of voiced speech for narrowband models were shown to provide reasonable mappings of pitch and pitch dynamic information in the GCT space (Figure 3-5). Nonetheless, the limiting assumption

of local regions reflecting *parallel* harmonic lines was implied to cause errors in this mapping. Further investigation is necessary to formulate an improved model of dynamic pitch content from an analytical perspective. For instance, one potential direction is related to preliminary simulations performed in [16], in which it was demonstrated empirically that the 2-D "fanned" harmonic line structure could be reasonably well-modeled using a 2-D chirp-like function with an exponentially decaying basis along time and a sinusoidal basis along frequency. Similarly, in relation to the wideband models, the current model for incorporating *dynamic* formant content was derived under relatively strict conditions (see Section 4.9). Despite its consistency with empirically observed behavior on synthetic vowels (see Figure 4-10), further investigation is needed to provide a model for dynamic formant content that is less restrictive in its region of validity.

Another fundamental modeling issue is that the derived models have been exclusively developed in the spectrogram *magnitude* domain. Consequently, in signal processing applications such as co-channel speaker separation, we have had to invoke linearity assumptions of the magnitude spectrogram. Further investigation is required to better understand the role of the short-time Fourier transform (STFT) *phase* in the context of a 2-D processing framework.

Choice of Short-time Analysis Windows and Region Sizes in Representation: Throughout this thesis, we have employed fixed-size analysis windows in both short-time Fourier analysis to generate the STFTs and corresponding spectrograms as well as fixed-sized local time-frequency regions for GCT analysis. While the choice of window/region sizes can be motivated from analytical as well as empirical findings, we briefly describe here areas of future research in relation to these choices.

In the short-time analysis domain, we chose window sizes to match the canonical narrowband and wideband spectrograms, as these are the most commonly used time-frequency distributions in speech analysis. Nonetheless, further investigation is needed to explore more carefully the effects of this choice of window. In Figure 7-1 through Figure 7-4, we show the effects of varying the analysis window length in analyzing a periodic impulse train with pitch of 150 Hz. The window length begins in Figure 7-1 at 50 ms corresponding distinctly to a narrowband spectrogram and decreases through to Figure 7-4 to 1 ms corresponding to a wideband spectrogram. In Figure 7-3, observe that an "intermediary" case between narrowband and wideband representations can be observed in both the spectrogram and the GCT domain, *thereby providing analysis of multiple resolutions*. Specifically, the spectrogram exhibits both vertical and horizontal grating patterns in the time-frequency space while the GCT correspondingly contains peaks along both the vertical and horizontal axes. Similar observations were also made using this analysis procedure for a Gaussian white noise sequence in which components along the Ω-axis in the GCT were observed for a narrowband representation and transitioned to components along the υ-axis for a wideband representation. Further investigation is necessary to assess the meaning of these effects and their potential application for speech analysis through *multi-resolution* analysis (e.g., as in the auditory spectrogram).

In the GCT domain, the local region size was similarly motivated from analytical and empirical findings. Nonetheless, as can be shown in Appendix A, improvements in analysis/synthesis using *adaptive* local region sizes can be obtained using a "goodness metric" in relation to the underlying signal model. Though the results in Appendix A demonstrate only a small gain in analysis/synthesis, we believe that this adaptive framework can provide benefits in signal processing applications such as speech enhancement.

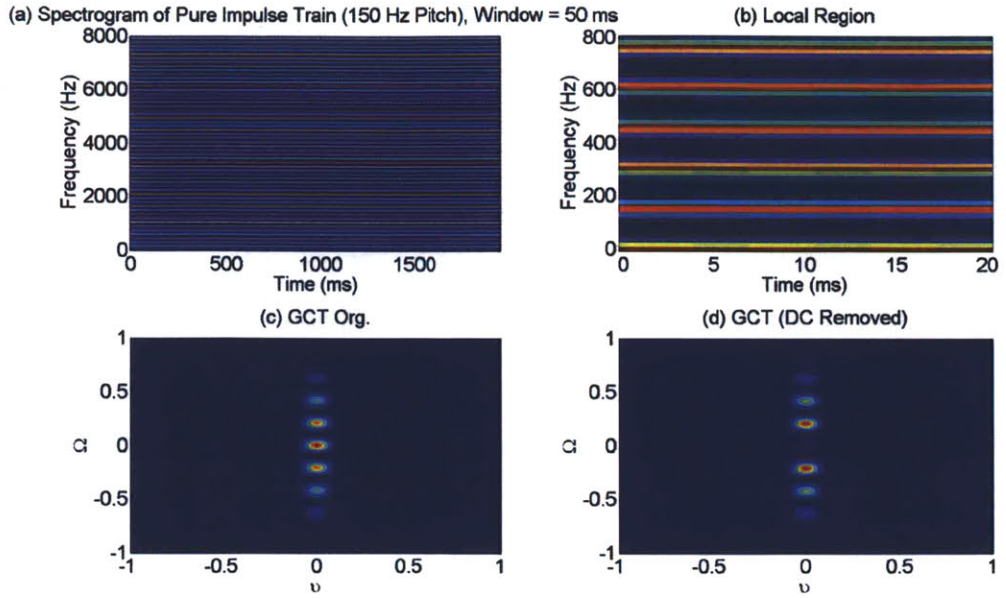

Figure 7-1. (a) Spectrogram of pure impulse train with 150-Hz pitch using a short-time analysis window of 50 ms (frame rate of 0.5 ms); (b) local region extracted for analysis; (c) GCT of (b); (d) GCT of (b) but with DC components removed for display purposes.

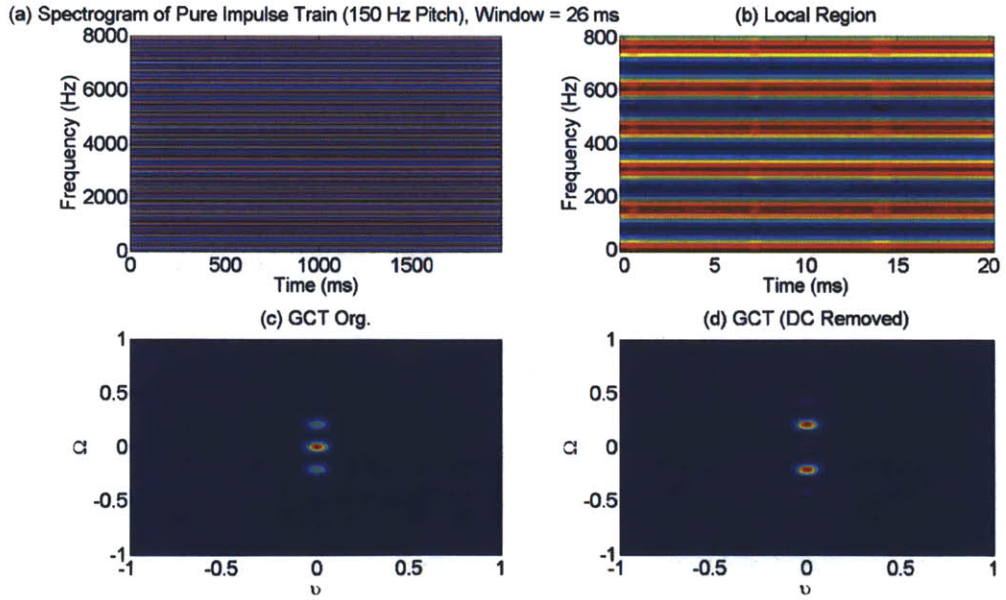

Figure 7-2. As in Figure 7-1 but for window of size 26 ms.

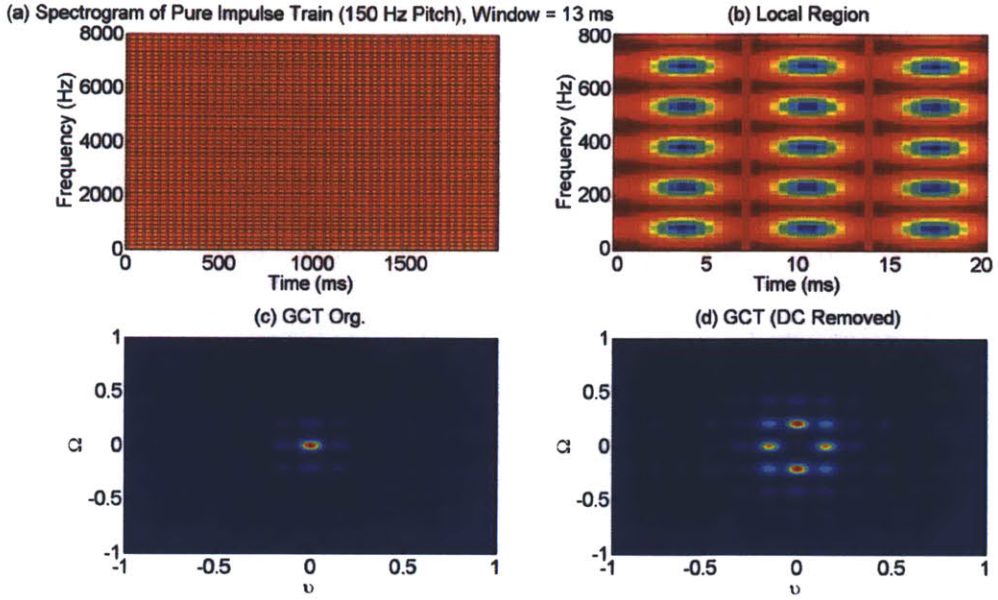

Figure 7-3. As in Figure 7-1 but for window of size 13 ms; observe in (c) and (d) components oriented along both the horizontal and vertical axes.

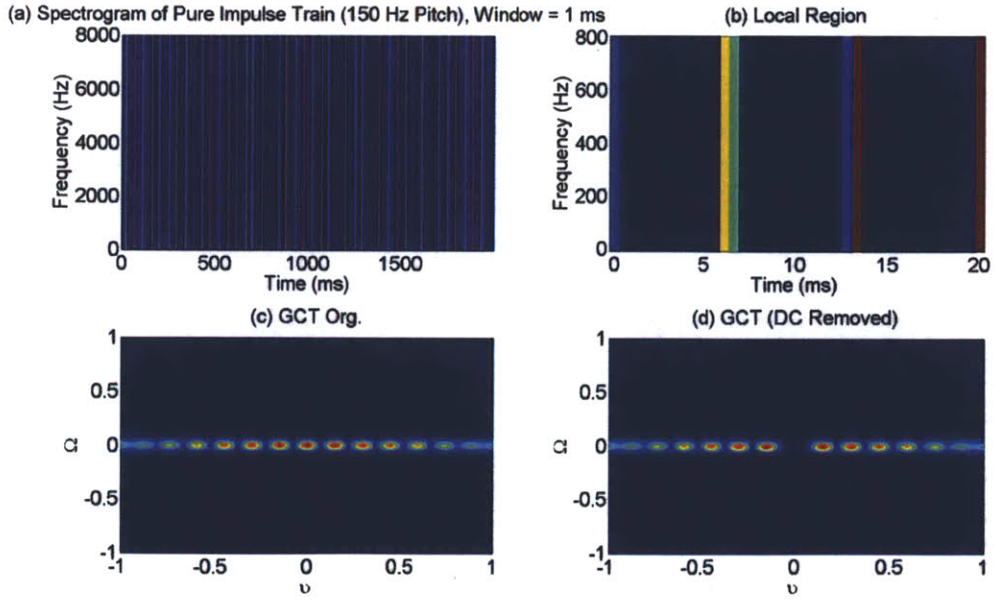

Figure 7-4. As in Figure 7-1 but for window of size 1 ms.

7.2.2 Co-channel Speaker Separation

In co-channel speaker separation, we have described in Chapter 6 a preliminary full system based on combining multi-pitch estimation and signal separation algorithms based on the GCT and the

standard pitch estimation algorithm (for single speakers) of Wavesurfer. As described in that chapter, two important issues must be addressed to develop a system capable of handling more general speech mixture conditions, particularly in relation to multi-pitch estimation. First, detection of the voiced-on-voiced versus voiced-on-unvoiced must be performed to ensure accurate determination of voicing mixture conditions; this step is crucial in both multi-pitch estimation and signal separation. A GCT-based approach for achieving this is schematized in Figure 7-5 based on detection of components in the GCT space along similar orientations. Note that this approach can be further generalized to perform detecting an arbitrary number of speakers present in a local time-frequency region of the spectrogram if it is known a priori that voicing exists in the region. Second, despite the ability of multi-pitch estimation using the GCT to obtain accurate estimates of "separate" and "close" voicing mixtures, an outstanding limitation is that both the pitch and pitch-dynamic cues used in the GCT approach are not sufficient to address the more general condition of speech mixtures with similar pitch values *and* in the presence of unvoiced regions (Figure 6-3). As suggested in Chapter 6, one approach could be to utilize alternative grouping cues such as spectral envelope to group individual speakers under these conditions.

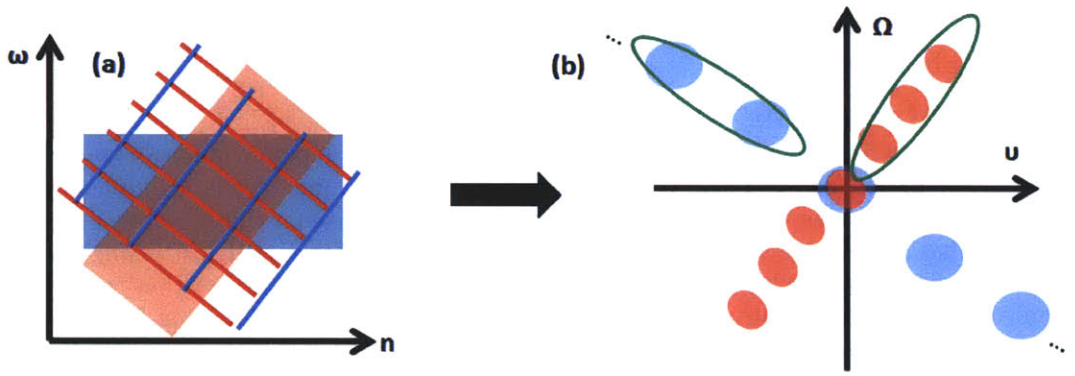

Figure 7-5. (a) Schematic of local time-frequency region of spectrogram containing two voiced speakers (blue and red); (b) GCT of (a) showing envelope replicas at distinct locations in the GCT space; detection of ordered components along the same orientation (green ellipses) may be used to used to detect the voiced-on-voiced voicing mixture condition as well as the number of speakers.

In addition to multi-pitch estimation, one outstanding issue in separation is that of mixtures of *unvoiced* speech (i.e., the unvoiced-on-unvoiced speech mixture condition). Due to the ambiguity of assignment in carrier positions to distinct speakers (see Chapter 3 and Chapter 4), our present development heuristically assigns half the amplitude of these local time-frequency regions to each speaker. A more sophisticated approach is to utilize the distributive nature of envelope replicas in the GCT domain (Figure 7-6). For instance, spectral matching approaches performed in the GCT domain could conceivably group envelope components of similar shape and orientation located at distinct locations to *distinct* speakers. An outstanding challenge in this approach, however, is the assignment of these groupings to individual speakers' pitch tracks for estimating their corresponding spectrograms.

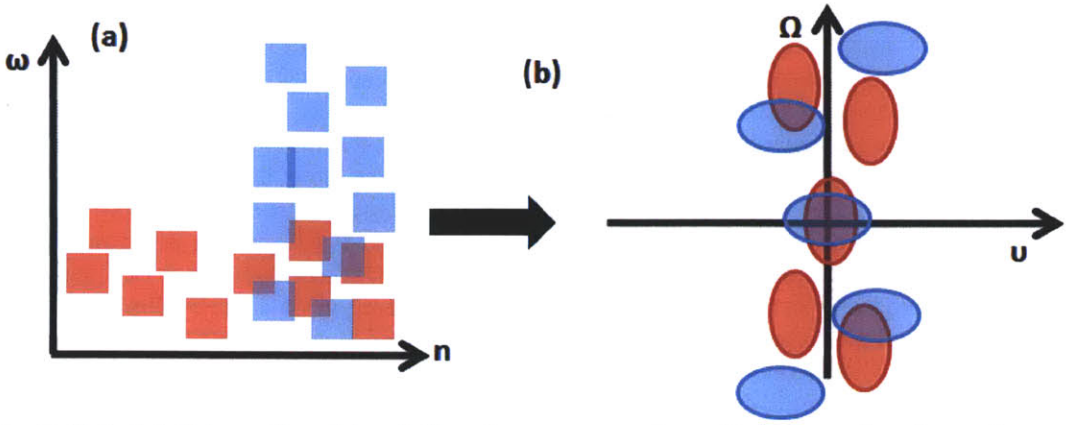

Figure 7-6. (a) Schematic of local time-frequency region of a narrowband spectrogram consisting of two unvoiced speakers, one with spectral shaping along ω (red) corresponding to e.g., a fricative and another along time (red) corresponding to a voiceless stop onset; (b) GCT of (a) showing distribution of envelope content along Ω. Observe that the envelope structures will be distinct in their orientation, which may be used for grouping of envelope components to distinct speakers.

7.3 New Directions

In this thesis, we have related the two-dimensional (2-D) GCT representation to fundamental properties of speech such as pitch, pitch dynamics and formant structure; in addition, we have briefly explored the mapping of noise content in the GCT domain. Consequently, the GCT framework has potential application in a variety of other speech signal processing tasks in addition to co-channel speaker separation as explored in this thesis. In this section, we describe several of these ideas as motivation for future work in speech analysis (Section 7.3.1), speech enhancement (Section 7.3.2), speech modification (Section 7.3.3), and speech parameter estimation (Section 7.3.4).

7.3.1 Speech Analysis Using 2-D Processing

The narrowband and wideband spectrograms are often viewed as the "canonical" time-frequency distributions from which others may be derived. Indeed, alternative distributions for representing speech are often viewed as "mixtures" of narrowband and wideband spectrograms. For instance, the auditory spectrogram is often viewed as being narrowband and wideband in the high and low-frequency regions, respectively. Another example of such a "mixture" is probabilistic latent factor analysis [40] in which the narrowband and wideband spectrograms are more explicitly "mixed" to form an alternative representation. The signal models derived in this thesis in local time-frequency regions could be applied to interpret such "mixtures" of the canonical and narrowband and wideband spectrograms.

Furthermore, the 2-D models in the corresponding GCT space developed in this thesis may also have implications in interpreting other 2-D processing approaches, such as the 2-D auditory model proposed in [42]. Indeed, preliminary work in [16] showed that the auditory representation did not provide a "coherent" representation of pitch and formant content as in the GCT representation. Models developed in this thesis could help in better interpreting these representations, particularly in relation to the multi-resolution property shown in Figures 7-1 through 7-4.

7.3.2 Speech Enhancement

Based on our understanding of where noise maps in the GCT space (e.g., Figure 3-8), one potential application of the GCT is speech enhancement. Consider for instance, the narrowband GCT representation of speech with additive interfering noise. Assuming again the linearity of the spectrogram magnitudes of speech and noise components in a mixture, analyzing a local time-frequency region according to the set of models results in an overlap of noise and speech content in the time-frequency space. Nonetheless, in the GCT space, the distribution of noise content along the Ω-axis may be exploited for enhancement. In particular, demodulation may be performed as in the co-channel speaker separation case to estimate the near-DC regions of the GCT containing the envelope of the speech component (Figure 7-7).

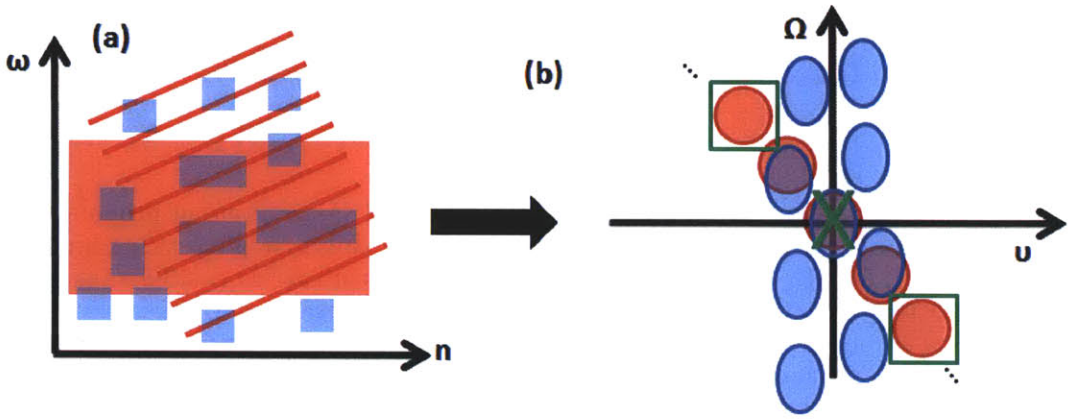

Figure 7-7. (a) Local time-frequency region of narrowband spectrogram computed for a single voiced speaker (red) with additive noise (blue); (b) GCT representation of (a) indicating overlap of components near GCT origin and also at carrier positions. Removal of near-DC components (green 'x') followed by demodulation of non-overlapped envelope components of voiced speech (green rectangle) could lead to enhancement.

7.3.3 Speech Modification

As emphasized throughout this thesis, the GCT-based signal models are based on an explicit mapping of basic speech parameters such as pitch, pitch dynamics, and formant structure to the GCT space. Consequently, another application of the models is in speech/voice modification. As a simple example, consider in Figure 7-8a and b the narrowband GCT representation of voiced speech. To modify the local time-frequency region to reflect an alternative desired/target pitch value and pitch dynamics, a simple algorithm could be based on first removing the replicas of the envelope content at the original carrier positions followed by multiplying the envelope component by a new set of carriers. Alternatively, in the wideband case, a simple algorithm for modifying the *bandwidth* of voiced speech could be to change the peak weights of carrier components as shown in Figure 7-8c and d.

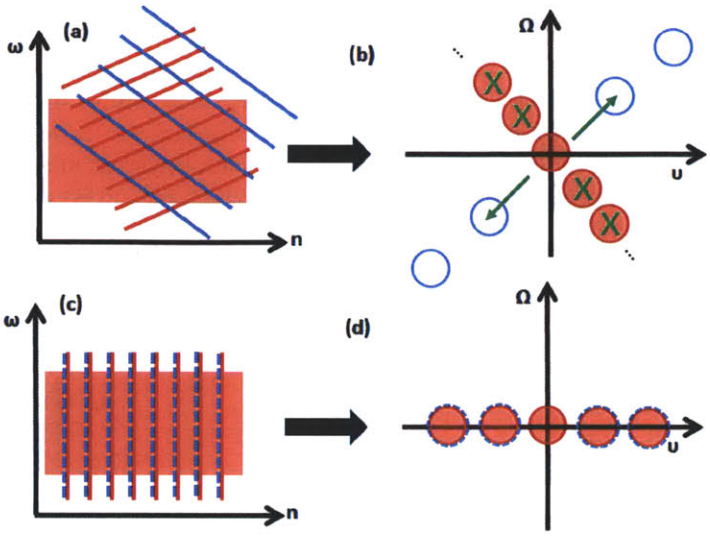

Figure 7-8. (a) Local time-frequency region of narrowband spectrogram of original voiced speech (red) and desired/target harmonic structure reflecting new pitch value and pitch dynamics (blue); (b) GCT of (a) showing removal of original carrier positions (green 'x') and desired carrier locations (blue hollow) of envelope; (c) Local time-frequency region of wideband spectrogram of original voiced speech (red) and desired/target temporal grating pattern reflecting formant bandwidth (blue dashed); (d) GCT of (c) showing envelope components at GCT origin and original carrier positions; modification of coefficient weights along υ-axis to change bandwidth of speech (blue dashed).

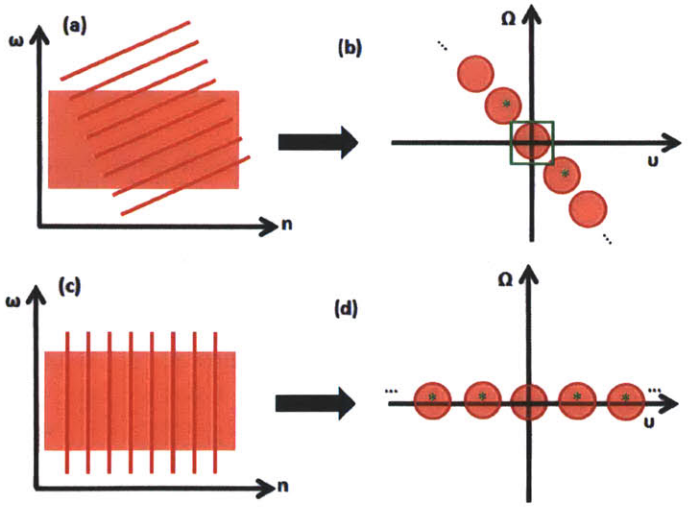

Figure 7-9. (a) Localized region of time-frequency region of narrowband spectrogram of voiced speech; (b) GCT of (a) with isolation of near-DC region corresponding to envelope (i.e., formant) (green rectangle) and peaks (green '*') for estimation of pitch and pitch dynamic information; (c) localized region of time-frequency region of wideband spectrogram of voiced speech; (d) GCT of (a) with peaks of multiple carrier positions to extract formant bandwidth content.

7.3.4 Speech Parameter Estimation

An alternative application to exploiting the GCT's representation of speech parameters is in their analysis and estimation. Indeed, in previous work [19], we showed that the narrowband GCT could be used to obtain improved formant frequency estimates of speech, particularly in conditions of high-pitch source signals. Specifically, envelope components at the GCT origin were separated from effects of the underlying pitch in based on both pitch values *and* pitch dynamics. Building on this work, we may also perform estimation of dynamic formant content using a simple approach of low-pass filtering in the GCT domain as shown in Figure 7-9b. Previous discussion in this thesis has also motivated estimation of pitch and pitch dynamic content from the narrowband GCT for single speakers based on location and orientation of carrier components in the GCT domain (see Section 5.2.1) (Figure 7-9b). Finally, estimation of speech parameters can also be performed using the wideband GCT in relation to formant bandwidth and pitch (Figure 7-9d).

Appendix A

Adaptive 2-D Processing of Speech

A.1 Introduction

In all of our described efforts using the Grating Compression Transform (GCT), we have assumed a uniform tiling (Figure 2-4) of the GCT space as a function of the frequency and time widths used in local region analysis. Region sizes were chosen based on analytical and/or empirical observations and held constant in processing such as analysis/synthesis of spectrograms. Herein we discuss an extension to our framework in *adaptive* GCT processing of speech. In particular, we aim to illustrate the feasibility and effect of *adaptively* modifying the local region sizes for GCT analysis as inspired by evidence for adaptation of receptive field tuning observed in mammalian behavioral tasks [49].

The overall general 2-D framework with adaptation is shown in Figure A-1. More specifically, we develop a method to adaptively "grow" local regions for the GCT based on a quantitative metric that assesses the relative "salience" of the proposed signal model in each localized region; this adaptation therefore allows for distinct resolutions of the GCT analysis based on the signal analyzed. In our analyses, we consider exclusively the narrowband GCT representation; nonetheless, due to the sinusoidal-series modulation interpretation of both narrow and wideband GCTs, similar principles may be applied to the wideband representation in future work.

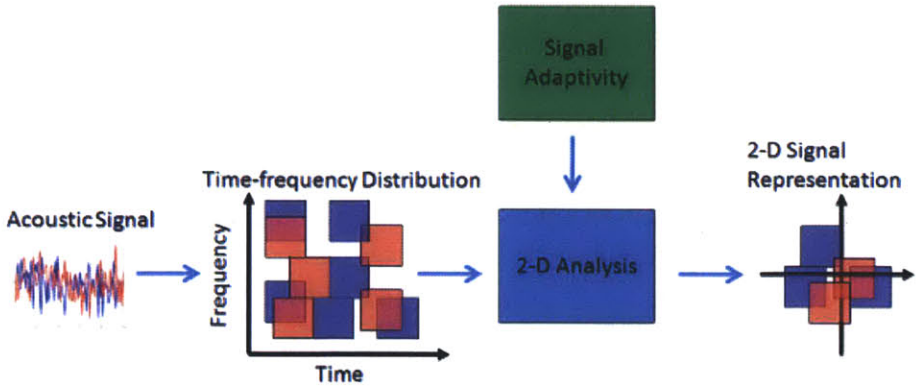

Figure A-1. 2-D Processing framework with time-frequency distribution, 2-D signal representation, and signal adaptivity.

A.2 Motivation

As motivation for adaptation, consider four synthetic signals shown in Figure A-2: Synth1 - a pulse train with rising pitch (150 to 200 Hz), Synth2 - a pulse train with fixed pitch of 200 Hz, both exciting a formant structure, Synth3 - Gaussian white noise, and Synth4 – a single impulse. The formant structure contains formant frequencies (bandwidths) of 669, 2349, 2972, 3500 Hz (65, 90, 156, 200 Hz). We perform analysis/synthesis on these waveforms using a method similar

to that described in Section 3.2 in spectrogram reconstruction. As a distinction in methodology, carrier positions are obtained using peak-picking of GCTs computed on local regions of the log spectrogram. As described in Section 5.3, the log operator has an effect of "flattening" the spectral envelope; as will subsequently discussed, this effect is used to derive a quantitative metric used in adaptation. Local region sizes were varied in time ranging from 20 to 50 ms in 2-ms steps while in frequency from 625 Hz to 1000 Hz in 62.5-Hz steps. As a quantitative metric for comparison, we compute the global signal-to-noise ratio (SNR) between the original and resynthesized waveforms.

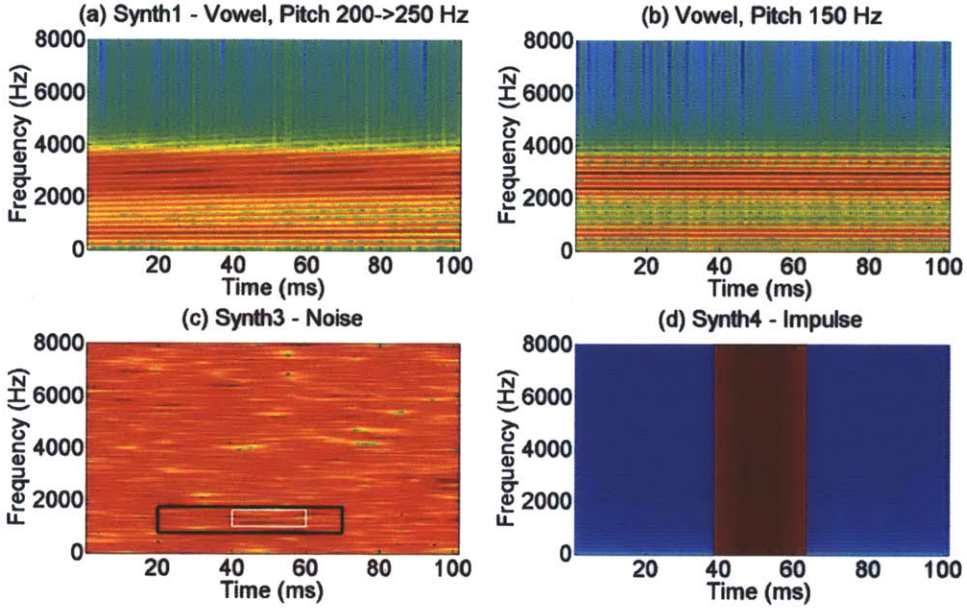

Figure A-2. Spectrograms of Synth1 through 4; (a) Synth1 - vowel with rising pitch; (b) Synth2 - vowel with fixed pitch; (c) Synth - noise; (d) Synth4 – single impulse; white and black boxes denote the extremal local region sizes; log spectrograms shown for display purposes.

Table A-1. Signal-to-noise ratios (dB) using distinct fixed region sizes (time - ms by frequency - Hz) with series-based analysis/synthesis. Optimal region sizes for each signal are indicated in bold along the diagonal.

Size/Signal	Synth1	Synth2	Synth3	Synth4
20 ms by 625 Hz	**6.95**	4.66	7.40	16.37
42 ms by 625 Hz	6.60	**4.76**	6.02	3.86
20 ms by 687.5 Hz	6.60	4.55	**7.45**	16.78
20 ms by 812.5 Hz	6.58	4.31	6.91	**18.86**

Table A-1 lists SNR values for all four signals across four distinct regions sizes. These sizes correspond to those that *maximized* the SNR for each distinct signal such that the diagonal of the table are the maximal SNR values in resynthesis. Observe from these results that four distinct sizes are obtained for each signal. Furthermore, a sub-optimal selection of the region size for distinct signals can result in substantial SNR degradations (e.g., the optimal size for Synth2 leads to an SNR of 3.86 dB for Synth4, ~15 dB below the optimal size for Synth4). These results

demonstrate that a "fixed tiling" of 2-D space exhibits limitations in reconstruction based on distinct properties of the signal itself.

A.3 Adaptation Algorithm and Analysis/Synthesis

The previous discussion motivates an *adaptive* "tiling" of the 2-D analysis space based on properties of the signal and region analyzed. Analogous techniques have been developed in the adaptive short-time analysis literature and typically compute a local metric indicating a measure of stationarity (e.g., "spectral kurtosis" [50]). An alternative metric based on the "relative salience" of 2-D carrier frequencies in the GCT with respect to the rest of the GCT content. We reason that this metric quantitatively assesses the extent to which the series-based 2-D amplitude model is valid for a given region. We then use this metric to guide an adaptive region-growing and selection method. In preliminary analyses, we observed that an analogous 2-D spectral kurtosis did not result in improvements in adaptive processing relative to the proposed "salience ratio".

Let $s_{log}[n,m]$ ($s_{log-hp}[n,m]$) denote a local region of the narrowband log-spectrogram (high-pass filtered) for a given signal such that its corresponding GCT is $S_{log}(\omega,\Omega)$ ($S_{log-hp}(\omega,\Omega)$). The dominant peak of the $S_{log-hp}(\omega,\Omega)$ ($|S_{log-hp}(\omega_d,\Omega_d)|$) magnitude is used to derive the carrier parameters ω_s, θ, and ψ_1 of a 2-D sinusoid denoted as $c_1[n,m]$. The resulting sinusoidal carrier $c_1[n,m]$ is scaled such that its GCT magnitude $|C_1(\omega,\Omega)|$ has a dominant peak value of $|S_{log-hp}(\omega_d,\Omega_d)|$. Additional carriers (for both voiced and unvoiced) speech are obtained by scaling the carrier parameters as in Chapter 3. We define a "salience ratio" (SR) as the ratio of following energies

$$SR = 10log_{10}\frac{E_c}{\iint_{\omega,\Omega}|S_{log}(\omega,\Omega)|^2 d\omega d\Omega - E_c} \quad \text{(A.1)}$$

$$E_c = \iint_{\omega,\Omega}\sum_{k=1}^{N}|c_k(\omega,\Omega)|^2\, d\omega d\Omega \quad \text{(A.2)}$$

E_c is the energy difference in the local region of the *original* (non-filtered) narrowband spectrogram and the carriers and N is the number of carriers. This metric relates the relative energy contributions of the carrier positions in the signal model to the overall region analyzed.

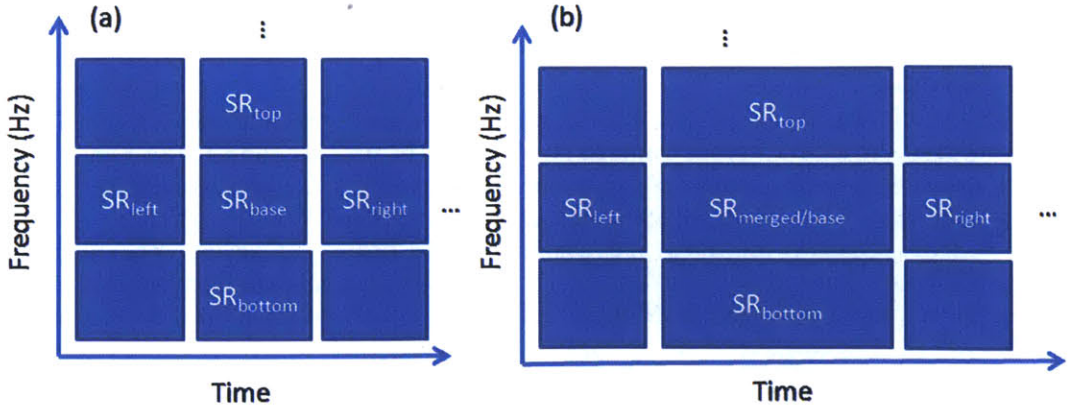

Figure A-3. (a) Base tiling showing base region and its neighbors; (b) Schematic of base region grown from (a) and its new neighbors.

To adapt and select region sizes based on SR, we first perform GCT analysis across the spectrogram of the signal analyzed using a fixed region size with a modified 2-D Hamming

window that satisfies the constant overlap-add property. We refer to this as the *base* tiling, and in each region, we compute the SR metric. The result of this initial analysis is a 2-D grid of SR values (Figure A-3a). Each base region is grown by examining its neighbors' SR values. Denote SR_{base} as the SR value of a base region with neighboring SR_{top}, SR_{bottom}, SR_{left}, SR_{right} as shown in Figure 4. Furthermore, denote $SR_{merged,neighbor}$ as the SR value computed using the combined windows of the base and one if its neighboring regions. The base region is then recursively merged with its neighbors using the following algorithm:

A1) Compute $SR_{merged,neighbor}$ of the base region for all of its neighbors (top, bottom, left, right)
A2) Determine the *maximum* of the four SR values computed in A1) (denoted as $SR_{merged,max}$) with its corresponding neighbor *max_neighbor.*
A3) If $SR_{merged,max} < \max(SR_{base}, SR_{top}, SR_{bottom}, SR_{left}, SR_{right})$,
terminate the algorithm by creating a new region SR_{merged} equal to the base region. Otherwise, merge base region with *max_neighbor* to form SR_{merged} with corresponding SR value $SR_{merged,max}$. Determine the new neighbors of SR_{merged} by its four edges. Use SR_{merge} as the *base* region in A1) to complete the recursion.

The algorithm iteratively grows each base region until the SR value of any resulting merged region is less than that of the unmerged region. The order of the base regions merged is based ordering the SR values of all base regions in descending order. In case the neighbor of a base region has already been incorporated into a previously merged region, it is *excluded* from the SR computations and comparison in the algorithm. Note that the neighbors of any region are strictly those along its four vertical edges such that only rectangular regions are grown (Figure A-3b).

After all base regions have been processed by the algorithm, the resulting set of *merged regions* is used in 2-D demodulation for analysis/synthesis. The 2-D Hamming windows in each merged region are summed and used to extract the appropriate coordinates of the spectrogram and demodulated individually as described in Section 2. The results are summed across all merged regions to reconstruct the spectrogram; the constant-overlap-add property of the 2-D Hamming windows guarantees a unity system if demodulation is not performed. The reconstructed spectrogram is combined with the phase of the original spectrogram and inverted for waveform reconstruction.

A.4 Evaluation and Results

A.4.1 Specific Methods

To assess the utility of the proposed methods, we performed analysis/synthesis using the GCT on 32 sentences of the TIMIT corpus sampled at 16 kHz. The data set consisted of 8 males and 8 females speaking 2 distinct sentences each. Short-time and GCT analyses were performed as in Chapter 3 but using a base tiling (fixed region size) region size of 20 ms by 625 Hz. We use the proposed series model with fixed and adaptive region sizes As quantitative metrics for comparison, we computed 1) global and 2) normalized segmental (25-ms non-overlapping) signal-to-noise ratios (SNR) and 3) PESQ scores of the synthesized waveforms in relation to the original.

A.4.2 Results

Figure A-4 shows the original and reconstructed spectrograms of a TIMIT utterance using the series-based reconstruction with fixed region size and that from the adaptive method. For display/comparison purposes, we plot the *log* spectrograms in these figures. In addition, we show the *non*-base tilings from region growing and a single base tiling (20 ms by 625 Hz) for

comparison purposes. Table A-2 lists average metrics across all utterances. The enforcement of harmonic structure in unvoiced regions presumably causes poorer performance relative to the analysis/synthesis method presented in Chapter 3. We did not perceive a difference of the waveforms from fixed and adaptive series methods in informal listening. Nonetheless, the ability of the adaptive method in providing sdistinct tilings (Figure A-4) of the 2-D analysis space with a modest SNR gain motivates future work in exploring optimal region-growing methods and/or alternative salience metrics to further improve analysis/synthesis. .

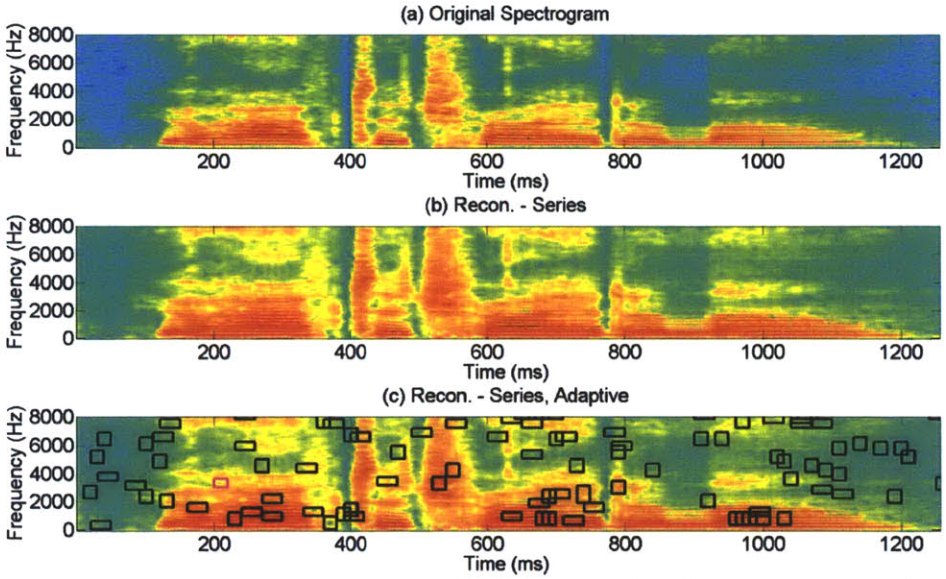

Figure A-4. (a) Original spectrogram of utterance "You'll have to try it alone"; spectrogram plotted on log scale; (b) reconstruction of spectrogram of utterance in (a) using series method (fixed-size) and (c) adaptive regions; black tilings denote *non*-base tilings from region growing (base tilings, 20 ms by 625 Hz, are excluded), a single base tiling (red, 210 ms, 3000 Hz) is shown for comparison purposes. Spectrograms plotted on log scale.

Table A-2. Average global and segmental SNRs and PESQ values.

	PESQ	Global SNR (dB)	Segmental SNR (dB)
Series (fixed)	3.69	9.67	10.14
Series (adaptive)	3.70	9.78	10.19

A.5 Conclusions

This section has developed an auditory-inspired method for adaptively selecting region sizes using a region growing method and a local salience metric for GCT processing. Although we obtained only minor gains in SNR for analysis/synthesis, we anticipate these tilings may provide more benefit in applications such as speech enhancement, analogous to the one-dimensional adaptive scheme presented in [50]. A limitation of the current method is that region growing is done in a greedy fashion; specifically, the choice of region growth is based only on an increase in the salience ratio for a local region and its neighbors without consideration of subsequent iterations of the region growing method. Future work will explore approaches to optimally select region sizes with the aim of further improving performance in analysis/synthesis.

Appendix B

Sinusoidal-based Speaker Separation System

In this thesis, we have used as a frame-based signal representation in contrast to a 2-D approach the sinusoidal system (sinusoidal-based) in the co-channel speaker separation task. Herein we briefly describe modifications to the original separation system proposed and described in [23] as used in this work. We consider a mixture of two speakers resulting in three distinct voicing state mixtures (as in Chapter 3): voiced-on-voiced, voiced-on-unvoiced, and unvoiced-on-unvoiced. The three primary modifications are 1) removal of matched sinusoidal frequencies, 2) diagonal matrix loading of the least-squares formulation, and 2) handling of voicing and unvoiced speech regions. As all other steps are identical to the original method, we refer the reader to [23] for further details and briefly summarize here the basic system and its extensions.

B.1 Basic System

The basic sinusoidal-based separation system is a *frame-based*/short-time technique in which short-time segments of the mixture waveform are processed. Two *voiced* signals were assumed to be present within each short-time segment such that sinusoidal parameters can be estimated for resynthesis of individual speakers' waveforms; the system requires prior knowledge of the pitch tracks of individual speakers for signal separation [23]. In particular, the Fourier transform of a short-time segment of the mixture waveform is assumed to correspond to a *sum* of harmonic components arising from two sets of sinusoids, where each set corresponds to an individual speaker; the frequency positions of the sinusoids in the Fourier transform are based on a simple mapping of pitch values to harmonic locations. In Figure B-1, we schematically illustrate the spectrum of a mixture consisting of two speakers. For separation, prior pitch knowledge is used to map the locations of these sinusoids in the mixture spectrum. Subsequently, a least-squared-error (LSE) fit is performed to solve for the sinusoid amplitudes of individual speakers. This formulation results in a matrix equation that we denote for conciseness as $H\boldsymbol{b} = \boldsymbol{x}$, where H is a symmetric matrix obtained in the LSE formulation, $\boldsymbol{b}$ is a vector of the sinusoidal amplitudes, and $\boldsymbol{x}$ is a vector of the mixture amplitudes. Inverting H gives a solution to $\boldsymbol{b}$, i.e., $\boldsymbol{b} = H^{-1}\boldsymbol{x}$, thereby resulting in two spectral estimates of the individual speakers. The inverse Fourier transform is then applied to the individual spectra to result in waveform estimates of the individual speakers.

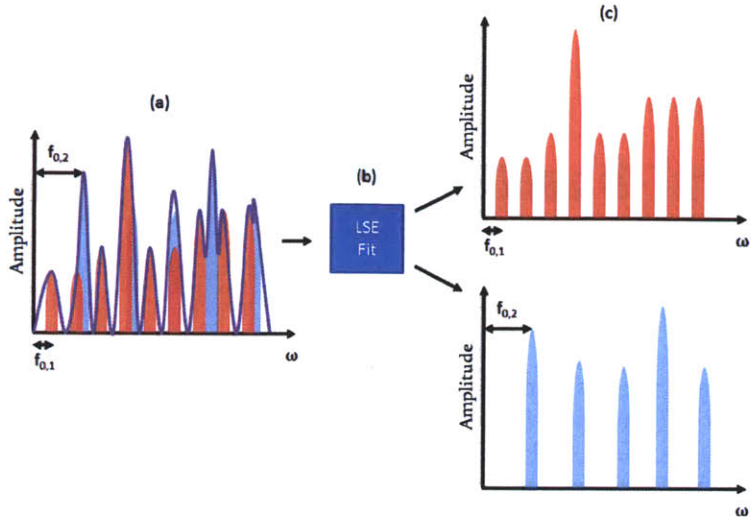

Figure B-1. Schematic illustrating sinusoidal-based separation for mixture of two voices in which the (a) spectrum of a mixture (maroon) of two individual speakers (red and blue) with distinct pitch values $f_{0,1}$ and $f_{0,2}$; (b) least-squares fitting to obtain sinusoidal amplitudes results in (c) spectrum estimates of the individual speakers.

B.2 Extensions

B.2.1 Diagonal Loading the Least-Squared Error Matrix

In the case of two voiced speakers, the sinusoidal-based approach matches that described in [23] and results in a least-squared error (LSE) formulation to solve for sinusoidal parameters. In an extension to this method, we perform two additional steps to the original algorithm to account for potential singularities in the resulting least-squares matrix. First, if two single sinusoidal parameters are matched between two speakers exactly in the resulting Fourier transform domain, the pair is removed. If all harmonics are removed between two speakers (i.e., when the pitch values of the two speakers are identical), the least squares formulation is abandoned in favor of interpolation as described in [23]. After the "harmonic match removal" procedure, the resulting matrix H [23] is checked for near singularities and diagonal loaded using the method identical to that applied in the GCT-based separation approach; we refer the reader to Section 3.4 for details of this method. The threshold λ is swept from 10^1 through 10^5 as in that section on a development set of speech mixtures; the "best" value as based on average global SNR on this development set was found to be 10^1. The latter step seeks to account for conditions when the pitch values are "close" but not equal.

B.2.2 Accounting for Unvoiced Speech in Speech Mixtures

The basic sinusoidal-based separation system does not account for the presence of *unvoiced* speech in a mixtures of speakers. In conditions of the mixture of voiced and unvoiced speech, we therefore develop an extension in which the sinusoidal-based system is solved for a *single* speaker using the harmonic parameters obtained from the pitch track of the voiced speaker. Singularity checking and diagonal loading of the resulting H matrix is also done in this condition as in the voiced-on-voiced case; in this case, singularities may arise due to near-zero values of the frequency positions of the sinusoids in the short-time spectrum. The resulting short-time waveform estimate is used as the estimate of the voiced speaker and the *difference* (i.e., the

residual) between the original mixture waveform and this estimate is used as that for the unvoiced speaker. Figure B-2 shows a schematic of this procedure. Finally, in the unvoiced-on-unvoiced case the mixture waveform is halved due to the ambiguity of assignment of sinusoids.

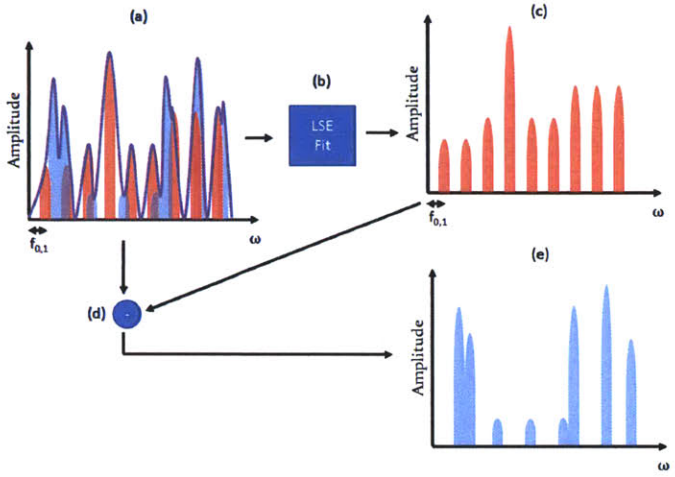

Figure B-2. (a) Mixture spectrum (maroon) consisting of a voiced (red) speaker with pitch $f_{0,1}$ and unvoiced speaker (blue); (b) LSE fit to obtain estimate of (c) voiced speaker; (d) subtraction of voiced estimate from original mixture spectrum to obtain (e) estimate of unvoiced speaker.

Appendix C

Two-dimensional Signal Processing Properties

In this thesis, we have utilized extensively properties of two-dimensional (2-D) signal processing to characterize speech in the time-frequency space and corresponding GCT domain. Herein we briefly summarize several key properties as they relate to those referred to in our derivations and development of the modulation model. We refer the reader to [29] for details of their derivation and emphasize here the results pictorially.

C.1 Harmonic Lines

In our analysis of spectrograms, we have invoked use of a sinusoidal series to model harmonic structure in voiced speech. For illustrative purposes, we consider a single sinusoid present in the time-frequency space oriented as having oscillations across frequency as in the narrowband spectrogram (Figure C-1a). The periodicity ω_s of the sinusoid across frequency will be *inversely* related to the spacing Ω_s (or "frequency") of the resulting impulses in the 2-D Fourier transform (i.e., GCT domain). Extending the sinusoid now to a sinusoidal *series* in Figure C-1b, note that the number of harmonics in the resulting GCT domain will *decrease* with *increasing* ω_s. In relation to pitch in the narrowband representation, the GCT therefore will have more harmonics for *higher* pitch values. As the "dual" to the narrowband case, consider from Figure C-2 a sinusoid oriented such that oscillations occur across *time* as in with the wideband spectrogram computed for voiced speech. The inverse relation of the spacing between harmonic lines and the "frequency" of harmonic components in the GCT domain still holds. Consequently, as argued in the wideband representation, the GCT will exhibit more harmonics for *lower* pitch values.

C.2 Bandwidth of Envelope Content

In addition to harmonic content, our analysis of spectrograms also utilized generalized "bandlimited" (in the GCT domain) functions to represent the envelope content i.e., corresponding to formant structure and onset/offsets. As a simple example, consider a 2-D rectangle located in the time-frequency space. As an initial condition, we illustrate in Figure C-3a a rectangle with time width Δn longer than frequency width denoted by $\Delta\omega$. The corresponding GCT of this rectangle will have bandwidth along the v-axis (Ω-axis) denoted by Δv ($\Delta\Omega$) that is inversely related to Δn ($\Delta\omega$). Observe that increases in $\Delta\omega$ (Δn) result in decreases in $\Delta\Omega$ (Δv) (Figure C-3 and Figure C-4). These arguments can be used to characterize the expansion/contraction of envelope content in the GCT space as was done in particular for formant structure (see Chapter 3) and onset/offset content.

C.3 Rotations

Consider next rotating either the single sinusoid or rectangle by an angle θ as was used to model pitch dynamics in the narrowband spectrogram (Figure C-5a). For the sinusoid, the spacing between harmonic lines ω_s is now along an axis coinciding with θ while the vertical and horizontal distances between the harmonic lines is $\omega_s \cos(\theta)$ and $\omega_s \sin(\theta)$, respectively. This

rotation in the time-frequency space corresponds to a rotation of the impulses in the GCT domain such that they are θ away from the Ω-axis. For the rectangle (Figure C-5b), this corresponds to a similar rotation such that the "widths" of the function in time-frequency space as well as the bandwidths in the GCT space become functions of θ. Similarly, the rotation of the sinusoidal content was used to argue for its relation to pitch dynamics in the narrowband representation. Furthermore, in both the narrow and wideband representations, rotation of a rectangle representing formant structure was used to incorporate formant dynamics into the model.

C.4 Modulation

Finally, we illustrate in Figure C-6 the concept of *modulation* in two dimensions. Specifically, a grating pattern-like *carrier* component (e.g., a sinusoid resting on a DC pedestal) is *modulated* (i.e., multiplied) by a slowly varying envelope component resulting in the modulation model. In the corresponding 2-D Fourier transform space (e.g., the GCT), the carrier component exhibits an impulse at the origin reflecting the DC value and two impulses reflecting the spatial frequency and orientation of the 2-D sinusoid. The slowly varying envelope component is mapped to a component near the origin in the 2-D Fourier space. The 2-D Fourier transform of the modulation product exhibits *replicas* of the Fourier transform of the envelope at the positions of the carrier.

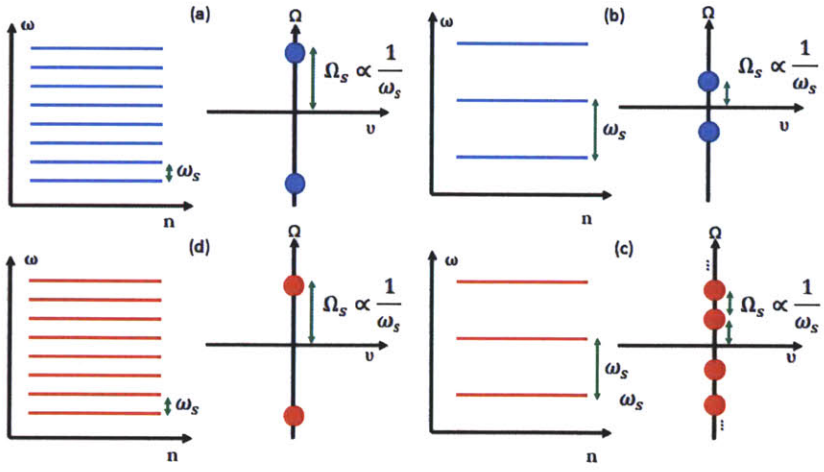

Figure C-1. A *single* sinusoid schematized in blue oriented across ω_s with (relatively) (a) small and (b) large ω_s values and their corresponding GCT representation; observe the inverse relation between ω_s and Ω_s; (c – d) illustrate a sinusoidal series schematized in red with the corresponding GCT representation showing fewer harmonic components for small ω_s versus large ω_s.

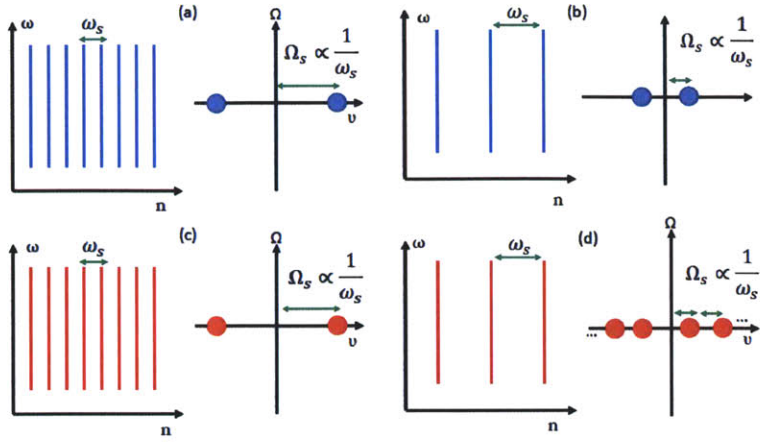

Figure C-2. A sinusoid schematized in blue oriented across n with (relatively) (a) small and (b) large ω_s values and their corresponding GCT representation; note the inverse relation between ω_s and Ω_s; (c – d) illustrate a sinusoidal series schematized in red with the corresponding GCT representation showing fewer harmonic components for small ω_s versus large ω_s.

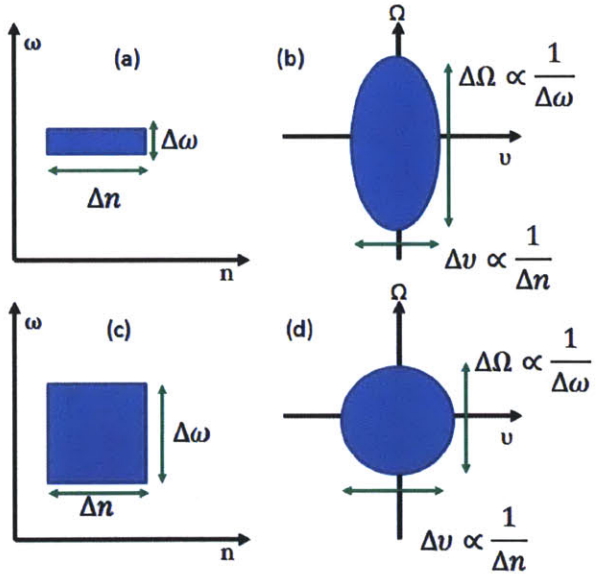

Figure C-3. (a) Schematic illustrating rectangle with longer time duration than frequency duration; (b) GCT of (a) with corresponding inverse relation in bandwidth along υ and Ω-axes with respect to Δn and $\Delta\omega$, respectively; (c) expansion of (a) along ω and (d) corresponding *reduction* in $\Delta\Omega$.

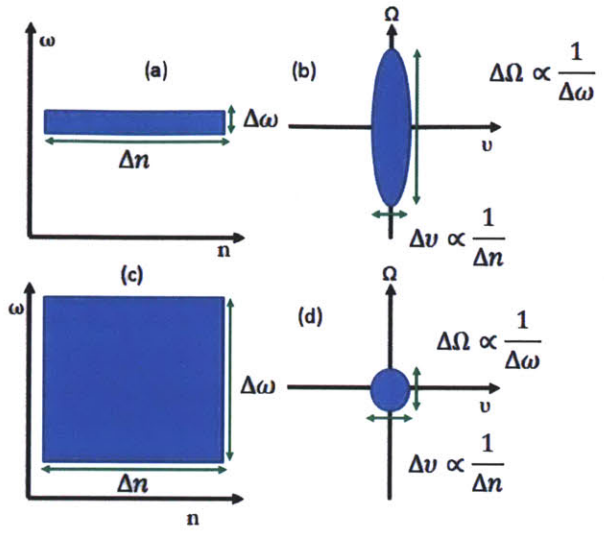

Figure C-4. (a) Relative to Figure C-3a, an expansion along n and (b) corresponding *reduction* in $\Delta\upsilon$; (b) expansion relative to Figure C-3a along *both* n and ω resulting in reduction of bandwidth in (d) for both $\Delta\upsilon$ and $\Delta\Omega$.

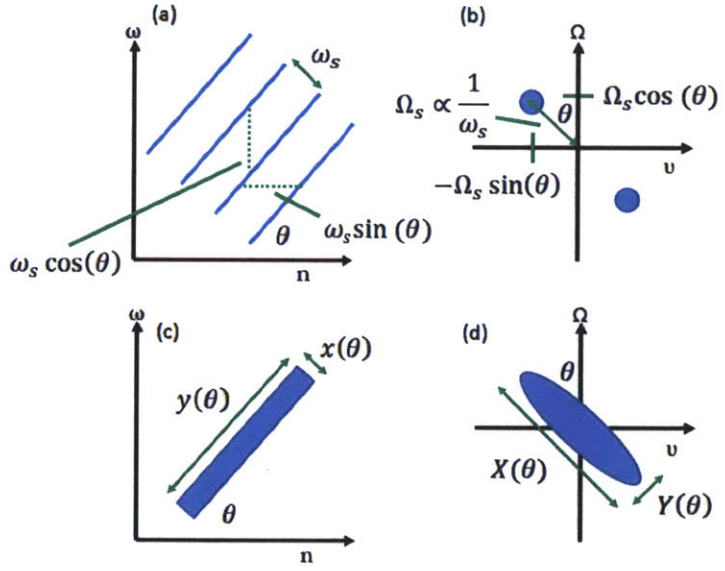

Figure C-5. (a) Rotation of single sinusoidal by angle θ such that ω_s is oriented along θ and vertical and horizontal distances between sinusoid peaks are also functions of θ; (b) GCT of (a) illustrating corresponding rotation of impulses away from the Ω-axis; (c) rotation of a rectangle by θ and corresponding GCT with rotation; "bandwidths" of the function are no rotated by θ in the GCT as well.

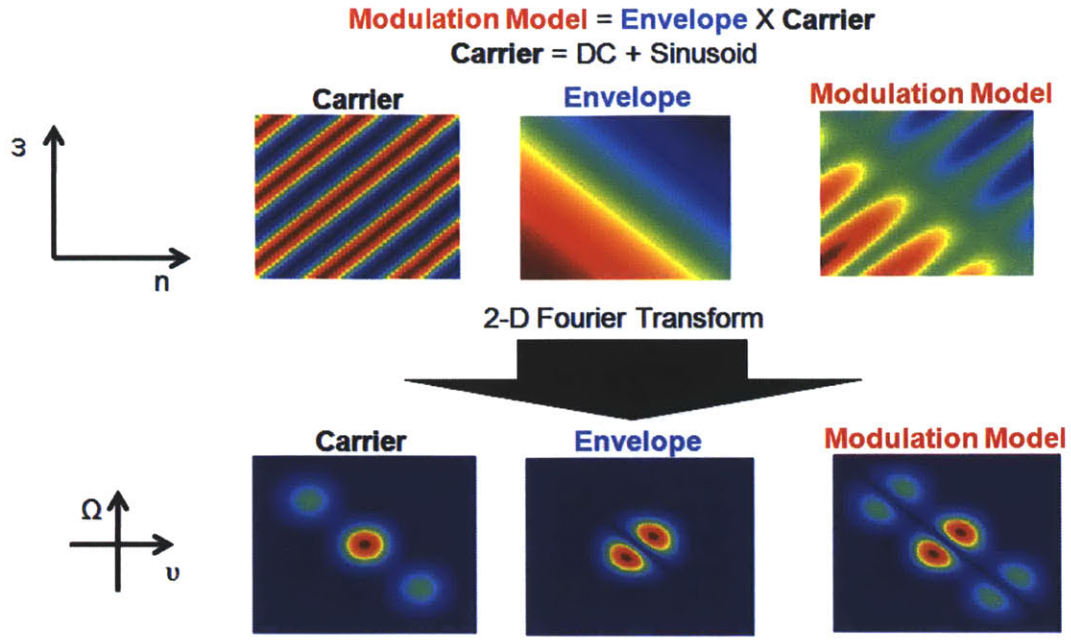

Figure C-6. Simulations illustrating concept of *modulation* in two dimensions; the modulation model is the product of a grating pattern e.g., a sinusoid resting on a DC pedestal multiplied or *modulated* by a slowly varying envelope structure (here, a slowly-varying sinusoid with no DC component). In the 2-D Fourier space, the carrier consists of an impulse at the origin reflecting the DC pedestal and two peaks reflecting the spatial frequency and orientation of the sinusoid; the 2-D Fourier transform of the envelope is *replicated* at the locations of the peaks corresponding to the carrier due to modulation.

Appendix D

Computational Complexity

In this thesis, we have developed algorithms using two-dimensional (2-D) processing of speech in local time-frequency regions of the narrowband and wideband spectrograms. Due to the 2-D nature of our framework, the computational complexity of our algorithms will be greater than traditional short-time/frame-based processing of speech. In this section, we describe and assess analytically the complexity of a typical Grating Compression Transform-based (GCT) processing algorithm.

D.1 Complexity as a Function of Number of Discrete-Fourier Transforms

In delineating the 2-D versus short-time/frame-based methods, we view the *primary* computational cost as arising from computing 2-D discrete-Fourier transforms (DFT) across multiple localized regions of the short-time Fourier transform (STFT) magnitude; therefore, we focus on this component in our calculations. Additional steps such as matrix inversions used in least-squared-error (LSE) fitting and search methods (e.g., in peak-picking) are excluded, since such steps may also be used in traditional frame-based processing techniques. For instance, in the sinusoidal-based separation system, matrix inversion is similarly required to perform LSE fits to the short-time spectrum (see Appendix B).

D.2 Analysis Parameters

For completeness, Table D-1 lists all of the parameters used in short-time and GCT-based analysis along with their corresponding abbreviations; as will be subsequently discussed, only a subset of these parameters will be relevant for analysis of complexity. In Table D-1, N_{GCT} refers to the *length* of the DFT applied along a single dimension such that corresponding 2-D DFT is of size N_{GCT}^2. Table D-2 and Table D-3 list the specific parameter values used in processing the waveform for narrowband- and wideband-based GCT analysis. As a reference for measurement, we consider processing a waveform sampled at 16 kHz with duration of 1 second such that $L = 16000$.

Table D-1. Listing of parameters and their abbreviations in GCT analysis of spectrograms.

Parameter	Abbreviation
Time length of signal	L
Time length of short-time analysis window (STFT)	L_w
Frame size (STFT)	dt_{STFT}
DFT length for STFT	N_{STFT}
Frequency width of local time-frequency region	f_{GCT}
Time width of local time-frequency region	t_{GCT}
Step size along frequency in GCT analysis	df_{GCT}
Step size along time in GCT analysis	dt_{GCT}
DFT length of 2-D GCT computation	N_{GCT}

Table D-2. Table of parameters for narrowband-based GCT analysis.

	L	L_w	dt_{STFT}	N_{STFT}	f_{GCT}	t_{GCT}	df_{GCT}	dt_{GCT}	N_{GCT}
Time/Frequency (ms/Hz)	1000	32	1		875	20	~218	5	
Samples	16000	512	16	512	28	20	7	5	512

Table D-3. Table of parameters for wideband-based GCT analysis.

	L	L_w	dt_{STFT}	N_{STFT}	f_{GCT}	t_{GCT}	df_{GCT}	dt_{GCT}	N_{GCT}
Time/Frequency (ms/Hz)	1000	2.5	0.625		500	37.5	125	9.375	
Samples	16000	40	10	512	16	60	4	15	512

D.3 Complexity Analysis

To analyze the complexity of GCT analysis, we begin by first noting that the complexity of a 1-D discrete-Fourier transform (DFT) used to compute the STFT (i.e., the spectrogram) using the standard fast Fourier transform (FFT) algorithms has complexity of $N_{STFT}\log_2(N_{STFT})$ [2]. Furthermore, the 2-D DFT (computed via the FFT) used in computing the GCT has complexity of $N_{GCT}^2 log_2(N_{GCT})$ [29].

To compute the STFT for a waveform of length L samples using a short-time analysis window of length $L_w < N_{STFT}$ and frame rate dt_{STFT}, we must compute N_{slices} 1-D DFTs, where N_{slices} is the number spectral slices and approximately given by

$$N_{slices} \approx \frac{L}{dt_{STFT}}. \tag{D.1}$$

The complexity of the STFT computation is therefore

$$Complexity_{STFT} = O(N_{slices}(N_{STFT}\log_2(N_{STFT}))) \tag{D.2}$$

where $O()$ denotes big-O notation. Rewriting this in terms of the underlying parameters of analysis, we obtain

$$Complexity_{STFT} = O(\frac{L}{dt_{STFT}}(N_{STFT}\log_2(N_{STFT}))). \tag{D.3}$$

Building on this result, the number of 2-D FFT computations required is denoted as $N_{patches}$ corresponding to the total number of local time-frequency regions to be analyzed by the GCT:

$$N_{patches} \triangleq N_{patches(time)}N_{patches(frequency)} \tag{D.4}$$

$$N_{patches(time)} \approx \frac{N_{slices}}{dt_{GCT}} \tag{D.5}$$

$$N_{patches(frequency)} \approx \frac{\frac{N_{STFT}}{2}}{df_{GCT}} \tag{D.6}$$

The complexity of GCT analysis of a spectrogram is

$$Complexity_{GCT} = O(Complexity_{STFT} + N_{patches}(N_{GCT}^2\log_2(N_{GCT}))) \tag{D.7}$$

i.e, the complexity of computing the STFT added to that of GCT analysis across all local time-frequency regions. Rewriting this in terms of the parameters used in analysis, we have

$$Complexity_{GCT} = O\left(\frac{L}{dt_{STFT}}(N_{STFT}\log_2(N_{STFT}))\right. \\ \left. + \left(\frac{N_{STFT}}{2(df_{GCT})}\right)\left(\frac{L}{dt_{GCT}dt_{STFT}}\right)(N_{GCT}^2\log_2(N_{GCT}))\right). \quad \text{(D.8)}$$

Simplifying this, we have that

$$Complexity_{GCT} = O\left(\frac{L(N_{STFT})}{dt_{STFT}}\left(log_2(N_{STFT}) + \frac{N_{GCT}^2\log_2(N_{GCT})}{2(df_{GCT})(dt_{GCT})}\right)\right). \quad \text{(D.9)}$$

Substituting the numerical values of the analysis parameters into (D.3) and (D.9) for the *narrowband* representation, we have that GCT processing of a 1-second waveform sampled at 16 kHz requires 1.73×10^{10} operations relative to 4.61×10^6 operations in computing the STFT alone; this corresponds to a $\frac{1.73\times10^{10}}{4.61\times10^6} \approx 3750$-fold increase in complexity. Similarly in the wideband case, GCT processing requires 1.61×10^{10} operations while the STFT computation requires 7.37×10^6 operations resulting in a $\frac{1.61\times10^{10}}{7.37\times10^6} \approx 2200$-fold increase in complexity.

D.4 Empirical Measurements

To estimate the computation time required in STFT and GCT-based processing, we simulated 1-D and 2-D DFT computations on test signals consisting of Gaussian white noise. Simulations were performed on an Intel 2.93 Gigahertz processor using the MATLAB software package [51]. Computation time was recorded using the 'tic' and 'toc' operations available in MATLAB.

1000 1-D DFT computations of length $N_{STFT} = 512$ are computed on a white noise sequence of 512 samples. Next, we simulated 1000 2-D 512 by 512-point DFT computation of a 2-D white noise sequence of size 512 by 512. We compute the average of the simulations to obtain representative computation times of individual 1-D and 2-D FFT computations. We denote these values as $\bar{c}_{STFT}$ and $\bar{c}_{GCT}$, respectively. Finally, using these values, we estimate the total computation times $c_{total}(STFT)$ and $c_{total}(GCT)$ as

$$c_{total}(STFT) = N_{slices}\bar{c}_{STFT} \quad \text{(D.10)}$$

$$c_{total}(GCT) = c_{total}(STFT) + N_{patches}\bar{c}_{GCT}\,. \quad \text{(D.11)}$$

In our simulations, we obtained values of $\bar{c}_{STFT} = 1.313 \times 10^{-5}$ seconds and $\bar{c}_{GCT} = 9.7 \times 10^{-3}$, respectively. The resulting estimated computation times are shown in Table D-4 based on substituting the parameter values for the narrowband and wideband cases in solving for N_{slices} and $N_{patches}$. Both short-time analysis methods have a total computation time ~100 times less than the 1 second duration of the signal. In contrast, GCT analysis is ~66 to 70 times greater than the duration of the signal. While short-time analysis can be performed in real time, substantially greater reductions in computational complexity are required for GCT processing to be applicable to real-time applications.

One future direction for complexity reduction may be explored in relation to the substantial overlap of the local time-frequency regions in GCT analysis. In both the narrowband and

wideband parameters used in this thesis, dt_{GCT} and df_{GCT} are $\frac{1}{4}$ the size of the entire region. Consequently, in computing the 2-D DFT, samples are repeated along both the time and frequency directions. Recursive FFT algorithms may be incorporated to exploit this redundancy in data to reduce the complexity of the 2-D DFT computations [2]. In addition to complexity reduction, parallel processing techniques may also be used to reduce the run time of GCT-based algorithms.

Table D-4. Estimated computational time of processing a 1-second waveform with STFT and GCT-based processing for wideband and narrowband representation. Units of time are measured in *seconds*.

	Narrowband	Wideband
$c_{total}(STFT)$	1.31×10^{-2}	2.10×10^{-2}
$c_{total}(GCT)$	70.92	66.24

Bibliography

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

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

[3] Makhoul, J., "Linear prediction: a tutorial review." s.l. : Proceedings of the IEEE, 1975, Issue 4, Vol. 63.

[4] Quatieri, T.F. and McAulay, R., "Speech analysis/synthesis based on a sinusoidal representation." s.l. : IEEE Transactions on Acoustics, Speech, and Signal Processing, 1986, Issue 4, Vol. 34.

[5] Greenberg, S. and Kingsbury, B., "The Modulation Spectrogram: In Pursuit of an Invariant Representation of Speech." Munich, Germany : Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997.

[6] Shamma, S., Chi, T. and Ru, P., "Multiresolution Spectrotemporal Analysis of Complex Sounds." s.l. : Journal of Acoustical Society of America, 2005, Vol. 118.

[7] Quatieri, T.F., "2-D Processing of Speech with Application to Pitch Estimation." Denver, CO : Proceedings of the International Conference on Spoken Language Processing, 2002.

[8] Schimmel, S. and Atlas, L., "Coherent envelope detection for modulation filtering of speech." Philadelphia, PA : Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.

[9] S., Sukittanon, Atlas, L. and Filali, K., "Improved Modulation Spectrum through Multi-scale Modulation Frequency Decomposition." Philadelphia : Proceedings of the 2005 IEEE Conference on Acoustics, Speech, and Signal Processing, 2005.

[10] Hermansky, H. and Morgan, N., "RASTA Processing of Speech." s.l. : IEEE Transactions on Speech and Audio Processing, 1994, Issue 4, Vol. 2.

[11] Schimmel, S., Atlas, L. and Nie, K., "Feasibility of single channel speaker separation based on modulation frequency." Honolulu, HI : Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2007.

[12] Wang, K and Shamma, S., "Spectral Shaping in the Central Auditory System." s.l. : IEEE Transactions on Speech and Audio Processing, 1995, Issue 5, Vol. 3.

[13] Jeon, W. and Juang, B., "Speech Analysis in a Model of the Central Auditory System." s.l. : IEEE Transactions on Audio, Speech, and Language Processing, 2007, Issue 6, Vol. 15.

[14] Shamma, S., Slaney, M. and Mesgarani, N., "Discrimination of Speech from Non-speech Based on Multi-scale Spectrotemporal Modulation." s.l. : IEEE Transactions on Audio, Speech, and Language, 2006, Issue 3, Vol. 14.

[15] Elhilali, M. and Shamma, S., "Information-bearing components of speech intelligibility under babble-noise and bandlimiting distortions." Las Vegas, NV : Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2008.

[16] Wang, T.T., *Exploiting Pitch Dynamics for Speech Spectral Estimation Using a Two-Dimensional Processing Framework.* Cambridge, MA : SM Thesis, MIT Department of Electrical Engineering and Computer Science, 2008.

[17] Schutte, K., *Parts-based Models and Local Features for Automatic Speech Recognition.* Cambridge, MA : MIT Department of Electrical Engineering and Computer Science, 2009.

[18] Atlas, L and Janssen, C., "Coherent modulation spectral filtering for single-channel music source separation." Philadelphia, PA : s.n., 2005.

[19] Wang, T.T. and Quatieri, T.F., "High-pitch Formation Estimation by Exploiting Temporal Change of Pitch." s.l. : IEEE Transactions on Audio, Speech, and Language Processing, 2010, Issue 1, Vol. 18.
[20] Wang, T.T. and Quatieri, T.F., "Multi-Pitch Estimation by a Joint 2-D Representation of Pitch and Pitch Dynamics." Makuhari, Japan : roceedings of the 11th Annual Conference of the International Speech Communication Association, 2010.
[21] Wang, T.T. and Quatieri, T.F., "Towards Co-channel Speaker Separation by 2-D Demodulation of Spectrograms." New Paltz, NY : Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009.
[22] Ezzat, T., Bouvrie, J. and Poggio, T., "Localized Spectrotemporal Cepstral Analysis of Speech." Las Vegas, NV : Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.
[23] Quatieri, T.F. and Danisewicz, R., "An Approach to Co-channel Talker Interference Suppression Using a Sinusoidal Model for Speech." s.l. : IEEE Transactions on Acoustics, Speech, and Signal Processing, 1990, Vol. 38.
[24] Lee, T.W., et al., "Blind source separation of more sources than mixtures using overcomplete representations." s.l. : IEEE Signal Processing Letters, 1999, Issue 4, Vol. 6.
[25] Li, Y., Woodruff, J. and Wang, D.L., "Monaural musical sound separation based on pitch and common amplitude modulation." s.l. : IEEE Transactions on Audio, Speech, and Language Processing, 2009, Vol. 17.
[26] Roweis, S., "Factorial Models and Refiltering for Speech Separation & Denoising." Geneva, Switzerland : Proceedings of Eurospeech, 2003.
[27] Vishnubhotla, S and Espy-Wilson, C., "An Algorithm for Speech Segregation of Co-channel Speech." Taipei, ROC : IEEE International Conference on Acoustics, Speech and Signal Processing, 2009.
[28] Stevens, K.N., *Acoustic Phonetics.* Cambridge, MA : MIT Press, 1999.
[29] Lim, J., *Two-Dimensional Signal and Image Processing.* Englewood Cliffs, NJ : Prentice Hall, 1990.
[30] Proakis, J., *Digital Communications.* Singapore : McGraw-Hill Book Co., 1995.
[31] Rice, S., *Mathematical Analysis of Random Noise.* s.l. : Bell Systems Technical Journal, 1954. 24:46-156.
[32] Van Trees, H., *Detection, Estimation, and Modulation Theory, Part I.* New York, NY : Wiley, 1968.
[33] Wu, M., Wang, D.L. and Brown, G., "A Multi-pitch Tracking Algorithm for Noisy Speech." s.l. : IEEE Transactions on Audio, Speech, and Language Processing, 2003, Vol. 11.
[34] Wohlmayr, M., Stark, M and Pernkopf, K., "A Mixture Maximization Approach to Multipitch Tracking with Factorial Hidden Markov Models." Dallas, TX : Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2010.
[35] Rifkin, R. and Lippert, R., *Notes on Reglarized Least Squares.* s.l. : MIT CSAIL Technical Report-2007-023, 2007.
[36] Jancovic, P and Kokuer, M., "Separation of Harmonic and Speech Signals Using Sinusoidal Modeling." New Paltz, NY : Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2007.
[37] Fisher, W., Doddington, G. and Goudie-Marshall, K., "The DARPA Speech Recognition Research Database: Specifications and Status." s.l. : Proceedings of the DARPA Workshop on Speech Recognition, 1986.
[38] , http://www.speech.kth.se/wavesurfer/. [Online]
[39] G., Hu and Wang, D.L., "Monaural speech segregation based on pitch tracking and amplitude modulation." s.l. : IEEE Transactions on Neural Networks, 2004, Vol. 15.

[40] Nam, J, et al., "A Super-Resolution Spectrogram Using Coupled PLCA." Makuhari, Japan : roceedings of the 11th Annual Conference of the International Speech Communication Association, 2010.
[41] Malyska, N and Quatieri, T.F., "Spectral Representations of Non-modal Phonation." s.l. : IEEE Transactions on Audio, Speech, and Language Processing, 2008, Issue 1, Vol. 16.
[42] Chi, T., Ru, P. and Shamma, S., "Multiresolution Spectrotemporal Analysis of Complex Sounds." s.l. : Journal of Acoustical Society of America, 2005, Vol. 118.
[43] Brown, G.J. and Cooke, M., "Computational auditory scene analysis." s.l. : Computer Speech and Language, 1994, Vols. 8: 297-336.
[44] Duda, R., P., Hart. and Stork, D., *Pattern Classification.* New York, NY : John Wiley and Sons, 2001.
[45] Bar-Shalom, Y., Li, X. and Kirubarajan, T., *Estimation with Application to Tracking and Navigation: Theory, Algorithms, and Software.* New York, NY : Wiley, 2001.
[46] Benincasa, D and Savic, M., "Voicing State Determination of Co-channel Speech." Seattle, WA : Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998.
[47] Mahgoub, Y and Dansereau, Y., "Voicing-State Classification of Co-channel Speech Using Nonlinear State-space Modeling." Philadelphia, PA : Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.
[48] Talkin, D., "A Robust Algorithm for Pitch Tracking (RAPT)." s.l. : in Speech Coding and Synthesis, ed. Klein, W. and Paliwal, K. Elsevier Science, 1995.
[49] Fritz, J., et al., "Rapid Task-Related Plasticity of Spectrotemporal Receptive Fields in Primary Auditory Cortex." s.l. : Nature Neuroscience, 2003, Issue (11): 1216-23, Vol. 6.
[50] Rudoy, D., Basu, P. and Wolfe, P., "Superposition Frames for Adaptive Time-Frequency Analysis and Fast Reconstruction." s.l. : IEEE Transactions on Signal Processing, 2010, Issue (5):2581-2596, Vol. 58.
[51] , "MATLAB." s.l. : The Mathworks, Inc. 1984 - 2008.
[52] Wang, T.T. and Quatieri, T.F., "2-D Processing of Speech for Multi-Pitch Analysis." Brighton, UK : Proceedings of the 10th Annual Conference of the International Speech Communication Association, 2009.
[53] Wang, T.T. and Quatieri, T.F., "Two-dimensional Speech Signal Modeling." s.l. : in review, IEEE Transactions on Audio, Speech, and Language Processing.