# HIGH-RESOLUTION SINUSOIDAL ANALYSIS FOR RESOLVING HARMONIC COLLISIONS IN MUSIC AUDIO SIGNAL PROCESSING

BY

ANDREAS F. EHMANN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Associate Professor Mark Hasegawa-Johnson, Chair
Professor Emeritus James Beauchamp
Associate Professor Minh Do
Professor J. Stephen Downie

# ABSTRACT

Many music signals can largely be considered an additive combination of multiple sources, such as musical instruments or voice. If the musical sources are pitched instruments, the spectra they produce are predominantly harmonic, and are thus well suited to an additive sinusoidal model. However, due to resolution limits inherent in time-frequency analyses, when the harmonics of multiple sources occupy equivalent time-frequency regions, their individual properties are additively combined in the time-frequency representation of the mixed signal. Any such time-frequency point in a mixture where multiple harmonics overlap produces a single observation from which the contributions owed to each of the individual harmonics cannot be trivially deduced. These overlaps are referred to as *overlapping partials* or *harmonic collisions*. If one wishes to infer some information about individual sources in music mixtures, the information carried in regions where collided harmonics exist becomes unreliable due to interference from other sources. This interference has ramifications in a variety of music signal processing applications such as multiple fundamental frequency estimation, source separation, and instrumentation identification.

This thesis addresses harmonic collisions in music signal processing applications. As a solution to the harmonic collision problem, a class of signal subspace-based high-resolution sinusoidal parameter estimators is explored. Specifically, the direct matrix pencil method, or equivalently, the Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) method, is used with the goal of producing estimates of the salient parameters of individual harmonics that occupy equivalent time-frequency regions. This estimation method is adapted here to be applicable to time-varying signals such as musical audio. While high-resolution methods have been previously explored in the context of music signal processing, previous work has not addressed whether or not such methods truly produce high-resolution sinusoidal parameter estimates in real-world music audio signals. Therefore, this thesis answers the question of whether high-resolution sinusoidal parameter estimators are really high-resolution for real music signals.

This work directly explores the capabilities of this form of sinusoidal parameter estimation to resolve collided harmonics. The capabilities of this analysis method are also explored in the context of music signal processing applications. Potential benefits of high-resolution sinusoidal analysis are examined in experiments involving multiple fundamental frequency estimation and audio source separation. This work shows that there are indeed benefits to high-resolution sinusoidal analysis in music signal processing applications, especially when compared to methods that produce sinusoidal parameter estimates based on more traditional time-frequency representations. The benefits of this form of sinusoidal analysis are made most evident in multiple fundamental frequency estimation applications, where substantial performance gains are seen. High-resolution analysis in the context of computational auditory scene analysis-based source separation shows similar performance to existing comparable methods.

*To my parents, for their love and support*

# ACKNOWLEDGMENTS

There are many people whom I would like to thank in helping and guiding me along a long road that culminated in this work. I first need to give my thanks to each member of my doctoral committee in turn, as each of them played a role in helping me throughout the doctoral process.

First, Dr. Mark Hasegawa-Johnson was instrumental in helping me organize all of my ideas into a single, cohesive work. I am often in awe of the expansiveness of his knowledge. It seems to me that for any concept he is confronted with, Dr. Hasegawa-Johnson has not only the ability to grasp it, but also the insight and perceptiveness to understand and question the very fine nuances of the idea. I only hope that at some point in my life, I can achieve such a level of knowledge and understanding.

Next, I would like to thank Dr. James Beauchamp. It was under Dr. Beauchamp that I began my graduate studies, and when I look back to my general knowledge in the domain of audio and music processing, and my capabilities in terms of writing and expressing my thoughts, I sometimes wonder if I am the same person as I was back then. I feel that I most certainly am not, and this is largely owed to Dr. Beauchamp. His experience and pure intuition in the field of music audio processing is invaluable.

Quite a few years ago, I took a class entitled "Vector Space Signal Processing." The class was one of the most challenging classes I have ever taken, but also the most rewarding. This class was taught by Dr. Minh Do, and it was here that I first learned about the notion of signal subspace-based techniques in signal processing. It was actually a final project in this class that served as a pilot study to the work presented in this thesis. Therefore, it was under Dr. Do that the seeds of this work were sewn.

Last, but most certainly not least, I owe a great deal of gratitude to Dr. J. Stephen Downie. As a member of the IMIRSEL laboratory at the Graduate School of Library and Information Science, he provided me with support and a place to call home during my time as a graduate student. I gained valuable first-hand experience with him in all aspects of academia, from research, to publication, to grant writing, to late-night symposia. Dr. Downie has become a life-long friend.

Naturally, those closest to me form a foundation for both emotional and intellectual support. My parents and my brother were instrumental during this whole process. My mother, at all turns, encouraged and supported me during my studies, as did my father and brother. Having a father who is an engineering professor is also invaluable, as throughout my life I have been exposed to an environment which fostered any intellectual achievement I might attain. And of course, my partner and the love of my life, Jana Žujović, was there to pick me up when I was down. Having someone so close to you who has a Ph.D in signal processing does not hurt either, as she and I have a shared experience in both our personal and professional lives.

Aside from the members of my committee and those closest to me, there is one individual that I owe a great deal of thanks to. My colleague, my former roommate, and my best friend during my graduate studies, Mert Bay served as my idea sounding board. Having someone work next to you who you can instantly bounce ideas off of, to see if they are sound, is essential to anyone. Aside from Mert, all members of the lab I have worked in over the years have been inspiring, and provided an environment in which I could be content and happy. Perhaps this is why I stuck around far longer than I needed to.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ANOVA | Analysis of Variance |
| DFT | Discrete Fourier Transform |
| ESPRIT | Estimation of Signal Parameters via Rotational Invariance Techniques |
| $f_0$ | Fundamental Frequency |
| MIR | Music Information Retrieval |
| MIREX | Music Information Retrieval Evaluation eXchange |
| MQ | McAulay-Quatieri Analysis |
| STFT | Short-time Fourier Transform |
| STHR | Short-time High Resolution |

# CHAPTER 1

# INTRODUCTION

The content-based analysis of digital multimedia objects is an area that has warranted an ever-increasing amount of attention in recent years. As the number of available digital multimedia objects continues to increase, so does the desire to be able to summarize, categorize, search, and analyze these objects. While humans are very adept at such tasks, the growth in available content has made these tasks intractable if performed manually. Therefore, there is an increasing emphasis on creating techniques that can perform such tasks automatically, by first extracting and inferring salient aspects of the multimedia objects.

One of the most expansive forms of multimedia objects is digital music audio. While the nature of music allows for large variation among pieces, many pieces also share commonality along a number of facets. First, a large proportion of music can be viewed as a mixture of individual sources. These sources usually correspond to individual musical instruments (e.g., the two violins, viola, and cello of a string quartet). Second, many musics are built upon well-established fundamental principles and rules that govern the combination of these individual sources. These principles influence such aspects as the temporal nature of the music (tempo, meter, rhythm), how pitches or notes are distributed (melody, harmony, etc.), and the overall organization of the piece (form or structure), among others. Therefore, a goal in content-based analysis is the automatic extraction of these properties. While some of these aspects can be determined by analyzing a piece as a whole, some properties can potentially be better deduced if reliable information about the individual sources can be inferred (e.g., transcription, instrumentation, etc.).

Music audio signals are often analyzed by producing a time-frequency representation of the signal. Such representations are used because both temporal aspects of music such as rhythm and frequency-dependent aspects such as harmony can be simultaneously captured (at some very basic level). A key challenge in music audio signal processing is that the very principles by which music sources are combined in a piece make inferring information about the constituent sources difficult. Most notably, the rhythm and harmony rules that govern the use of simultaneous pitches from different sources produce a large overlap among the sources' individual time-frequency representations. It is, in fact, these overlaps that cause some musical pitches to sound more "pleasant" than others when played simultaneously. For nontrivial time-frequency points that are common to multiple sources, the corresponding time-frequency points in a music mixture will represent the additive combinations of each of the sources. Due to resolution limits inherent in time-frequency analyses, the contributions owed to each source in the mixture at these overlapped points cannot be trivially deduced from the mixture. If the information found in such points is critical in inferring information about a source, difficulties can arise because the properties at these points have become unreliable due to interference from other sources.

## 1.1 Thesis Goals

Because the nontrivial points in a time-frequency representation of a single, pitched, musical instrument tone can largely be attributed to sinusoidal *partials* or *harmonics* of the fundamental frequency of the tone, overlapped time-frequency points in mixtures of sources are often referred to as *overlapping partials* or *harmonic collisions*. This thesis aims to recover the salient properties of each of the overlapping, or collided, harmonics directly from music mixtures. To achieve this end, this thesis evaluates whether *signal subspace* techniques are capable of resolving the parameters of sinusoids (i.e., harmonics) that are closely spaced in frequency. Many sinusoidal parameter estimators based on signal subspaces are said to have the property of *super resolution*. This sort of high-resolution analysis allows for the estimation of sinusoidal parameters that would otherwise be very difficult from the direct analysis of more traditional frequency representations such as the Fourier transform. Such signal subspace techniques are parametric in that it is required that the signal conform to some underlying model. In this case, the underlying model is a sinusoidal model, which aligns well with the harmonic nature of pitched musical sources. Therefore, if a signal has a strong fit to this underlying model, there is the potential that the parameters of closely spaced sinusoids can be resolved.

Whether or not real-world musical signals closely enough satisfy the underlying model of signal subspace sinusoidal parameter estimation techniques to allow for accurate, high-resolution analysis first requires the design of a signal subspace-based sinusoidal parameter estimator well suited for musical signals. Drawing inspiration from previously established methods, this thesis presents a signal subspace-based sinusoidal estimator suited for time-varying signals. While similar techniques are already established, there has, to date, been no thorough evaluation of the high-resolution properties of such techniques for musical mixtures. Although the accuracy of these high-resolution properties can be assessed directly, this thesis also evaluates the potential benefits of high-resolution sinusoidal analysis in specific music signal processing application areas. A music signal processing application that can yield benefits if closely spaced sinusoids can be correctly resolved is multiple fundamental frequency estimation. This thesis presents and evaluates a multiple fundamental frequency estimator that operates on the parameter estimates of the signal subspace-based sinusoidal analysis system. In addition, a musical source separation method based on the presented sinusoidal analysis system is also designed and evaluated. If inference about the properties of individual sources is a step in content-based music analysis, then the evaluation of source separation performance provides an indication of the potential of the high-resolution sinusoidal analysis technique.

## 1.2 Thesis Organization

This thesis is organized as follows. Chapter 2 covers background information. The background serves to better define the problem this thesis aims to address, namely harmonic collisions. The prevalence of harmonic collisions in music mixtures is explored by analyzing symbolic music information. The background covers some of the existing approaches that have been proposed to handle these collisions in different application areas. Most importantly, the background chapter lays a theoretical and practical foundation for the variety of techniques and principles that are used throughout this thesis.

Chapter 3 presents the proposed sinusoidal analysis system. The sinusoidal parameter estimation technique used is the direct matrix pencil method. The direct matrix pencil method for sinusoidal parameter estimation is largely synonymous with the Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) method. This method is adapted to time-varying signals by producing the estimates on short frames of the signal, as done with most techniques that generate time-frequency representations. The sinusoidal analysis method is evaluated on a number of different types of signals.

Chapter 4 introduces a multiple fundamental frequency estimator designed to operate on the proposed high-resolution sinusoidal analysis method. The strategy adopted is a cancel-and-iterate approach where an estimate of a predominant fundamental frequency is made. The contributions of this predominant fundamental frequency estimate are then canceled, or removed, from the observed spectrum to produce a residual. The process is subsequently repeated on the residual spectrum and iterated until all fundamental frequency estimates are made. The system is evaluated and compared against other baseline systems including a current state-of-the-art technique.

Chapter 5 covers a computational auditory scene analysis (CASA) inspired method for producing source separations from the sinusoidal analysis. The system aims to group sinusoidal partials together based on similarities in their properties to form source estimates. These groups of partials can then be used to drive a synthesis method to produce separated source signals. The evaluation of the separated source signals provides some indication of how accurate the parameter estimates of the sinusoidal analysis system are.

Finally, Chapter 6 summarizes and discusses the findings of the previous chapters. Conclusions based on these findings are drawn. Potential future directions for the work presented in this thesis are presented based on the implications of the findings.

# CHAPTER 2

# BACKGROUND

The work presented in this thesis concerns itself primarily with the analysis and processing of musical audio signals. In order to gain a firm foundation for the work in the remaining chapters, this chapter discusses the nature of musical sounds, the nature of music itself, existing methods for analyzing musical audio, and some applications of these analysis methods. Section 2.1 introduces the sinusoidal model of pitched musical sounds which is used throughout this thesis. Section 2.2 covers time-frequency analysis and the most common methods for estimating the parameters of the sinusoidal model. Special attention should be paid to this section as it introduces a key problem inherent in signal analysis, notably the uncertainty principle which limits the simultaneous time and frequency resolution one can achieve in time-frequency representations. This section also introduces the direct matrix pencil method for estimating sinusoidal parameters, which is the analysis method explored and examined throughout this work. Section 2.3 discusses aspects of the nature of (Western) music and how rhythm and harmony rules interact to produce a large amount of overlap in time and frequency among musical sources. These overlaps, called *harmonic collisions*, create many difficulties in music signal processing applications and are the main challenge this thesis aims to address. Finally, Sections 2.4 and 2.5 cover two example application areas: multiple fundamental frequency estimation and musical audio source separation. Existing approaches to dealing with harmonic collisions in these two application areas are covered in their respective sections.

## 2.1 The Sinusoidal Model for Pitched Instrument Sounds

Musical instruments that create pitched sounds (i.e., musical notes) produce periodic or quasi-periodic waveforms. It has long been known from the advent of the Fourier series that such signals can be decomposed into a sum of sinusoidal *partials*. In the general case, a signal $x(t)$ composed of $K$ partials can be expressed as

$$x(t) = \sum_{k=1}^{K} A_k(t) \cos\left( \int_{\tau=0}^{t} 2\pi f_k(\tau)\, \mathrm{d}\tau + \phi_k \right) + n(t) \tag{2.1}$$

where $A_k(t)$, $f_k(\tau)$, and $\phi_k$ represent the time-varying amplitude, instantaneous frequency, and initial phase of the $k^{th}$ partial, respectively. Because real sounds generally contain some manner of noise component (e.g., the breath noise in a flute), a residual $n(t)$ is often also included. In perfectly periodic sounds, the partials obey a harmonic relationship. Each partial is an integer multiple of a fundamental frequency $f_0$. Therefore, Equation 2.1 can be rewritten as

$$x(t) = \sum_{k=1}^{K} A_k(t) \cos\left( \int_{\tau=0}^{t} 2\pi k f_0(\tau)\, \mathrm{d}\tau + \phi_k \right) + n(t) \tag{2.2}$$

When partials obey a harmonic relationship, they are refered to as *harmonics*.

Two simplifications are often made to the sinusoidal model. First, the noise component, $n(t)$, is often ignored. While noise components can be important to timbre, especially during attack portions of sounds, most musical information is carried in the harmonics. Second, the time-varying parameters of the sinusoidal model, namely the frequencies and amplitudes of harmonics, are assumed to be constant over short spans of time (e.g., less than 50 ms). Such an assumption lends itself well to short-time analysis methods where a signal is segmented into short (overlapping) frames. Although the sinusoidal parameters are assumed constant within each frame, these parameters are allowed to vary between frames. Therefore the time-varying nature of harmonic amplitudes and frequencies can be represented. For a time-frame $m$, the simplified sinusoidal model of a discrete-time signal sampled at a rate of

$f_s$ can be expressed as

$$x^{(m)}[n] = \sum_{k=1}^{K} A_k^{(m)} \cos(2\pi f_k^{(m)} n/f_s + \phi_k^{(m)}) \tag{2.3}$$

where $A_k^{(m)}$, $f_k^{(m)}$, and $\phi_k^{(m)}$ are the constant amplitude, frequency, and initial phase of the $k^{th}$ harmonic in frame $m$. Equation 2.3 is the model for pitched musical instrument tones used throughout this thesis.

## 2.2 Time-Frequency Analysis and Methods for Estimating Sinusoidal Parameters

A key goal in musical signal analysis is to estimate the salient parameters in Equation 2.3 for one or more musical sources in a music mixture. Estimation of sinusoidal parameters is referred to as *sinusoidal analysis* or *harmonic retrieval*. Sinusoidal analyses are most commonly derived directly from the time-varying spectrum of the sound or sound mixture. However, classes of harmonic retrieval techniques based on *signal subspace* techniques also exist. These techniques assume or determine an underlying model of the signal, onto which the signal is subsequently projected. Some signal subspace sinusoidal estimation techniques have the property of *super-resolution*, which refers to the ability to estimate the parameters of very closely spaced sinusoids (in frequency) within a small region of time support. Normally, the uncertainty principle of time-frequency analyses provides an inescapable bound on the simultaneous time and frequency resolution. However, it is important to note that the Fourier transform is in itself a completely nonparametric technique. That is, it assumes no underlying model of the signal. If a valid sinusoidal model is chosen or determined, and the signal very closely fits this model, super-resolution becomes possible.

This section covers the short-time Fourier transform, the uncertainty principle, and traditional methods for sinusoidal analysis. In addition, signal subspace harmonic retrieval techniques are also covered. Special focus is placed upon the direct matrix pencil method for estimating sinusoidal parameters. This technique is also commonly known as the Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) algorithm.

### 2.2.1 The short-time Fourier transform

The analysis of the behavior of a signal in both time and frequency simultaneously is referred to as time-frequency analysis. While time-frequency analyses comprise a variety of techniques including wavelet transforms [1], constant-Q transforms [2], etc., the most fundamental time-frequency representation is the short-time Fourier transform (STFT). A detailed coverage of analysis and synthesis using the STFT can be found in [3]. The STFT entails segmenting a sequence $x[n]$ into short, usually overlapping frames through the use of a compactly supported, sliding window function. A discrete Fourier transform (DFT) is performed on each frame. The end result of the STFT is a time-frequency representation of the signal that describes its time-varying spectrum. The time-frequency units of the STFT are spaced linearly in both time and frequency. For a signal, $x[n]$, the STFT produces a time-frequency representation for time-frame index $t$ and frequency index $l$ (a DFT *bin*) calculated as

$$X[t,l] = \sum_{n=0}^{N-1} x[n-tH]w[n]e^{-j\frac{2\pi l}{N}n} \tag{2.4}$$

The length of each time-frame is $N$. The window function $w[n]$ has a region of support of $n \in [0, N-1]$. The shape of the window function plays an important role, and is usually chosen to have good spectral characteristics such that side-lobes are suppressed, and crosstalk between neighboring DFT bins is low. A *hop* factor, $H$, controls how many samples each frame skips forward, and thus the amount of overlap between adjacent frames.

### 2.2.2 The uncertainty principle

The uncertainty principle, which holds for all time-frequency decompositions, states that perfect frequency resolution cannot be achieved in a limited span of time support, and vice versa. Denoting the standard deviation of a signal over time, $\Delta_t$ (a measure of the duration of the signal), and the standard deviation of the signal's spectrum, $\Delta_\omega$ (a measure of the signal's bandwidth), the uncertainty principle can be expressed as [4]

$$\Delta_t \Delta_\omega \geq \frac{1}{2} \tag{2.5}$$

The result of Equation 2.5 is that as duration shortens, bandwidth increases, and vice versa.

In the practical case of a STFT of a discrete-time signal calculated via Equation 2.4, the uncertainty principle manifests itself through the bandwidth of the main lobe of the window function, $w[n]$, used in analysis. For example, assume a signal, $x[n]$, composed of a mixture of two sinusoids, is sampled at $f_s = 44100$ Hz. The frequency of the first sinusoid is $f_1$ and the second, $f_2$, resulting in angular frequencies $\omega_1 = 2\pi f_1/f_s$ and $\omega_1 = 2\pi f_2/f_s$ Therefore, the signal $x[n]$ can be expressed as

$$x[n] = \cos(\omega_1 n) + \cos(\omega_2 n) \tag{2.6}$$

$X(e^{j\omega})$, the spectrum of $x[n]$, can be expressed as

$$X(e^{j\omega}) = \frac{1}{2}[\delta(\omega - \omega_1) + \delta(\omega + \omega_1) + \delta(\omega - \omega_2) + \delta(\omega + \omega_2)] \tag{2.7}$$

Assume in this example a frame length of $N = 2048$ samples (46 ms frame size) is used to truncate the signal and represent a single STFT frame. Furthermore, assume a Hamming window, $w[n]$, serves as the analysis window resulting in a windowed version of the signal $\hat{x}[n] = w[n]x[n]$. Because the window function is multiplied with the signal, and because the spectra of sinusoids are delta functions, the spectrum of the window function is modulated to be centered on each sinusoid of the signal $x[n]$. Denoting the Fourier transform of the Hamming window function $W(e^{j\omega})$, the resulting spectrum of the windowed signal, $\hat{X}(e^{j\omega})$, can be expressed as

$$\hat{X}(e^{j\omega}) = \frac{1}{2}[W(e^{j(\omega-\omega_1)}) + W(e^{j(\omega+\omega_1)}) + W(e^{j(\omega-\omega_2)}) + W(e^{j(\omega+\omega_2)})] \tag{2.8}$$

The main lobe width of a length-2048 Hamming window is $4f_s/N$ (or four bins of a 2048-point DFT) with a 6.0-dB bandwidth of 1.81 bins [5]. While zero-padding can be used to produce a more finely sampled DFT, the frequency spread caused by the window length remains constant. Therefore, if the two sinusoids are close in frequency, the additive combination of the overlapping spectra of the modulated window functions produces only a single

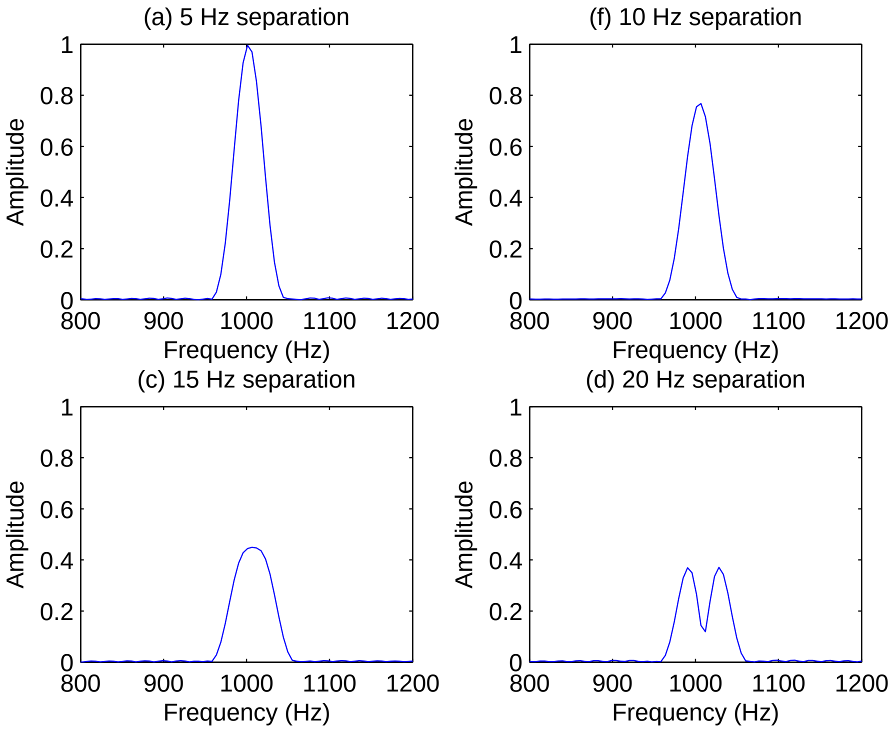

Figure 2.1: DFT of two sinusoids at spacings of (a) 5 Hz, (b) 10 Hz, (c) 15 Hz, and (d) 20 Hz.

prominent peak in the DFT. Equivalently, the STFT can be interpreted as a bank of bandpass filters, with each filter having the frequency response of the window function centered at each bin location of the DFT. Therefore, two closely spaced sinusoids will most strongly excite the same filter. Only at larger separations in frequency will the sinusoids be clearly resolvable.

Figure 2.1 shows the resulting spectra of two sinusoids at varying levels of separation in frequency. In this figure, the length-2048 Hamming-windowed signal is zero-padded to a length of 8192 samples. The sample rate is 44100 Hz. The first sinusoid is at $f_1 = 1000$ Hz. The frequency of $f_2$ is adjusted at 5 Hz increments above 1000 Hz. In this particular example, two peaks become clearly visible at a separation of 20 Hz ($f_1 = 1000$ Hz and $f_2 = 1020$ Hz). However, inspection of the DFT in this case shows that the peaks occur at 990 Hz and 1028 Hz. Extreme beating due to a separation of 20 Hz causes a phase cancellation in the DFT bins between the two sinusoids. Although there is strong evidence of multiple sinusoidal peaks in this case, their properties cannot be immediately deduced by a simple inspection of the magnitude spectrum. Moreover, the original signal $x[n]$ represents a possible best case scenario in that the two sinusoidal components are of equal amplitude. If one sinusoid is significantly weaker than the other in amplitude, the bandwidth of the window function plays a more significant role in obfuscating the other sinusoid. Naturally, increasing the frame/window size narrows the bandwidth of the window. However, such an increase comes at the expense of time resolution in a short-time analysis as the time-varying spectral magnitude characteristics are averaged over the length of the window.

### 2.2.3 STFT-based sinusoidal analysis

As stated previously, one key goal of musical signal analysis is to recover the salient parameters of the sinusoidal model of Equation 2.3 for one or more sources. This is often carried out by analysis of the short-time Fourier transform of the signal as described in Section 2.2.1. Two prevalent methods for obtaining sinusoidal parameters are phase-vocoder analysis [6] and sinusoidal tracking methods [7, 8].

A phase vocoder models a signal as a sum of sine waves with time-varying amplitude and frequency [9]. Using a filterbank interpretation of the STFT, if only a single sinusoid is present in each channel, its sinusoidal parameters can be measured. The measurement of the sinusoidal parameters is most easily achieved when the length of the window function is a multiple of the fundamental period of the periodic signal. In this case, the channels (i.e., DFT bin frequencies) are perfectly centered on the harmonic frequencies of the signal. In this case, the phase-vocoder is said to be pitch-synchronous [10]. While pitch-synchronous phase vocoders are effective for isolated single tones, constant retuning of the window length to follow a musical passage with changing pitches becomes cumbersome.

Alternatives to phase vocoder techniques are those techniques that rely on sinusoidal tracking. Procedures introduced by McAulay and Quatieri (MQ) [7] and simultaneously by Smith and Serra [8] also rely on STFT analysis. Peak picking of the magnitude spectrum above a magnitude threshold is used to estimate the frequencies and amplitudes of sinusoidal components for each frame. Smith and Serra perform quadratic interpolation of the three frequency points surrounding a peak to refine the estimates. Peaks are tracked

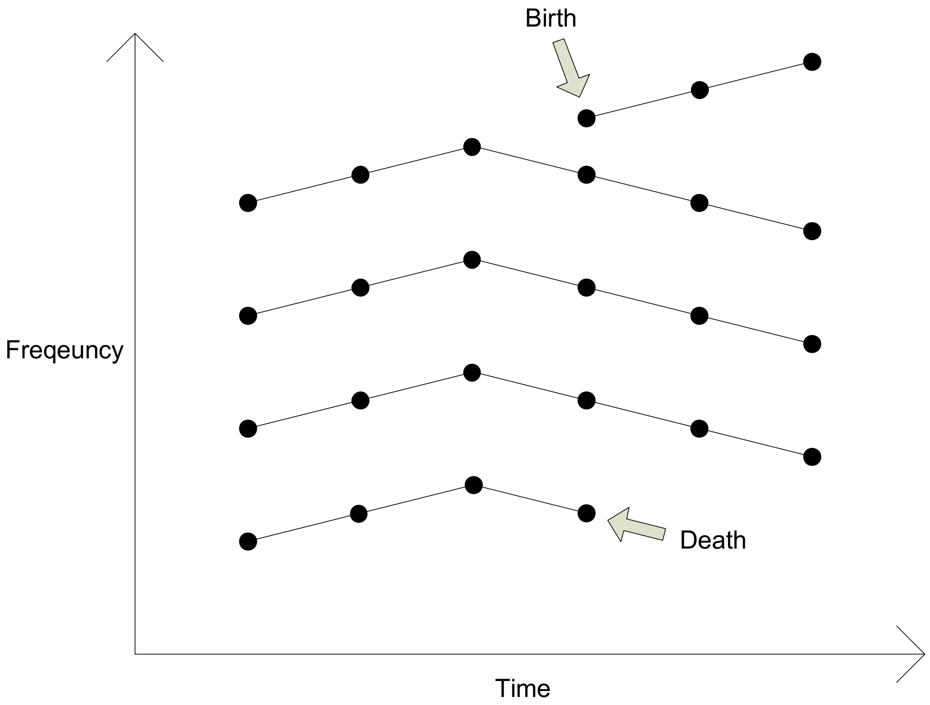

Figure 2.2: Sinusoidal tracking procedure. Time-frequency points corresponding to estimated sinusoids are tracked frame to frame if they are in close proximity in frequency. Tracks with no matches can be born or die.

and linked from frame to frame based on frequency proximity to produce frequency tracks. If no matches are found in a previous or subsequent frame, tracks are allowed to be *born* or *die*, respectively. The sound can be resynthesized from the frequency tracks using additive synthesis. A residual can be calculated by subtracting the sinusoidal synthesis from the original, provided a phase matching step is performed. A graphical representation of the sinusoidal tracking procedure can be seen in Figure 2.2. More sophisticated sinusoidal trackers include the use of hidden Markov models (HMMs) [11], [12] to track sinusoidal partials, or linear prediction techniques [13].

Revisiting the closely spaced sinusoids example in Figure 2.1(b), only a single visible peak is evident in a DFT frame. The sinusoidal tracking procedure of the MQ technique, where peak-picking is employed on each frame, will merge two sinusoids into a single track. Phase-vocoder analyses are also largely reliant on single harmonics being present in each filter channel to produce reliable estimates of the underlying sinusoidal parameters of the signal. Therefore, time-frequency resolution limits make resolving the sinusoidal parameters of collided harmonics due to multiple sources difficult and tend to merge closely-spaced sinusoids into single sinusoidal tracks.

### 2.2.4 Signal subspace-based harmonic retrieval and ESPRIT

A newer class of harmonic retrieval methods are based upon the principle of signal subspaces. They are also sometimes referred to as *super-resolution* techniques. Signal subspace methods concern themselves with decomposing a signal into a signal subspace (in this particular case, the sinusoidal harmonics) and a noise subspace. These techniques include Prony's method [14], Pisarenko [15], Tufts-Kumaresan [16] [17], ESPRIT [18], MUSIC [19], etc. Their use in music analysis was first explored in the context of analyzing isolated, percussively excited musical tones (e.g., piano, guitar, etc.) [20]. This thesis focuses on one of these signal subspace techniques, the direct matrix pencil method, also commonly known as ESPRIT.

Short-time subspace-based sinusoidal estimators that track harmonics have previously been developed. In [21] and [22], sinusoidal parameter estimates derived from ESPRIT are tracked using MQ-like methods. These approaches are practically equivalent to the sinusoidal analysis methodology used in this thesis, with differences only in implementation details. In [23], Badeau employs direct tracking of the signal subspaces as opposed to the tracking of parameter estimates at every frame. In fact, Badeau's thesis [24] can largely be considered a seminal work in regards to the application of ESPRIT to music signals. Badeau's work covers a broad range of topics, including model order estimation, spectral whitening and its performance benefits in ESPRIT estimation, resolution bounds, subspace-based tracking, and the effects of non-stationarity of sinusoidal frequencies. While the analysis method presented in this thesis follows directly from previous work, and is very largely inspired and influenced by the work of Badeau, these works leave what is perhaps the most important question unanswered. The motivation for using ESPRIT-based parameter estimation is its high-resolution potential. To date, no thorough evaluation of this potential has been performed. Most previous experiments focus on the analysis and synthesis of monophonic music signals. However, because traditional methods of sinusoidal analysis such as MQ are well established and known to work reasonably well for these signals, the use of high-resolution sinusoidal analysis is somewhat poorly motivated. Additionally, examples of the one or two polyphonic mixtures previously examined provide no indication of whether or not super-resolution is actually taking place. Thus, this work strives to answer the question: Is high-resolution sinusoidal analysis really high-resolution when applied to music signals? While the conceptual underpinnings of the analysis method presented in this thesis are not novel, it is hoped that the combination of the work of Badeau and this thesis, in concert, serve to establish the theoretical and practical implications of short-time high resolution sinusoidal analysis.

As a brief digression before the workings of the ESPRIT method are explained in detail, it should be noted that signal subspace-based techniques generate an additional subspace that is referred to as a noise subspace. It was previously stated that this thesis ignores the noise component of signals. Furthermore, many types of musical sounds are not restricted to be pitched, and thus harmonic, in nature. While this thesis focuses on signal subspaces, there has been some work in using these types of methods to extract the noise subspaces of signals. An interesting use of such subspace-based techniques can be found in [25], where the audio signal is projected onto the noise subspace to extract drum sounds.

In Section 2.1 the sinusoidal model for pitched sounds was presented. The direct matrix pencil method for estimating sinusoidal parameters uses a similar underlying sinusoidal model with one slight difference: The amplitudes of harmonics are not constant, but rather exponentially damped. Naturally, constant amplitudes are supported by such a model as a damping factor of zero produces constant amplitude. To simplify the notation in the following discussion, and to add the exponentially damped behavior of the sinusoidal amplitudes, the formulation of Equation 2.3 is slightly adapted. First consider that the signal is composed of complex exponentials. Therefore, a model composed of $K$ complex sinusoids will support $K/2$ real sinusoids (one complex sinusoid at positive frequency and one at negative frequency). A signal, $x[n]$, of length $N$, and consisting of $K$ exponentially-damped complex sinusoids, can be expressed as

$$x[n] = \sum_{k=1}^{K} R_k z_k^n \quad \text{for } n = 0, ..., N-1 \tag{2.9}$$

where

$$\begin{aligned}
R_k &= A_k e^{j\phi_k} \\
z_k &= e^{-\alpha_k + j\omega_k}. \\
A_k &= \text{Amplitude of } k^{th} \text{ harmonic} \\
\phi_k &= \text{Initial phase of } k^{th} \text{ harmonic} \\
\alpha_k &= \text{Exponential damping factor of } k^{th} \text{ harmonic} \\
\omega_k &= \text{Frequency of } k^{th} \text{ harmonic.}
\end{aligned}$$

As before, the goal of this harmonic retrieval technique is to estimate the sinusoidal parameters for all harmonics. The matrix pencil method for estimating sinusoidal parameters is now presented. The following derivations follow directly from [26], and are included here for completeness. First define two matrices, $X_0$ and $X_1$, containing the samples of $x[n]$ as

$$[X_0] = \begin{bmatrix} x[0] & x[1] & \cdots & x[L-1] \\ x[1] & x[2] & \cdots & x[L] \\ \vdots & \vdots & \ddots & \vdots \\ x[N-L-1] & x[N-L] & \cdots & x[N-2] \end{bmatrix}_{(N-L)\times L} \tag{2.10}$$

and

$$[X_1] = \begin{bmatrix} x[1] & x[2] & \cdots & x[L] \\ x[2] & x[3] & \cdots & x[L+1] \\ \vdots & \vdots & \ddots & \vdots \\ x[N-L] & x[N-L+1] & \cdots & x[N-1] \end{bmatrix}_{(N-L)\times L} \tag{2.11}$$

where $L$ is an analysis parameter called the *pencil parameter*. If $x[n]$ conforms to the model of Equation 2.9, $X_0$ and $X_1$ can be expressed as

$$[X_0] = [Z_L][R][Z_R] \tag{2.12}$$

and

$$[X_1] = [Z_L][R][Z][Z_R] \tag{2.13}$$

where

$$[Z_L] = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_K \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{N-L-1} & z_2^{N-L-1} & \cdots & z_K^{N-L-1} \end{bmatrix}_{(N-L)\times K} \tag{2.14}$$

$$[Z_R] = \begin{bmatrix} 1 & z_1 & \cdots & z_1^{L-1} \\ 1 & z_2 & \cdots & z_2^{L-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_M & \cdots & z_K^{L-1} \end{bmatrix}_{K\times L} \tag{2.15}$$

$$[R] = \begin{bmatrix} R_1 & 0 & \cdots & 0 \\ 0 & R_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_K \end{bmatrix}_{K\times K} \tag{2.16}$$

$$[Z] = \begin{bmatrix} z_1 & 0 & \cdots & 0 \\ 0 & z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & z_K \end{bmatrix}_{K\times K} \tag{2.17}$$

The factorization of Equations 2.12 and 2.13 can be easily verified by substitution. The damped sinusoids $z_k \in \{z_1, z_2, ..., z_K\}$ are referred to as *poles*. The matrix pencil is formed as

$$X_1 - \lambda X_0 \tag{2.18}$$

Substituting Equations 2.12 and 2.13 into 2.18, the matrix pencil can be written as

$$X_1 - \lambda X_0 = Z_L R[Z - \lambda I] Z_R \tag{2.19}$$

where $I$ is the $K \times K$ identity matrix. In general, the rank of the pencil will be $K$. However, if $\lambda = z_1, z_2, ..., z_K$, a row/column of the pencil becomes zero, and the rank of the pencil is reduced to $K-1$. These rank reducing values of $\lambda$ represent the poles present in the signal. In order to solve for all rank reducing values of $\lambda$, the problem is formulated as a generalized eigenvalue problem as follows:

$$(X_1 - \lambda X_0)\mathbf{q} = 0 \tag{2.20}$$

$$\mathbf{p^H}(X_1 - \lambda X_0) = 0 \tag{2.21}$$

with $\mathbf{q} \in R\{X_0^H\}$ and $\mathbf{p} \in R\{X_0\}$. Equations 2.20 and 2.21 hold when $\lambda = z_k \in \{z_1, z_2, ..., z_K\}$, $q = q_k =$ the $k^{th}$ column of $Z_R^+ = Z_R^H(Z_R Z_R^H)^{-1}$, and $p = p_k =$ the $k^{th}$ row of $Z_L^+ = (Z_L^H Z_L)^{-1} Z_L^H$. The following method can then be used to solve for the generalized eigenvalues. Left-multiplying Equation 2.20 with $X_0^+$ gives

$$\begin{aligned}
&X_0^+ X_1 \mathbf{q} - \lambda X_0^+ X_0 \mathbf{q} = 0 \Rightarrow \\
&X_0^+ X_1 \mathbf{q} - \lambda \mathbf{q} = 0 \Rightarrow \\
&(X_0^+ X_1 - \lambda I)\mathbf{q} = 0
\end{aligned} \tag{2.22}$$

Therefore, the system poles can be solved for by simply calculating the eigenvalues of the square matrix $X_0^+ X_1$. Since $X_0^+ X_1$ is rank $K$, it will contain $K$ nonzero eigenvalues (the poles) and $L-K$ zero eigenvalues.

Heretofore, the discussion of the matrix pencil method for estimating the parameters of damped sinusoids has focused on the noiseless case. In the presence of noise, the signal matrices $X_0$ and $X_1$ are formed as before. However, the full pseudoinverse $X_0^+$ is instead replaced with a rank-$K$ truncated pseudoinverse. Expressing the SVD of $X_0$ as $U_0 \Sigma V_0^H$, the rank-$K$ truncated pseudoinverse is $X_{0K}^+ = V_{0K} \Sigma_K^{-1} U_{0K}^H$, where $\Sigma_K$ contains only the largest $K$ singular values, and $U_{0K}$ and $V_{0K}$ contain the $K$ corresponding rows and columns, respectively. Once again, performing the eigenvalue decomposition of $X_{0K}^+ X_1$ will yield the $K$ signal poles and $L-K$ zero eigenvalues. Since only $K$ eigenvalues are nontrivial, the computation can be simplified to that of an eigenvalue decomposition of an $K \times K$ matrix as opposed to an $L \times L$ matrix. Substituting the decomposition of $X_{0K}^+$ into Equation 2.22 for $X_0^+$, we get

$$V_{0K} \Sigma_K^{-1} U_{0K}^H X_1 \mathbf{q} = \lambda \mathbf{q} \tag{2.23}$$

Because $V_{0K}^H V_{0K} = I$ and $V_{0K} V_{0K}^H \mathbf{q} = \mathbf{q}$, left-multiplying Eq 2.23 with $V_{0K}^H$

yields

$$\Sigma_K^{-1} U_{0K}^H X_1 V_{0K} (V_{0K}^H \mathbf{q}) = \lambda (V_{0K}^H \mathbf{q}) \tag{2.24}$$

Therefore, the system poles can be solved for by performing an eigenvalue decomposition on the $K \times K$ square matrix $\Lambda = \Sigma_K^{-1} U_{0K}^H X_1 V_{0K}$.

Solving for the eigenvalue of matrix $\Lambda$ gives the poles $\{z_1, z_2, ..., z_K\}$. From these values, the frequencies and damping factors can be calculated directly as

$$\begin{aligned} \omega_k &= \angle z_k \\ \alpha_k &= -log|z_k| \end{aligned} \tag{2.25}$$

The harmonic amplitudes and initial phases are solved for as a least squares problem. The signal $x[n]$ can be expressed in matrix form in terms of the poles and complex amplitudes as

$$\mathbf{x} = Z_p \mathbf{r} = \\ \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_K \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{N-1} & z_2^{N-1} & \cdots & z_K^{N-1} \end{bmatrix} \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_K \end{bmatrix} \tag{2.26}$$

We now wish to recover $\mathbf{r} = [R_1, R_2, ..., R_K]^T$. This inverse problem can be solved for in the least squares sense as

$$\mathbf{r} = (Z_p^H Z_p)^{-1} Z_p^H \mathbf{x} \tag{2.27}$$

The amplitude, $A_k$, and initial phase, $\phi_k$, can be recovered from each complex amplitudes $R_k$ by taking the magnitude and phase, respectively, as

$$\begin{aligned} A_k &= |R_k| \\ \phi_k &= \angle R_k \end{aligned} \tag{2.28}$$

Until now, there has been little discussion as to how the key parameters, namely the pencil parameter $L$ and the number of sinusoids $K$, are chosen. Good values of $L$ have been found empirically to be $N/3$ to $N/2$ [27, 28]. While the number of sinusoids is not known *a priori*, over-estimating does not lead to adverse effects. Therefore, $K$ should be chosen sufficiently large.

If a signal fits the model described in Equation 2.9, the direct matrix

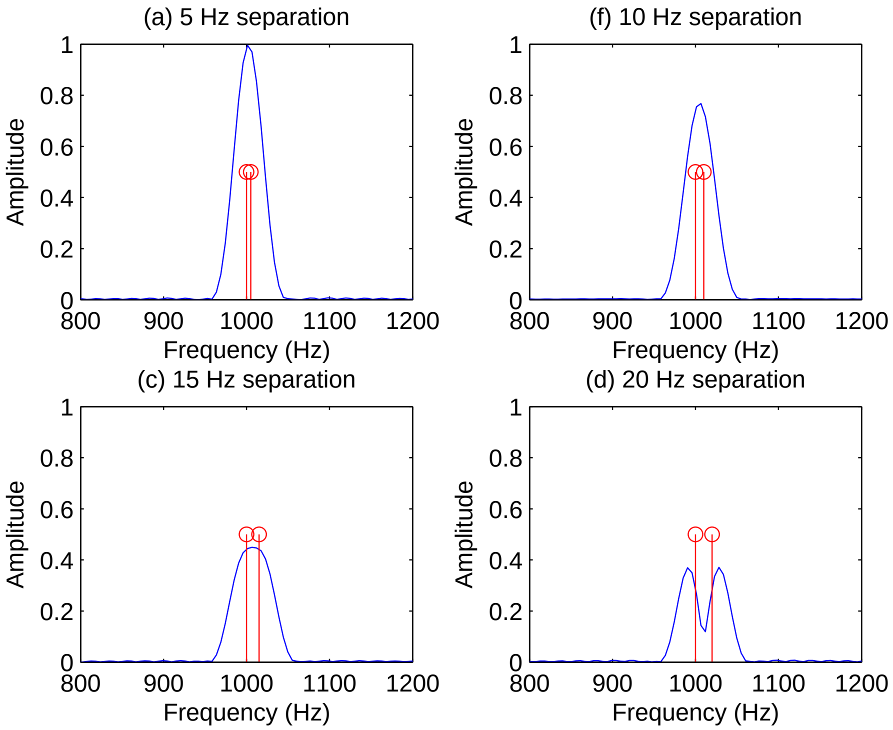

Figure 2.3: Matrix pencil estimation of two sinusoids at spacings of (a) 5 Hz, (b) 10 Hz, (c) 15 Hz, and (d) 20 Hz. The estimates are represented as the stems. The DFT of the mixture is also shown.

pencil method has the ability to resolve and calculate the parameters of closely spaced sinusoids. Revisiting once again the closely spaced sinusoids examples in Section 2.2.2, Figure 2.1, if the direct matrix pencil is employed on the length-2,048 frame, the sinusoidal parameters are perfectly recovered for all spacings in frequency. In this example, a rectangular window is used to truncate the signal for matrix pencil analysis. Figure 2.3 demonstrates the matrix pencil estimates of the same sinusoids in Figure 2.1 superimposed on the DFT. With perfect model fit, sinusoids at arbitrarily close proximity in frequency can be resolved (up to numerical precision limits). In reality, musical signals do not completely match the piece-wise constant frequency and amplitude sinusoidal model of Equation 2.3. Although in a short span of time frequencies of the sinusoid will not vary greatly, they are indeed not stationary. Deviation from a constant-frequency model can be considered as a form of *model noise*. Discussion of the effects of both signal and model noise will be reserved for later in Chapter 3.

## 2.3 Musical Mixtures

The discussion in Section 2.2 demonstrated that if two musical sources have energy at the same point in a time-frequency representation, the individual contributions of each source become difficult to resolve. Because musical sources are sparse in the frequency domain (i.e., most of their energy is located only at harmonic positions of the fundamental), one could expect that these harmonic collisions may be a rare and insignificant problem. However, the very nature of Western music composition leads to a high degree of overlap among sources' time-frequency representations. In this section, a brief overview of the nature of musical mixtures and musical scales is introduced to explain why harmonic collisions are so prevalent. In addition, an analysis of symbolic music data is performed in order to roughly quantify the rate of occurrence of these harmonic collisions in Western music.

### 2.3.1 Rhythm, harmony, and musical scales

Music theory describes music through a variety of elements such as rhythm, harmony, melody, texture, form, etc. Two of these aspects, rhythm and harmony, play a significant role in the placement of musical sources' harmonics in time and frequency.

Rhythm refers to the temporal arrangement of musical sounds and silences. In general, music is subdivided along the time axis by some fundamental unit of time, usually referred to as a *beat*. Beats themselves can be further subdivided. Rhythm refers to an underlying, and usually repeating, pattern of the temporal arrangement of musical notes. When music contains multiple sources, musical notes are often arranged to begin and end very closely in time so as to maintain some form of structured composite rhythm. In other words, most musical sources tend to show a high degree of temporal alignment in musical mixtures.

Just as rhythmic rules govern the temporal arrangement of music pieces, harmony rules govern the use of simultaneous pitches. The notion of harmony as it relates to musical notes played simultaneously is largely based on *consonance* and *dissonance*. Perceptually, consonance refers to a combination of two musical pitches that sound "pleasant" when played together. Dissonant pitches do not share this property, and are functionally used in

Table 2.1: The 12 simple intervals relative to C.

| **Note Name** | **Number of Semitones** | **Interval Name** | **Frequency Ratio**† |
|---|---|---|---|
| C | 0 | Unison (P1) | 1:1 |
| C♯ / D♭ | 1 | Minor Second (m2) | 16:15 |
| D | 2 | Major Second (M2) | 9:8 |
| D♯ / E♭ | 3 | Minor Third (m3) | 6:5 |
| E | 4 | Major Third (M3) | 5:4 |
| F | 5 | Perfect Fourth (P4) | 4:3 |
| F♯ / G♭ | 6 | Augmented Fourth (A4)<br>Diminished Fifth (d5) | 45:32 |
| G | 7 | Perfect Fifth (P5) | 3:2 |
| G♯ / A♭ | 8 | Minor Sixth (m6) | 8:5 |
| A | 9 | Major Sixth (M6) | 5:3 |
| A♯ / B♭ | 10 | Minor Seventh (m7) | 9:5 |
| B | 11 | Major Seventh (M7) | 15:8 |
| C | 12 | Perfect Octave (P8) | 2:1 |

†Frequency ratios are approximate for equal temperament tuning

music to introduce tension. In general, sounds that are consonant have a large number of coinciding harmonics.

To better understand how consonance arises, the notion of musical pitches, scales, and intervals must be explained. Western music is built on twelve pitches per octave. In equal temperament, these twelve pitches equally subdivide the octave on a logarithmic scale. An *interval* is the relationship between two pitches. In simplest terms, a *key* refers to which pitch serves as the harmonic center of piece (*tonic*), as well as the mode (the group of music intervals) used to define the key (e.g., major or minor). In general, the names of the pitches are unimportant. The intervals, and their relation to the tonic of the musical piece, carry the significant musical information of the piece. Because there are 12 musical tones, there exist 12 possible base intervals. When pitches are described in terms of frequency, many of the intervals can be expressed as simple integer ratios of the fundamental frequencies of the two notes. The most basic interval is the unison, where both pitches are identical, and have fundamental frequency ratios of 1:1 (e.g., both notes are C at identical pitch heights). A perfect octave occurs at ratios of 2:1 (e.g., C and C one octave above). A perfect fifth occurs at ratios of 3:2, and perfect fourths at ratios of 4:3. Table 2.1 contains a list of intervals with their separations in semitones relative to the root, the corresponding note names for a root of C, their common interval names, and the ratios of fundamental frequencies relative to the root. Recall from Section 2.2 that pitched, musical tones produce harmonics of their fundamental frequency. When two pitches are played at simple integer ratios of one another, many of their harmonics coincide. For example, with the perfect fifth and a ratio of fundamental frequencies of 3:2, every second harmonic of one source will coincide with every third harmonic of the other source. The coincidence of partials gives rise to consonance and explains how harmony can make harmonic collisions prevalent in music.

### 2.3.2 Analysis of symbolic music data

An analysis of symbolic music data was performed in order to better understand how prevalent harmonic collisions are in music. A dataset of 1,252 music pieces in MIDI format, derived from the “Symbolic Key Finding” task of the 2005 Music Information Retrieval Evaluation eXchange (MIREX) [29], was used in this analysis. The corpus of MIDI files comprises pieces of Baroque, Classical, and Romantic music. Two separate analyses were performed. First, histograms of the occurrence of pitch intervals for major and minor modes were constructed. This analysis provides insight into how often pitches that can be considered consonant or dissonant with respect to the tonic of the musical key occur. Second, an analysis of the MIDI files was performed to estimate how often harmonics of one musical source would be corrupted by other sources.

#### Frequency of occurrence of note intervals

Histograms of the occurrence of note intervals were constructed in order to estimate the distribution of interval occurrences in Western music. The dataset of 1,252 MIDI pieces was marked up with the tonic and mode of the piece. Half of the pieces are major, the other half, minor. First, all pieces in the same mode (major or minor) were normalized to be the same key (C) by a simple transposition operation. Separate histograms were constructed for each of the two modes. Only simple, and not compound, intervals were considered. In other words, the octave of each note was not considered,

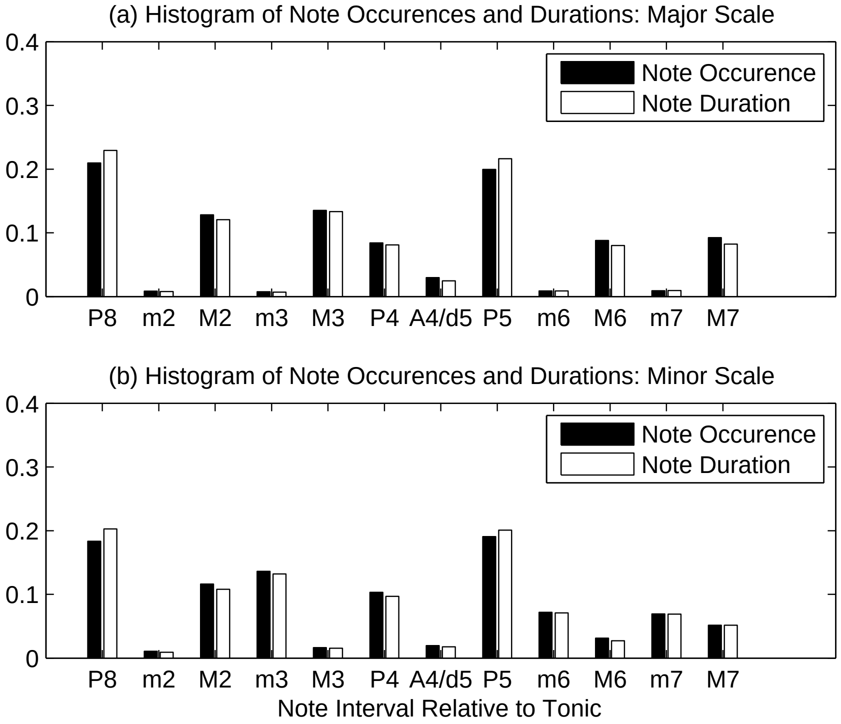

Figure 2.4: (a) Histogram of interval occurrences and duration-weighted occurrences for major scales. (b) Histogram of interval occurrences and duration-weighted occurrences for minor scales.

resulting in a histogram containing 12 bins, one for each simple interval. Each note of each MIDI file for one of the two modes was counted and placed in its corresponding histogram bin. Therefore, each occurrence of the note G was placed in the bin corresponding to the fifth (G being the fifth of C). In addition to a simple counting of the occurrence of each note, a second histogram was constructed where each note was weighted by its duration. The resultant histograms (one for simple note occurrence, the other weighted by duration) can be seen in Figure 2.4 for each of the two modes.

As seen in Figure 2.4 the tonic and fifth are (as expected) the two most frequently occurring notes. The third (which defines the major or minor triad) also plays a prominent roll for each mode. The histograms also show that major seconds and fourths occur frequently across modes. Many of these frequently occurring intervals produce the potential for a large number of coinciding and thus, colliding, harmonics.

### Frequency of occurrence of harmonic collisions

The previous analysis focused on simply estimating the note/interval distribution of musical keys. This analysis did not take into account whether these pitch-intervals were occurring simultaneously or not. Klapuri for instance demonstrated that in the case of a simultaneous major triad 47%, 33%, and 60% of the harmonic partials of the root, major third, perfect fifth, respectively, are overlapped by the other notes in the chord [30]. Therefore, an analysis was performed to measure the simultaneous occurrence of intervals and harmonic collisions in Western music. First, the dataset of MIDI files had to be filtered to produce a valid subset of multisource pieces. All pieces that were solo piano, harpsichord, or organ were removed. Moreover, the General MIDI standard contains a patch for string ensembles. All string ensemble MIDI channels were removed. The resulting set of MIDI files contained 382 pieces with more than one source.

For each MIDI file, a simple time-frequency representation of each individual source, based solely on the symbolic information, was constructed. A hop size of 10 ms and a frequency bin spacing of 21.5 Hz (equivalent to a 2,048 point DFT of a 44.1 kHz sampled signal) was used. For each note of the source, the corresponding time-frequency points were marked based on the note's onset time, offset time, and the expected harmonic locations of the pitch (up to 10 kHz). Overlaps in the time-frequency representations, and therefore harmonic collisions, were counted when contributions of two or more sources existed at the same time-frequency points. Figure 2.5 shows the time-frequency representation of an excerpt of the first 5 seconds of a piece containing violin, viola, and French horn. The time-frequency points shared by multiple sources are also shown.

In the analysis of the 382 MIDI files, it was found that 29.6% of all nontrivial fime-frequency points (i.e., time frequency points with a contribution due to at least one source harmonic) were occupied by more than one source. However, when sources were considered individually, it was found that, on average, 50.1% of an individual source's nontrivial time-frequency points were interfered with by other sources.

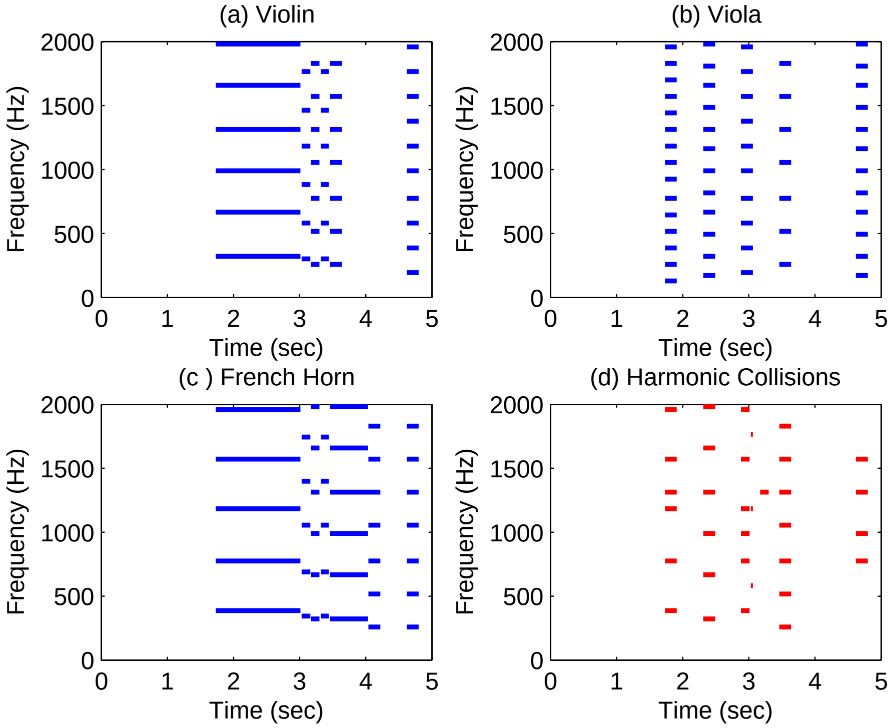

Figure 2.5: MIDI-derived time-frequency representations for a piece containing (a) violin, (b) viola, and (c) French horn. The time-frequency points shared by two or more sources can be seen in (d).

## 2.4 Multiple Fundamental Frequency Estimation

Multiple fundamental frequency (multi-$f_0$) estimation concerns itself with estimating the fundamental frequency, $f_0$, of sources in polyphonic music mixtures. Multi-$f_0$ estimation produces a low-level musical transcription of a piece. It is a simplified version of multipitch analysis (though $f_0$ is very strongly related to pitch, the complexities of pitch perception do not produce a true one-to-one mapping between the two). In its own right, however, multi-$f_0$ analysis can be thought of as an important first step to true multipitch estimation, and later, music transcription. Aside from music transcription, multi-$f_0$ estimation has additional applications in general music information retrieval (MIR), as well as source separation.

This section forms a foundation for multi-$f0$ estimation by first introducing the $f_0$-estimation of audio containing only a single source. Methods of multiple-$f_0$ estimation are subsequently presented. Finally, methods for evaluating $f_0$ estimation techniques are discussed.

### 2.4.1 Single, monophonic, fundamental frequency estimation

Early work in $f_0$ estimators focused on signals containing only one $f_0$ at any given time. Signals with a single $f_0$, or pitch, at any given time are called *monophonic* signals. In general, single-$f_0$ estimators comprise time-domain methods, and frequency-domain (or, more accurately, time-frequency domain) methods. Time-domain techniques for single-$f_0$ estimation include, among others, autocorrelation function (ACF) methods [31], average magnitude difference functions (ADMF) methods [32], and variants of squared distance function (SDF) based methods such as YIN [33]. Frequency-domain approaches include spectral autocorrelation [34], cepstral methods [35], the harmonic product spectrum [36], and techniques that attempt to match observed spectra with some form of harmonic model.

While many of the techniques may seem to differ on the surface, there is a strong underlying equivalence among many of the approaches. For example Tolonen and Karjalainen observed that autocorrelation functions greatly resemble cepstral analysis. The autocorrelation can be calculated as the inverse Fourier transform of the squared-magnitude spectrum. Cepstral analysis, on the other hand, is simply the Fourier transform of the log magnitude spectrum. Therefore, the main underlying difference is the extent to which the magnitude spectrum is compressed or expanded. Moreover, Klapuri showed that both these methods are implicit realizations of a model that emphasizes frequency partials at harmonic locations of the magnitude spectrum [37]. Therefore, these approaches share a strong similarity with harmonic pattern matching techniques. For example, some harmonic pattern matching techniques introduce a concept of a harmonic *comb* or *seive* [38]. Combs are constructed by placing weighting functions at harmonic locations of an $f_0$ hypothesis. The spectrum is subsequently weighted by the comb and integrated over frequency (i.e., correlated) to produce a *salience* score for a given $f_0$ hypothesis.

Figure 2.6 shows an example of a generic harmonic comb pattern matching technique for a single frame of a French horn tone. The comb is constructed by placing Gaussians with a 10 Hz standard deviation at harmonic locations of the fundamental frequency. When extracted harmonic amplitudes are weighted by the comb and then integrated, combs that match the true $f_0$ will produce a high salience score. The harmonic comb corresponding to the

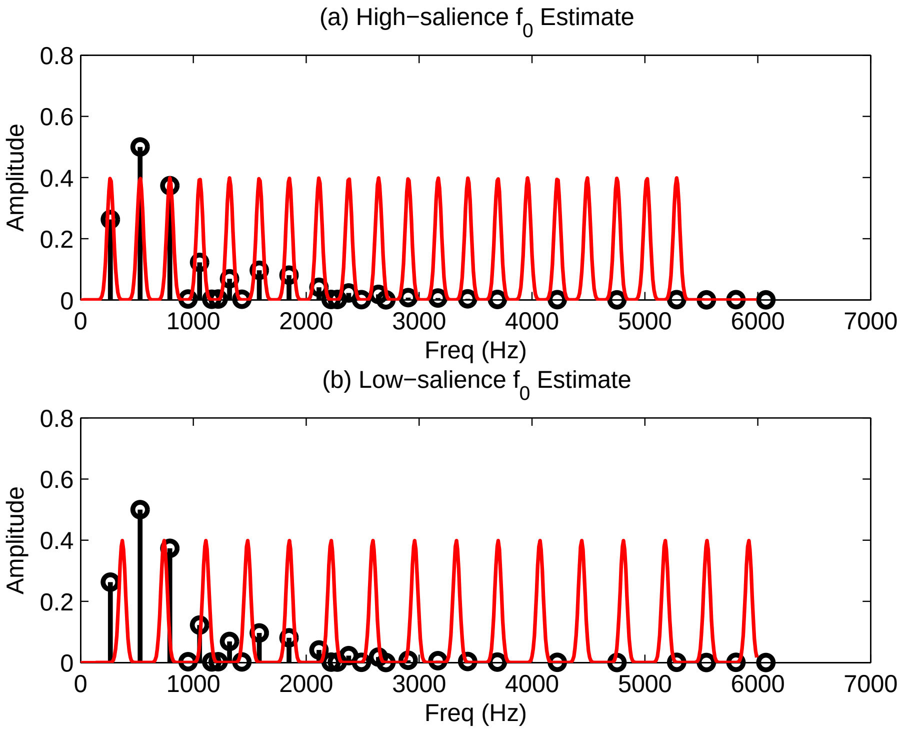

Figure 2.6: Harmonic combs of two $f_0$ hypotheses. Harmonic partials of a French horn tone at an $f_0$ of 261 Hz are shown as stems. The harmonic comb of an $f_0$ hypothesis of 261 Hz is shown in (a). The harmonic comb of an $f_0$ hypothesis of 370 Hz is shown in (b).

true $f_0$ is seen in Figure 2.6(a). The harmonic comb of an $f_0$ hypothesis with little overlap of the observed spectrum is seen in Figure 2.6(b).

### 2.4.2 Multiple, polyphonic, fundamental frequency estimation

When signals contain multiple sources at a given time, they are referred to as *polyphonic* signals. For polyphonic signals, sometimes the main $f_0$ contour of interest is that of the predominant melody. Poliner *et al.* provide a good overview of melody estimation techniques [39]. The estimation of multiple, simultaneous, fundamental frequencies is a significantly more challenging problem than single-$f_0$ estimation and melody estimation. According to Yeh, multi-$f_0$ estimation methods can largely be classified into two categories: Iterative-cancellation techniques and joint estimation techniques [40].

Iterative-cancellation methods involve estimating the $f_0$ of a predominant source, and then canceling the contributions of the that source. For exam-

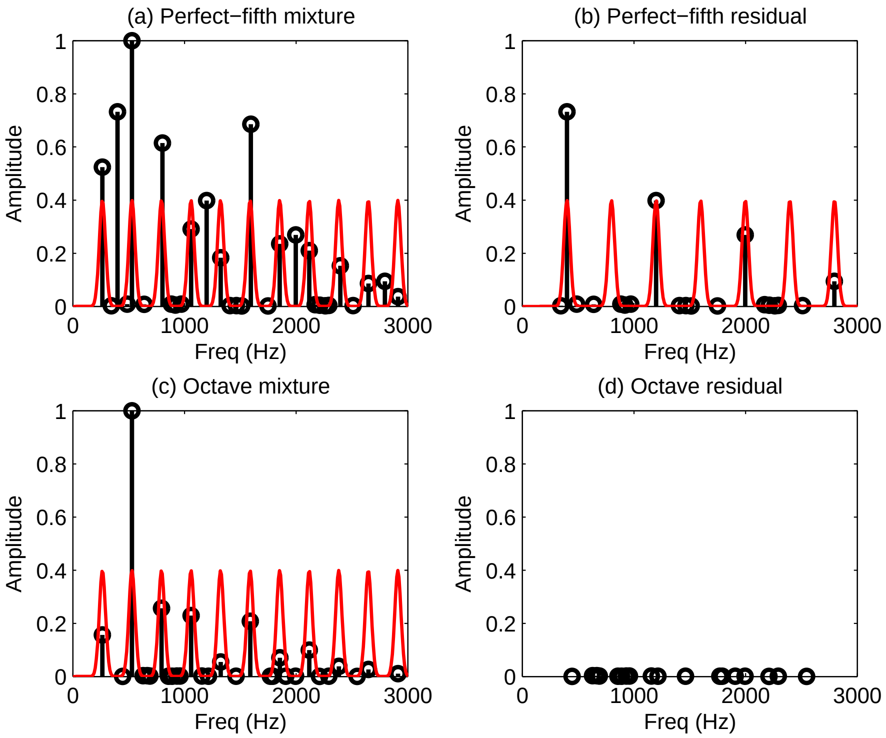

Figure 2.7: Example of iterative cancellation multiple-$f_0$ estimation techniques for a perfect-fifth and octave mixture. The harmonic partials of the perfect-fifth mixture, along with the comb pertaining to the highest salience $f_0$ estimate, can be seen in (a). The residual after the harmonics corresponding to the $f_0$ estimate in (a) are canceled, and the comb of the second dominant $f_0$ estimate are shown in (b). The harmonic partials of the octave mixture, along with the comb pertaining to the highest salience $f_0$ estimate can be seen in (c). Cancellation of the harmonics of this candidate $f_0$ produces virtually no residual, and a subsequent $f_0$ cannot be estimated as shown in (d).

ple, all harmonics corresponding to an estimated $f_0$ might be fully removed from the observed spectra. This procedure is iterated until all sources are accounted for. The main advantage of this technique is its low computational burden. However, due to harmonic collisions, the cancellation of individual sources has a propensity to also remove valuable information that may be necessary to estimate the $f_0$ of other sources. Figure 2.7 demonstrates the cancellation effects that can occur with mixtures containing harmonic collisions. In this example, spectra are represented as harmonic partials using peak-picking of the DFT magnitude spectra. As stated earlier, colliding partials generate a single peak, and thus a single visible harmonic. In the case of this perfect-fifth mixture, the first predominant $f_0$ found is that of the lower (in pitch) tone. Once the harmonics corresponding to this $f_0$ are fully canceled, every other harmonic of the higher tone (a perfect fifth above) is also canceled. In this case however, there is enough of a residual to correctly support a second, and correct, $f_0$ hypothesis. The case gets more complicated for octave mixtures. Cancellation of the predominant $f_0$ estimate results in full cancellation of all significant harmonics. The residual comprises only weak, spurious partials, and no subsequent $f_0$ estimate can be made.

Parsons introduced an early attempt to detect collisions based on symmetry of DFT bins surrounding a harmonic peak in the spectrum, as well as well behaved phase of those bins [41]. Klapuri proposed partial cancellation methods where the contribution of an estimated source is not fully removed. Two methods explored by Klapuri include partial cancellation based on various spectral smoothness principles, and the cancellation of only lower frequency harmonics [42, 43].

Joint estimation techniques aim to estimate all combinations of $f_0$ candidates. As the number of sources increases, so do the number of possible $f_0$ combinations. Therefore, such approaches come at the expense of a higher computational cost. An example of a joint estimation technique for two-tone mixtures is the two-way mismatch method [44]. Proximity of partials (in frequency) are evaluated against the locations of the hypothesis of joint $f_0$s and vice versa. Another example of joint estimation techniques by de Cheveigné cancels all contributions due to a joint hypothesis of multiple $f_0$s [45] to produce a residual. The joint-$f_0$ hypothesis with minimum residual is chosen as the estimate. Yeh proposes a hybrid approach between joint and iterative estimation to reduce the required search space of joint-$f_0$ hypotheses [40]. Yeh also points out that the main underlying advantage of joint estimation techniques is that they are better adapted to handling harmonic collisions. With a joint-$f_0$ hypothesis, the locations of expected harmonic collisions are easily determined, and the contributions of those harmonics can be shared among the individual $f_0$ estimates in a joint hypothesis. Approaches based on non-parametric techniques such as non-negative matrix factorization (NMF) [46] and statistical modeling techniques such as harmonic temporal clustering [47] can also be considered joint-estimation techniques.

### 2.4.3 Evaluation measures for multiple fundamental frequency estimation performance

In musical contexts, the estimated $f_0$s are usually deemed correct if they are within a semitone tolerance window centered on the ground-truth $f_0$. This allowed error tolerance is roughly a ±3% error in frequency. Because the most common errors in $f_0$ are octave and suboctave errors, additional evaluations sometimes map all $f_0$s to a single octave. In this case, the evaluation can be said to be an evaluation of the *chroma* accuracy. Chroma accuracy makes sense from a musical perspective in that the note that pertains to a given frequency carries similar musical meaning regardless of its octave. In multiple-$f_0$ estimation, additional factors must be taken into account because usually the number of simultaneous $f_0$s is not known beforehand. Therefore, aspects such as false alarm and false negative rates must be measured as well. Poliner and Ellis propose a series of evaluation measures that take into account both accuracy and detection rates [48]. Multiple-$f0$ estimation has been an evaluation task of the annual MIREX campaign since 2007 [49]. Bay *et al.* provides an overview of the performance metrics and current state-of-the-art performance in musical multi-$f0$ estimation [50].

## 2.5 Musical Audio Source Separation

Audio source separation aims to separate individual sources from audio mixtures. In a musical context, consider an example of a string quartet that usually contains two violins, a viola, and a cello. A goal of musical audio source separation is to separate the signals corresponding to the individual musical instruments. For the simplest monaural (one channel) mixture, given a signal $x[n]$ composed of $J$ sources, $x[n]$ can be expressed as the linear superposition of sources

$$x[n] = \sum_{j=1}^{J} x_j[n] \tag{2.29}$$

where $x_j[n]$ represents the $j^{th}$ source. The goal of source separation is to extract each source, $x_j[n]$ (e.g., the cello), from the composite mixture, $x[n]$. A more relaxed constraint on source separation is to recover a source signal that *sounds* identical to each true source. This section presents an overview of audio source separation approaches. Focus is placed upon computational auditory scene analysis because it interacts nicely with the sinusoidal model of instrument tones, and is intuitive to conceptualize. Finally, discussions on how source separation techniques are typically evaluated are presented.

### 2.5.1 Overview of source-separation methods

In general, source separation techniques can be classified into four classes: Unsupervised methods, model-based methods, multichannel methods, and perceptually-inspired auditory scene analysis methods. These four classes are by no means disjoint, as some techniques can be considered to belong to more than one of these generalized classes.

Unsupervised methods for audio source separation attempt to learn characteristics of source instruments directly from the data. Examples of such audio source separation techniques are ones related to independent component analysis (ICA) [51], nonnegative matrix factorization (NMF) [52, 53, 54, 55], and independent subspace analysis (ISA) [56, 57, 58, 59]. While ICA is generally a multichannel method, related methods such as NMF and ISA can work on monaural mixtures. For instance, ISA uses ICA on spectrograms to factorize the observed spectrogram and then cluster the factors into sources. The NMF method aims to decompose observed spectra into a small number of basis spectra, that when additively combined represent the observed spectra. The non-negativity constraint comes into play in that magnitude spectra are strictly positive, and they are additively combined during mixing. As an added constraint, it is generally desired for these decompositions to be as sparse as possible. Additional constraints can include the harmonicity of basis spectra or temporal smoothness constraints [60].

Model based methods use models of sources to aid source separation. With large dictionaries of instrument models spanning both pitch and dynamics, matching pursuit [61] can be used to decompose a music signal in terms of the dictionary elements [62]. Bay and Beauchamp [63] use pre-stored spectra to deal with harmonic collisions by replacing the harmonic amplitudes of collided/corrupted harmonics with ones taken from the best matching spectra in their library. Bayesian schemes relying on prior parametric models have also been explored in various contexts [64, 65, 66, 67, 68]. Various assumptions regarding the smoothness of spectral envelopes and the sinusoidal model have also been attempted to aid in the harmonic collision problem [69, 70, 71, 72]. Methods dependent on multi-pitch estimation and subsequent least-squares estimation of collided harmonic amplitudes are presented in [73] and [74].

When more than one channel of recording is available, a variety of techniques can be used in source separation. Beamforming [75] and other microphone array techniques perform well when there are an adequate number of sensors. In general however, music is distributed in stereo, and thus, only two channels are available. Techniques have been developed to discover the key mixing parameters of each source (interchannel intensity and time differences) from two-channel recordings [76, 77, 78, 79, 80]. Time-frequency points exhibiting the same interchannel differences can then be grouped together and separated. Harmonic collisions however affect the estimation of the mixing parameters and subsequently the unmixing. Perfect unmixing can only be achieved under the condition of W-disjoint orthogonality [81], that is, in cases where no harmonic collisions take place. Viste and Evangelista pay specific attention to resolving harmonic collisions in stereo source separation by noting the shapes of harmonic envelopes and beating patterns [82].

Following Bregman's [83] seminal work on auditory scene analysis, various approaches that attempt to model, by computer, what humans are believed to use as the primary cues for grouping and segregating sounds have been researched. Objects are formed from elementary time-frequency points or harmonic tracks using grouping cues that include common fate (common onset/offset [84], AM and FM modulation of harmonic components [85]), harmonic concordance, frequency proximity, and spatial localization. These objects are then grouped again to form streams. For example, an individual note can be considered an object, and a stream of these notes a musical passage. Computer based auditory scene analysis (CASA) has been attempted by a variety of researchers [86, 87]. Treatments specific to music can be found in [88, 89]. A key criticism of many CASA based approaches is that they approach the problem in a bottom-up manner. A signal is first analyzed to form a low-level time-frequency representation. Time-frequency points are then subsequently grouped together into objects and the objects fused into sources. Any error along the way leads to error propagation as higher level representations are built. Ellis proposed a top-down approach to CASA to alleviate some of these problems [90].

### 2.5.2 CASA and elementary grouping cues

CASA-based source separation from harmonic tracks involves assigning the harmonics into groups that represent their respective sources. In the case of music, a natural first step is to group harmonics into note objects, and then subsequently fuse notes into streams that potentially represent musical passages. The means by which this happens has been studied extensively in the field of auditory scene analysis. The following discussion will focus exclusively on the grouping of harmonic partials into note-objects. Therefore, only harmonic partials that are occurring simultaneously are considered. The general principle of grouping is that the harmonics of a source share many similarities in their properties and behavior over time. Usually these similarities are measured across various facets. Summary measures then combine individual facets (cues) by, for example, linear combination. Examples by which some of these similarities can be quantified are now presented.

**Common frequency modulation**

To quantify the similarity of frequency contours for harmonic grouping, Brown and Cooke [91] proposed a similarity measure. Denote two sinusoidal partials $S_i^t$ and $S_j^t$ where $t$ serves as the time (frame) index. Because a sinusoidal partial has multiple salient parameters, denote the functions $f$ and $A$ as ones that recover the frequency and amplitude of the argument, respectively. Therefore, the frequency of partial $S_i$ at time-step $t$ is expressed as $f(S_i^t)$ and its amplitude $A(S_i^t)$. The similarity between the frequency trajectories of these two tracks, $s_f\left(f(S_i), f(S_j)\right)$ can be calculated as

$$s_f\left(f(S_i), f(S_j)\right) = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} \exp\left( - \frac{\left[ \frac{f(S_i^t)}{\bar{f}(S_i)} - \frac{f(S_j^t)}{\bar{f}(S_j)} \right]^2}{2\delta_f^2} \right) \tag{2.30}$$

where $t_1$ and $t_2$ are the first and last frames that $S_i$ and $S_j$ overlap, $\bar{f}(S_i)$ and $\bar{f}(S_j)$ are the mean frequencies of the partials over the time interval $[t_1, t_2]$, and $\delta_f$ is a tolerance factor. Note that $s_f\left(f(S_i), f(S_j)\right)$ is bounded between zero (dissimilar contours) and unity (identical contours).

### Common amplitude modulation

To quantify amplitude modulation similarity, a measure identical to that used for frequency in Equation 2.30 can be used. Denoting the amplitudes of the trajectories $S_i^t$ and $S_j^t$ as $A(S_i^t)$ and $A(S_j^t)$, their average over the time span $[t_1, t_2]$ as $\bar{A}(S_i)$ and $\bar{A}(S_j)$, and a tolerance factor as $\delta_a$, the measure for amplitude similarity is expressed as

$$s_a\left(A(S_i), A(S_j)\right) = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} \exp\left( - \frac{\left[ \frac{A(S_i^t)}{\bar{A}(S_i)} - \frac{A(S_j^t)}{\bar{A}(S_j)} \right]^2}{2\delta_a^2} \right) \tag{2.31}$$

Not all musical instruments have the property that the amplitude envelopes of the harmonics share a common shape. Nevertheless, there are many instruments which indeed do have similar amplitude envelopes for all of their harmonics. Moreover, in the presence of amplitude modulation such as tremolo, the common amplitude modulation measure is useful.

### Harmonic concordance

The frequencies of the partials of pitched instruments obey a harmonic relationship. The frequency $f_i$ of the $i^{th}$ harmonic of a pitched instrument is therefore $mf_0$, where $m$ is (nearly) a positive integer and $f_0$ is the fundamental frequency of the source. If two partials from a single source are compared, $f_i$ at $m$ times the fundamental, and $f_j$ at $n$ times the fundamental, the ratio of $f_i/f_j$ is expected to approach a rational fraction $m/n$, and $m$ and $n$ are expected to to be small. A measure of harmonicity, or harmonic concordance, should favor harmonics whose frequency ratios can be expressed as simple integer ratios. Another desired property of such measures is to avoid the need for explicitly calculating the fundamental frequencies of sources. One such measure has been proposed by Virtanen and Klapuri [71]. First, a minimum frequency, $f_{min}$ is calculated as the frequency of the lowest harmonic that is found in sinusoidal tracking. Bounds are then calculated for $m$ and $n$ as

$$m = 1, 2, ..., \left\lfloor \frac{f_i}{f_{min}} \right\rfloor, n = 1, 2, ..., \left\lfloor \frac{f_j}{f_{min}} \right\rfloor \tag{2.32}$$

Finally all possible ratios of $m/n$ are calculated and the one that most closely matches the observed frequency ratio is chosen. The harmonicity error, $d_h$, is

$$d_h(f_i, f_j) = \min \left| \log \left( \frac{f_i/f_j}{m/n} \right) \right| \tag{2.33}$$

The absolute value of the log is used to equally account for ratios above and below unity. An alternative measure of harmonicity was proposed by Every and Litwic [92]. A heuristic function is created that weights common frequency ratios (e.g., 2:1, 3:1) higher than less common ones (e.g., 9:4).

### 2.5.3 Synthesis of separated sources and binary masks

Source separation requires the synthesis of sounds separated from mixtures. The techniques used for synthesis depend greatly on the signal representation used in separation. If, for instance, each source is represented as a group of sinusoidal tracks, additive synthesis [93] can be employed to synthesize the separations. However, lower level time-frequency representations such as the STFT or cochleagrams are perhaps even more prevalent. A common engine for separation and synthesis that is prevalent in CASA-based and multichannel-based source separation techniques is the *binary mask*.

The definition of a binary mask perhaps most simply begins with the definition of the *ideal* binary mask (IBM). Ideal binary masks are constructed with full knowledge of the individual source signals. The ideal binary mask aims to assign time-frequency points where a source is dominant to that source. Assuming a signal is composed of two sources, $x_1[n]$ and $x_2[n]$, with time-frequency representations for time-frame $t$ and frequency bin $k$, $X_1[t, k]$ and $X_2[t, k]$, the ideal binary mask for each respective source, $M_1[t, k]$ and $M_2[t, k]$, can be expressed as

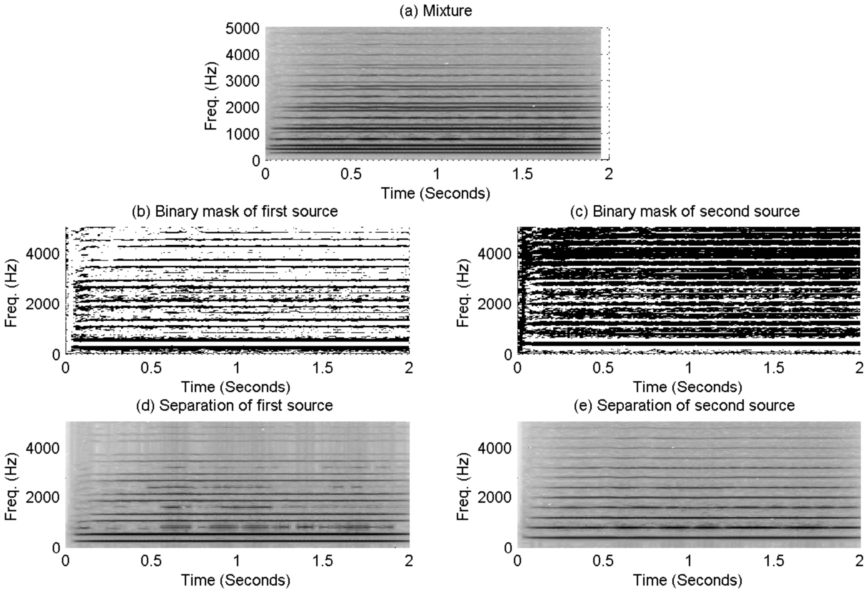

Figure 2.8: Example of ideal binary masks applied to a mixture of two tones. The spectrogram of the mixture is shown in (a). The binary masks of the two sources are shown in (b) and (c). Black points denote "1" and white points denote "0." The spectrograms of the separated sources are shown in (d) and (e).

$$\begin{aligned} M_1[t,k] &= \begin{cases} 1 & \text{if} \quad 20\log_{10}(X_1[t,k]) - 20\log_{10}(X_2[t,k]) > \Theta \\ 0 & \text{else} \end{cases} \\ M_2[t,k] &= \begin{cases} 1 & \text{if} \quad 20\log_{10}(X_2[t,k]) - 20\log_{10}(X_1[t,k]) > \Theta \\ 0 & \text{else} \end{cases} \end{aligned} \tag{2.34}$$

In most definitions, the threshold parameter, $\Theta$, is zero. At this threshold, each time-frequency point is assigned to only one source. However, softer thresholds such as $\Theta = -6$ have been shown to produce better results [94]. In this case, a time-frequency point can potentially be assigned to both sources if the corresponding time-frequency points of the individual sources are close in amplitude. Synthesis of a source can be achieved by applying the mask to the mixture (an element-by-element product of the mask and mixture) and using STFT-synthesis.

Figure 2.8 demonstrates the application of an ideal binary mask to a mixture of two tones at a perfect-fifth interval. The separations show that one of the sources is apparently missing harmonics. This is due to the fact that the other source is dominant in the mixture. Therefore, all energy at those corresponding time-frequency points is assigned to the other source. As a result, the dominant source has significant contributions of the other source present after separation. This phenomenon demonstrates the most common types of distortions in source separations when time-frequency overlaps (harmonic collisions) are present: distortions due to missing harmonics, and distortions due to interference of other sources.

Naturally, source separation with prior and perfect knowledge of the sources to be separated seems to be an endeavor with little point. However, IBMs have value in the fact that they provide good baselines of comparison in the evaluation of source separation results. Moreover, while the aim of source separation is to recover individual sources perfectly, a main goal of CASA can be reformulated so as to rather recover the ideal mask for each source [95]. Non-ideal binary masks can be constructed based on time-frequency trajectories (e.g., sinusoidal tracks) that have been grouped according to CASA principles. Many of the previously presented multichannel methods exploit interchannel characteristics of time-frequency points to generate time-frequency masks directly. Note that with ideal masks, time-frequency points are assigned fully to individual sources. Therefore, even in the case of ideal binary masking, time-frequency points containing the contributions of two or more sources will cause distortions. Source separation techniques that use templates or basis vectors such as the previously discussed non-parametric and parametric-model based methods to drive synthesis do not suffer from this problem.

### 2.5.4 Evaluation measures of source separation performance

The evaluation of audio source separation algorithms presents a key challenge: quantifying a subjective sound quality. The most common measure of separation performance is the signal-to-distortion ratio (SDR), which can be expressed as

$$SDR_{dB} = 10\log_{10}\frac{\sum_n x[n]^2}{\sum_n (x[n]-\hat{x}[n])^2} \tag{2.35}$$

where $x[n]$ and $\hat{x}[n]$ are the original source and separated signals, respectively.

A phase-sensitive measure of SDR for audio operates on the magnitude spectrum. Denoting the $l^{th}$ bin of the $t^{th}$ frame of $x[n]$ and $\hat{x}[n]$ as $X[t,l]$ and $\hat{X}[t,l]$, the spectral error ratio (SER) can be expressed as

$$SER_{dB} = 10\log_{10}\frac{\sum_{t,l}|X[t,l]|^2}{\sum_{t,k}(|X[t,l]| - |\hat{X}[t,l]|)^2} \tag{2.36}$$

The spectrum based approach is less sensitive to relative phase offsets that are usually perceptually irrelevant.

Overviews of evaluation techniques can be found in [96] where concerns of matching separated sources to the ground truth originals are also addressed. Perceptual measures such as those found in [97] and [98] process the signals with an auditory model and carry out the measures in the auditory domain. When separation is used as a front-end for transcription, accuracy can be used as a measure of performance [99]. Vincent *et al.* [100] add more low-level measures to the standard SDR by also measuring the interference and effects of other sources in the separated signals (signal to interference ratio, SER), separation artifacts (signal to artifact ratio, SAR), and spatial filtering distortion for multi-channel separation. An implementation to evaluate source separations using these measures is freely available [101]. Both SDR and SER have been used as the evaluation measures in recent source separation evaluation campaigns [102, 103].

## 2.6 Summary and Discussion

This chapter presented an overview of some of the goals, techniques, and challenges faced in musical audio signal processing. The concept of overlapping partials, or harmonic collisions, was explored from both a signal processing and musicological perspective. The treatment of harmonic collisions showed that, in short spans of time, the parameters of closely spaced sinusoids cannot be trivially extracted by simple inspection of the magnitude spectrum. Moreover, it was demonstrated that these harmonic collisions exist on the order of 50% of the time for a given musical source in a musical mixture. A high-resolution signal-subspace method, called the direct matrix pencil method, that has the ability to resolve the parameters of closely-spaced sinusoids, was introduced. However, this technique requires that the frequency of the sinusoids are stationary over the span of time it operates over. The reason harmonic collisions are a challenge in many domains of music signal processing is that they hinder the ability to infer information about individual sources in mixtures. Two example applications where these difficulties are evident, multi-$f_0$ estimation and musical audio source separation, were presented. Simple examples that show how harmonic collisions play a role in these application domains were demonstrated. The focus of this thesis now turns to this question: Do signal subspace-based sinusoidal parameter estimators have the potential to resolve harmonic collisions in music audio signals?

# CHAPTER 3

# SHORT-TIME HIGH-RESOLUTION SINUSOIDAL ANALYSIS

This chapter presents an implementation of a short-time high-resolution (STHR) sinusoidal analysis system. The system is built around the direct matrix pencil method for extracting sinusoidal parameters. Chapter 2 demonstrated that the direct matrix pencil method has the potential to resolve closely spaced sinusoids. By itself, however, the matrix pencil method does not support sinusoids with time-varying parameters. Therefore, the matrix pencil method is adapted here to operate on small time windows of the signal within which sinusoidal parameters are assumed stationary. Extracted sinusoidal estimates for each frame are linked across frames to generate sinusoidal tracks. The ultimate goal of this sinusoidal analysis system is for the system to be able to produce individual tracks of sinusoidal partials that are potentially very closely spaced. In traditional implementations of sinusoidal tracking based on the short-time Fourier transform, closely spaced sinusoids are merged into a single track. The usefulness of this form of super-resolution sinusoidal tracking is explored in later chapters.

Section 3.1 presents the proposed sinusoidal analysis system. Design considerations involving the individual components of the system are covered. Section 3.2 evaluates the sinusoidal analysis system for a range of input signals. The system is evaluated with synthetic signals that perfectly fit the model used by the direct matrix pencil method. The evaluations using synthetic signals serve to demonstrate the effects of the processes surrounding the matrix pencil estimation of sinusoidal parameters. The system is also evaluated on real musical instrument tones and tone mixtures. The ability of the system to produce a valid sinusoidal representation and its ability to detect collided harmonics are evaluated.

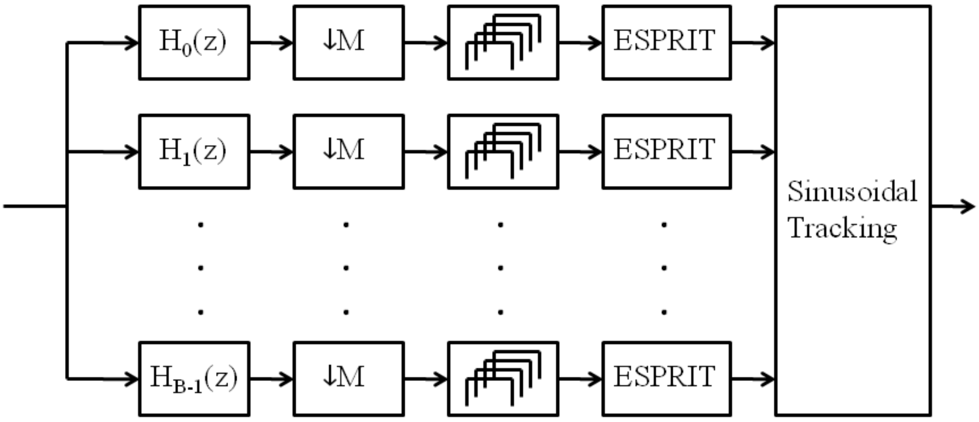

Figure 3.1: Block diagram of short-time high-resolution sinusoidal analysis. The signal is filtered into sub-bands. Each sub-band is decimated and windowed into overlapping frames with a rectangular window. The sinusoidal parameters within each band are extracted using ESPRIT and then sinusoidal tracking is performed to build partial tracks.

## 3.1 Sinusoidal Analysis System Overview

An overview of the short-time high-resolution sinusoidal analysis system is shown in Figure 3.1. The system consists of a filter bank that decomposes the signal into sub-bands. Each sub-band is downsampled and windowed with overlapping rectangular windows. For each window of each sub-band, direct matrix pencil, or ESPRIT, sinusoidal estimation is performed. The result of this analysis is a time-frequency representation comprising estimates of the parameters of prominent sinusoidal partials present in the signal through time. These individual estimates are then grouped with a tracker to form individual sinusoidal tracks.

This section covers each aspect of the sinusoidal analysis system in detail. Section 3.1.1 covers the preprocessing portion (filter bank, downsampling, and windowing). Section 3.1.2 explains details regarding the ESPRIT estimation and challenging issues that frequently arise. Finally, Section 3.1.3 introduces a method for tracking sinusoidal partials through time from the pole estimates generated by ESPRIT for each window.

### 3.1.1 Filter bank, downsampling, and windowing

The primary motivation for performing direct matrix pencil sinusoidal analysis on subsampled sub-bands is to greatly reduce computational load. Recall from Chapter 2 that ESPRIT is largely dependent on computationally expensive matrix operations such as singular value and eigenvalue decompositions. These matrices are dependent on the length of the signal and the order of the sinusoidal model (i.e., how many sinusoids are being estimated). Decomposing the signal into subsampled sub-bands breaks down the problem of extracting all sinusoids in a signal into a series of smaller subproblems. Operating on sub-bands allows the system to search for a smaller number of sinusoids in each band than would be necessary for the entire broadband signal. Downsampling produces a representation of the signal for a given span of time with a smaller number of samples. The filtering and downsampling operations make the problem more tractable.

In addition to reduced computational load, Tkacenko showed that there are two additional benefits to performing ESPRIT sinusoidal analysis on downsampled sub-bands of the signal [104]. First, when the sinusoids are in the presence of colored noise, the noise appears flattened within each narrow sub-band. Algorithms such as ESPRIT are largely dependent on the noise subspace being white (uncorrelated), and therefore, a form of whitening is carried out by sub-band filtering. In addition, the subsampling of each band effectively widens the separation of sinusoidal components, aiding in the resolution of closely spaced sinusoids.

The design of the filter bank raises important design considerations in regards to interactions with subsequent ESPRIT sinusoidal analysis. Noteably, the nature of the filters and the subsequent downsampling raise some significant concerns. Assume that the bank of $B$ filters in Figure 3.1 are of equal bandwidth and linearly spaced, and critical downsampling is performed on each band (the downsample factor, $M$, is equal to $B$). With non-ideal bandpass filters, the band edges at the lower and upper cutoffs represent only a -6 dB attenuation point. Therefore, if a sinusoid exists just outside the band edge for a filter, the attenuation might not be significant enough to fully suppress it. After downsampling, this sinusoid will alias and its sinusoidal parameters will be extracted. The result is a spurious sinusoidal partial at the wrong frequency.

To counteract aliasing effects, first observe that the underlying signal model used by the direct matrix pencil method is composed of complex sinusoids. Real sinusoids always generate a pair of complex conjugate poles. Therefore, to extract sinusoidal parameters using the matrix pencil method, only half of the spectrum (positive or negative frequency) needs to be analyzed. Thus, there is no restriction that the filter bank must be real. A complex filter bank can be designed by modulating a prototype lowpass filter with complex sinusoids to center each filter at its desired location and have coverage over only positive frequencies. With no negative frequency component for each bandpass filter, in-band aliasing after decimation is greatly suppressed. The lesser degree of in-band aliasing also relaxes the necessity for high-order sharp-cutoff filters. In addition, the number of sinusoids to be searched for in each band is effectively halved as the order of the model does not need to account for both the negative and positive frequency components of real sinusoids.

Figure 3.2 demonstrates aliasing effects for both real and complex filters. In this particular example, the filter bank is composed of four bands. The first subplot shows the ideal (real) filter bank. Subsequent subplots demonstrate downsampling of the non-ideal third band (index 2) for real and complex bandpass filters. Whereas the real filter displays a great deal of aliasing, as seen in Figure 3.2(c), the corresponding complex filter shows virtually no in-band aliasing, as shown in Figure 3.2(e). While spurious partials will still be analyzed and extracted with the direct matrix pencil method, these can be trivially pruned out. In the case of this particular band (the index of the filter is even), all extracted sinusoids that have negative frequency can be rejected prior to sinusoidal tracking. When the index of a filter is odd, the band of interest occupies the negative frequency portion after downsampling. This effect is shown in Figure 3.2(f). Therefore, extracted poles with positive frequency can be pruned for odd-indexed filters. The filter bank is implemented using a 512 order finite-impulse response linear-phase lowpass filter as the prototype. For a $B$-band, linearly-spaced filter bank, the prototype lowpass filter has bandwidth $\pi/(2B)$. The prototype filter, $h[n]$, is designed using the window design method with a Hamming window. The resultant group delay is 256 samples. The complex bandpass filter for band $b$, $h_b[n]$, is

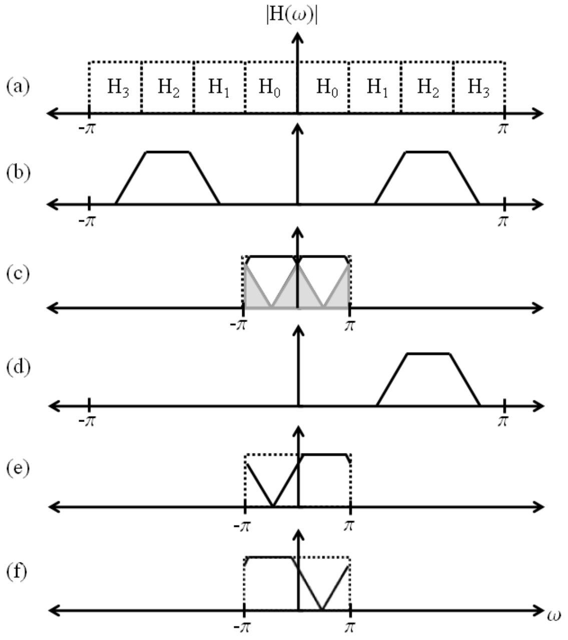

Figure 3.2: An example filter bank and the effects of downsampling. An ideal filter bank is shown in (a). A non-ideal filter corresponding to the third band (index 2) is shown in (b). The resultant downsampling and aliasing effects of the band in (b) are shown in (c). A complex bandpass filter for the third band is shown in (d). The resultant downsampling effects of the filter in (d) are shown in plot (e). Plot (f) shows the effects of downsampling on complex bands that have odd indices.

produced by modulating the prototype as follows:

$$h_b[n] = e^{j\pi\left(\frac{1}{2B}+\frac{b}{B}\right)n}h[n] \tag{3.1}$$

A linear phase prototype is used to impart an equal delay on all sinusoidal partials.

Figure 3.3 shows the individual magnitude responses and the net magnitude and phase response of a 16-band complex filter bank used to analyze signals sampled at a rate of 22.05 kHz. Extracted sinusoidal frequencies must be translated from those estimated in each subsampled sub-band signal to their true locations. For a bank of $B$ linearly spaced complex filters downsampled by a factor of $B$, denote the lower cutoff frequency (in Hz) of filter $b$, $f_L^{(b)}$, and the upper cutoff frequency $f_H^{(b)}$. The original sample rate of the signal is $f_s$. An extracted pole frequency, $\omega$ (in radians), from band $b$ can be

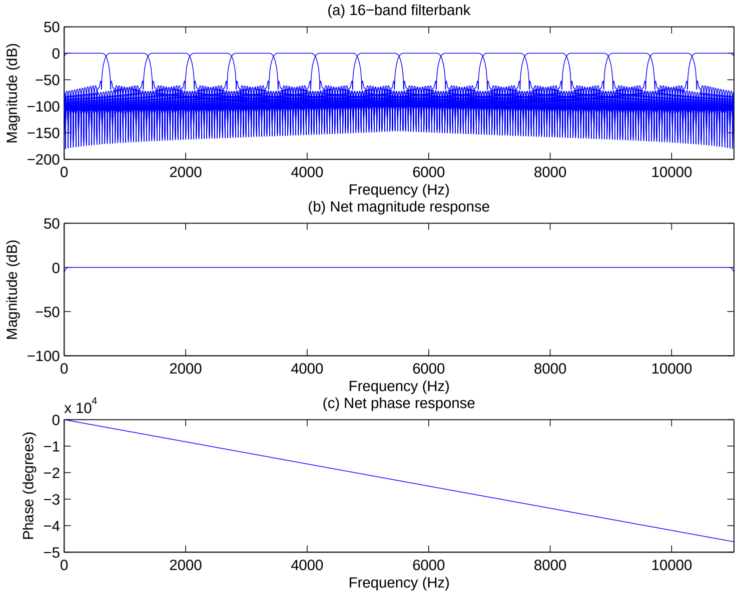

Figure 3.3: A 16-band complex filter bank. The magnitude responses of the individual filters are shown in (a). The net magnitude and phase responses are shown in (b) and (c) respectively.

translated to its true frequency, $f$ (in Hz), as

$$f = \begin{cases} \frac{\omega f_s}{2\pi B} + f_L^{(b)} & \text{for b even} \\ \frac{\omega f_s}{2\pi B} + f_H^{(b)} & \text{for b odd} \end{cases} \tag{3.2}$$

In addition to translating the frequencies, the effect of the magnitude response of the filter must be accounted for. This is especially true for harmonics that reside near the cutoff frequencies of the filters and are potentially attenuated by as much as 6 dB. The magnitude response of the prototype filter is stored in a lookup table with 0.1 Hz resolution. Each extracted harmonic amplitude is then divided by the stored magnitude response of the filter closest to that particular frequency. Because each filter is designed to have unit magnitude response, all extracted sinusoidal amplitudes must also be doubled (as only the positive frequency portion of a real sinusoid has been estimated).

Following filtering and downsampling, each band is windowed with overlapping rectangular windows. Depending on application, typical window lengths used are 46 ms or 93 ms. Assuming a 16-band downsampled filter bank, operating on a 22050 Hz sampled signal, the window lengths corresponding to 46 ms and 93 ms are 64 and 128 samples, respectively. A hop size equivalent to 1/8 the window length is used (87.5% overlap). Such a large overlap between adjacent frames may seem unnecessary as it greatly increases the amount of computation required. However, the ultimate goal of the sinusoidal analysis system is to track closely spaced sinusoidal partials. A relatively small hop size ensures that the parameters of each sinusoid do not vary greatly between successive frames. Relatively small changes in the sinusoidal parameters greatly aid in linking the pole estimates of each frame to the pole estimates of the next. Note that such a high overlap is only required when the signal is expected to have harmonic collisions. For the analysis of isolated tones or monophonic passages, the amount of frame overlap can be reduced thus reducing computational load.

### 3.1.2 ESPRIT, model order, and regularization

Direct matrix pencil/ESPRIT estimation is performed on each windowed frame of each sub-band of the preprocessing stage of Section 3.1.1. The details of the direct matrix pencil method sinusoidal estimation were covered extensively in Section 2.2.4. The main free parameters of matrix pencil estimation are the pencil parameter, and the model order. Model order refers to the number of sinusoids whose parameters are to be estimated. As previously stated, good values of the pencil parameter are $N/3$ to $N/2$ with $N$ being the length of the signal (window). The sinusoidal analysis system presented here uses a pencil parameter of $N/2$.

In general, the model order is not known *a priori*. To understand the effects of model order on a system, assume a signal is composed of $K$ sinusoids, and a model order of $\hat{K}$ is chosen. If $\hat{K} > K$, the true $K$ sinusoids and their parameters are guaranteed to be a subset of the $\hat{K}$ estimates. However, if the model order is underestimated, i.e. $\hat{K} < K$, there is no guarantee that the estimated $\hat{K}$ sinusoids match any of the true $K$ sinusoids. Thus, it is obvious that overestimation of model order is far superior to underestimation. The overestimation of model order does not come without some cost, however. While the true poles will be a subset of the $\hat{K}$ estimated poles, recall that the poles are defined only by the frequencies and damping factors of the sinusoids. The amplitudes and phases of the sinusoids are solved for as a least-squares projection of the signal onto the subspace spanned by the $\hat{K}$ poles. Over a finite time-support, sinusoids are only orthogonal if they are harmonics of the inverse period of the window length (e.g., the DFT basis). Since the $\hat{K}$ pole estimates are likely not orthogonal, some of the energy of the true $K$ sinusoids will be projected onto the remaining $\hat{K} - K$ erroneous sinusoids. If these erroneous sinusoids lie in close proximity to true sinusoids, the amount of energy they capture from the projection will be non-trivial, and they will resultantly have significant amplitude. The hope is that such spurious estimates, even if they have significant amplitude, will not display the sort of temporal continuity required to form sinusoidal tracks with slowly time-varying parameters.

Estimation of model order has been extensively researched. Both [105] and [106] provide good overviews of model order selection techniques. Notwithstanding the fact that overestimation of model order can skew amplitude estimates slightly, the proposed sinusoidal analysis system uses the maximum model order supported in each sub-band. As previously stated, overestimation is far less catastrophic than underestimation. Therefore, the system is not at the mercy of potentially underestimating model order with one of the existing order selection techniques. For a 16-band filter bank, operating on 22.05 kHz sampled signals and a 46 ms window size, 15 sinusoids are searched for in each band.

Recall once again that sinusoidal amplitudes and phases are determined from the estimated poles using a least-squares projection. As is frequently the case with inverse problems, there is the potential that the calculation of the amplitudes and phases is numerically sensitive. When estimated poles are close together, the condition number of the matrix $Z_p$ in Equation 2.27 can become large. Because real world signals do not fit the underlying signal model exactly (e.g. their harmonic frequencies do indeed slightly vary within a window, the harmonic amplitudes are not perfectly exponentially decaying, or the signal is in the presence of colored noise), there is a slight perturbation of the signal from the estimated model. The chief characteristic of ill-conditioned systems is that small perturbations or errors in the model can lead to very large perturbations of the resulting estimates, in this case, the harmonic amplitudes. These problems manifest themselves in the analysis or real musical mixtures as the occurrence of large “explosions” of amplitudes, sometimes by orders of magnitude. Such “explosions” can interfere with sinusoidal tracking, and if present in the output resynthesis of the signal, can cause disturbing audio artifacts.

The regularization of inverse problems, specifically Tikhonov regularization [107], provides a means for counteracting “exploding” amplitudes. Repeating once again Equation 2.27 for posterity, recall that the complex harmonic amplitudes contained in vector $\mathbf{r}$ can be recovered from the signal $\mathbf{x}$ (in vector form) and the matrix $Z_p$ (containing the $K$ estimated poles) as

$$\mathbf{r} = (Z_p^H Z_p)^{-1} Z_p^H \mathbf{x} \tag{3.3}$$

This least-squares solution is the one that minimizes the residual $\|Z_p\mathbf{r} - \mathbf{x}\|_2^2$. The most basic form of Tikhonov regularization adds an additional term to instead minimize

$$\|Z_p\mathbf{r} - \mathbf{x}\|_2^2 + \|L\mathbf{r}\|_2^2 \tag{3.4}$$

If $L = \alpha I$, with $I$ being the identity matrix, Tikhonov regularization favors solutions with smaller norms. The weight factor $\alpha$, known as the regularization parameter, balances the costs between minimizing the least-squares residual and the $l^2$ norm of the solution $\mathbf{r}$. The closed-form regularized least-squares solution to estimate the complex sinusoidal amplitudes becomes

$$\mathbf{r} = (Z_p^H Z_p + L^H L)^{-1} Z_p^H \mathbf{x} \tag{3.5}$$

Because estimating extremely large harmonic amplitudes is penalized through regularization, the amplitude explosion problem is largely mitigated.

The choice of the factor $\alpha$ that balances the residual cost and solution norm cost now becomes a free parameter. A host of methods such as the L-curve method exist for choosing an appropriate regularization parameter. However, such methods look to choose an ideal parameter for each specific least-squares problem. In the case of the short-time sinusoidal analysis system, there are hundreds and potentially thousands of least-squares problems being solved for every audio file. The tuning of the regularization parameter for each least-squares problem becomes intractable. Therefore, some form of a global estimate for the regularization parameter is desired. An optimization experiment for choosing a regularization parameter is presented in Section 3.2.

### 3.1.3 Sinusoidal tracking based off ESPRIT pole estimates

Estimates of sinusoidal parameters are produced for each frame of each subband using the direct matrix pencil method. After the frequency estimates are translated to their true locations via Equation 3.2 and the effects of the the magnitude response of each bandpass filter compensated for, sinusoidal estimates are linked frame to frame through the use of a sinusoidal tracker. At the frame level, the representation produced by ESPRIT is largely equivalent to the representations used by STFT-based trackers such as the MQ method presented in Section 2.2.3 (which operate off harmonic estimates derived from peak picking the DFT of each frame).

Because a large model order is used in the matrix pencil estimation of sinusoidal parameters, many extracted poles will be erroneous. In essence, the signal subspace will capture portions of the noise subspace, and some poles will serve to capture aspects of the noise. Examination of the estimated damping factor for the poles can be used to prune some of the erroneous pole estimates. It is expected that moderately stable sinusoids undergo relatively little amplitude change within a frame, and thus have damping factors close to zero. For a 16-band filter bank with an effective frame size of 46 ms corresponding to 64 samples, a pole with a damping factor with absolute value 0.05 would undergo a 27.8 dB change in amplitude during the frame's durations. Therefore, all poles with damping factors $|\alpha| > 0.05$ are immediately removed from consideration.

Traditional MQ tracking links sinusoidal estimates based solely on proximity in frequency. In essence, smooth frequency trajectories are enforced by only allowing linkages of sinusoidal estimates that are closely spaced in frequency between frames. In the case of high-resolution analysis where there is the potential that sinusoidal estimates exist closely spaced in a given frame (i.e., harmonic collisions are resolved), producing track linkages based solely on frequency can be error prone. If two source harmonics are close in frequency and perhaps cross at some point, there is the potential that the existing tracks corresponding to each source harmonic could switch and begin tracking the harmonic of the other source. Therefore, frequency proximity between an existing track and the subsequent pole estimates is a necessary but insufficient condition. Disambiguation of closely spaced tracks can be carried out by also enforcing smoothness of the harmonic amplitudes of each track. Recall that the primary justification for a high overlap factor between successive frames in the STHR analysis is to ensure that the key sinusoidal parameters, namely frequency and amplitude, do not very greatly frame to frame. The sinusoidal tracker used in the STHR sinusoidal analysis system operates on both harmonic frequencies and amplitudes to produce partial tracks.

The STHR sinusoidal tracker works as follows. Denote the sinusoidal tracks that are currently active at frame $t$: $S_1^t$, $S_2^t$, ..., $S_M^t$. The goal is to produce linkages of the $M$ active tracks to the $K$ extracted poles of frame $t+1$. Denote the $K$ poles at frame $t+1$: $z_1^{t+1}$, $z_2^{t+1}$, ..., $z_K^{t+1}$. Let the functions $f$ and $A$ denote the frequency and amplitude of the argument, respectively. Therefore the frequency of track $S_m^t$ is $f(S_m^t)$ and the amplitude corresponding to pole $z_k^{t+1}$ is $A(z_k^{t+1})$. There are a total of $T$ analysis frames and thus $t \in [0, T-1]$. Tracks are built in the following step-by-step manner.

**Step 1.** For each track, $S_m^t$, currently active at time $t$, find all pole candidates, $z_k^{t+1}$, in frame $t+1$ that are in close proximity in frequency to track $S_m^t$. That is, find all $z_k^{t+1}$ such that $\left|f(S_m^t) - f(z_k^{t+1})\right| < \Delta f_{max}$. All $z_k^{t+1}$ that fit this criterion (proximity in frequency) form a set of potential matches for track $S_m^t$. Denote the set of potential match candidates for track $S_m^t$, $C_m^{t+1}$. All $S_m^t$ that have no candidate matches are allowed to die.

**Step 2.** For track $S_m^t$ and candidate set $C_m^{t+1}$, find the $z_k^{t+1} \in C_m^{t+1}$ that minimizes amplitude difference. That is, find the $z_k^{t+1}$ that minimizes $\Delta A_{dB} = \left|20\log_{10} A(S_m^t) - 20\log_{10} A(z_k^{t+1})\right|$. If $\Delta A_{dB} < \Delta A_{max}$, produce a temporary linkage between track $S_m^t$ and pole $z_k^{t+1}$. If there exists no $z_k^{t+1} \in C_m^{t+1}$ that provides smooth amplitude continuity (i.e. there is too large a jump in amplitude to all candidate poles), allow $S_m^t$ to die.

**Step 3.** Repeat the first two steps for all $m \in [1, M]$ to account for all existing tracks.

**Step 4.** For all poles at time step $t+1$ that are uniquely assigned to one track at time $t$, make the temporary linkage permanent. All poles that have temporary linkages to more than one track must now be uniquely assigned. For a pole $z_k^{t+1}$ assigned to more than one track, a cost function, $J(z_k^{t+1}, S_m^t)$, that equally weights frequency and amplitude deviation between it and track $S_m^t$ is calculated as $J(z_k^{t+1}, S_m^t) = \left|\log\left(f(z_k^{t+1})/f(S_m^t)\right)\right| + \left|\log\left(A(z_k^{t+1})/A(S_m^t)\right)\right|$. The track $S_{\hat{m}}^t$ that minimizes $J$ is chosen as the proper linkage since it provides the closest match in terms of both frequency and amplitude. For all other tracks $S_m^t$ that had $z_k^{t+1}$ as a temporary linkage, remove $z_k^{t+1}$ from their respective candidate pools $C_m^{t+1}$, and go to Step 2. This process will determine if there are other potential viable candidates for tracks that had temporary linkages broken.

**Step 5.** All remaining poles that have not been accounted for and linked to a track must now be handled. Direct matrix pencil estimation has the potential to occasionally not report a pole in a given frame. In addition, even though the least-squares estimation of harmonic amplitudes is regularized, the potential for a gross error should also be considered. Therefore, all tracks that had deaths at time step $t-1$ are now considered as potential matches for unmatched poles in time step $t+1$. Remember that the hop size between frames is relatively small, and therefore, linkages between frames two time steps apart are still relevant. Steps 1-4 are repeated for all tracks that died at step $t-1$ and for all unaccounted for poles at step $t+1$. In the case where a gross amplitude error occurred due to numerical instability in the least-squares estimation of amplitudes, this additional tracking procedure also serves as a form of regularization. If a match is produced, estimates of the frequency and amplitude at time $t$ are produced by linearly interpolating between the track parameters at step $t-1$ and the matching pole estimates at time step $t+1$. All poles that have linkages to tracks at frame $t$ or tracks that died at track $t-1$ produce births of new tracks.

**Step 6.** $t = t + 1$. If $t < T$ go to Step 1.

**Step 7.** All tracks that are of duration less than $T_{min}$ are pruned. Short-lived tracks are likely due to spurious pole estimates that are fit to noise.

Figure 3.4 shows the pole estimates and resulting sinusoidal tracks of a mixture of two oboe tones an octave apart. It is evident that many spurious poles have been rejected due to tracking. Only poles that can be attributed to

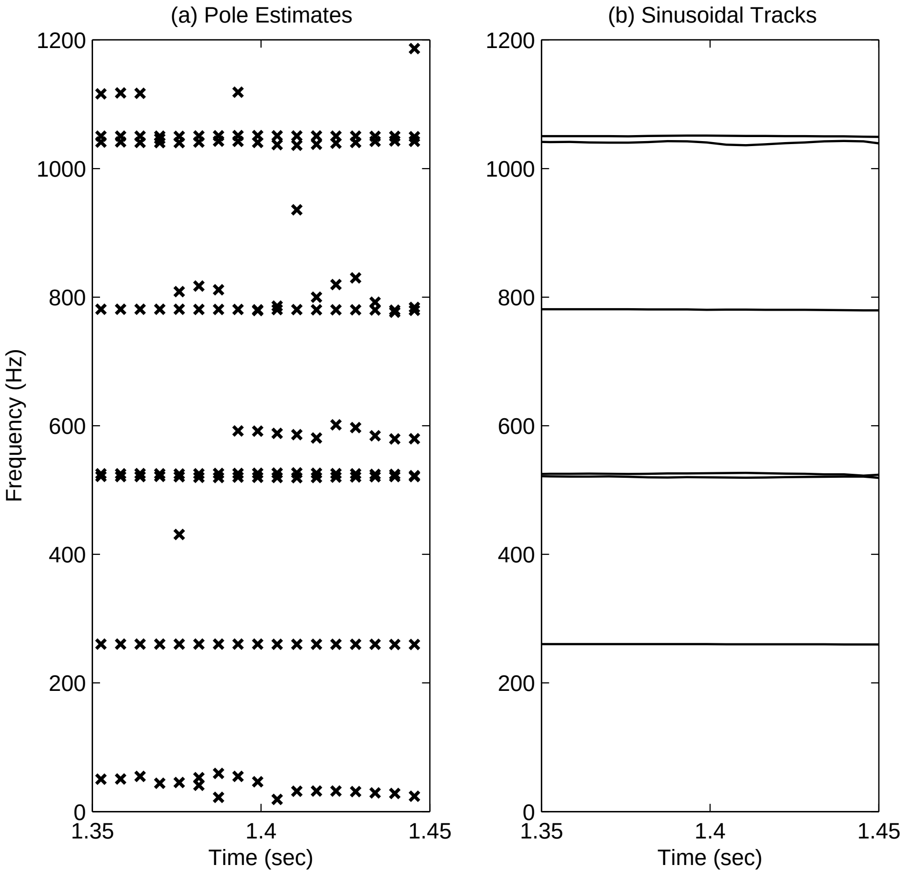

Figure 3.4: Derived pole estimates (a) and sinusoidal tracks (b) of a mixture of two oboe tones an octave apart.

actual harmonics of the signal maintain the necessary continuity and smoothness of parameters to produce viable sinusoidal tracks. This behavior is crucial due to the fact that the model order is generally overestimated in each sub-band for the ESPRIT analysis.

It is important to note that potential difficulties can arise in the tracking procedure if two closely spaced harmonics are also very close in amplitude. This ambiguity is largely unresolvable and represents the types of challenges faced in any bottom-up approach where objects are formed from very low level representations. Nevertheless this tracking procedure does serve to better track closely spaced sinusoids than traditional MQ tracking. The key free parameters are the maximum allowable frequency deviation, $\Delta f_{max}$, the maximum allowable amplitude deviation, $\Delta A_{max}$, and the minimum track length, $T_{min}$. The maximum allowable frequency deviation is set to $\Delta f_{max} = 0.01 f(S_m^t)$ (i.e., one percent of the track frequency). Although this represents a relatively small deviation, recall that the effective hop size used in most analyses is 5.8 ms. The maximum allowable amplitude deviation is $\Delta A_{max} = 6$ dB. All tracks of duration less than 10 frames are pruned ($T_{min} = 10$).

## 3.2 Experiments and Evaluation

This section presents the evaluation of the short-time high-resolution sinusoidal analysis system across a range of input signals. The ability of the method to correctly determine the sinusoidal parameters of synthetic signals is tested in Section 3.2.1. The synthetic signals are ones that perfectly match the underlying model used by the direct matrix pencil method. Section 3.2.2 covers a second experiment meant to aid in choosing a potential regularization parameter. Evaluations of whether STHR analysis provides valid sinusoidal representations for real-world musical tones and mixtures is explored in Section 3.2.3. Finally, a preliminary test of the system's ability to resolve collided harmonics is tested in Section 3.2.4.

### 3.2.1 Synthetic signals

The first test of the sinusoidal analysis system involves evaluating its performance on synthetic signals that perfectly fit the underlying model used by the direct matrix pencil method. The purpose of the test is to evaluate what effects the various preprocessing stages, namely the filter bank and downsampling, have on estimating the sinusoidal parameters. The signal used in this test is a bandlimited pulse train additively synthesized as

$$x[n] = \sum_{k=1}^{K} \cos(2\pi k f_0 n / f_s) \tag{3.6}$$

where $f_0$ is the fundamental frequency, $f_s$ is the sample rate, and $K$ is chosen to be the maximum value such that $K f_0 < f_s/2$. The fundamental frequency is varied at 1 Hz increments from 65 Hz (C2) to 2093 Hz (C7).

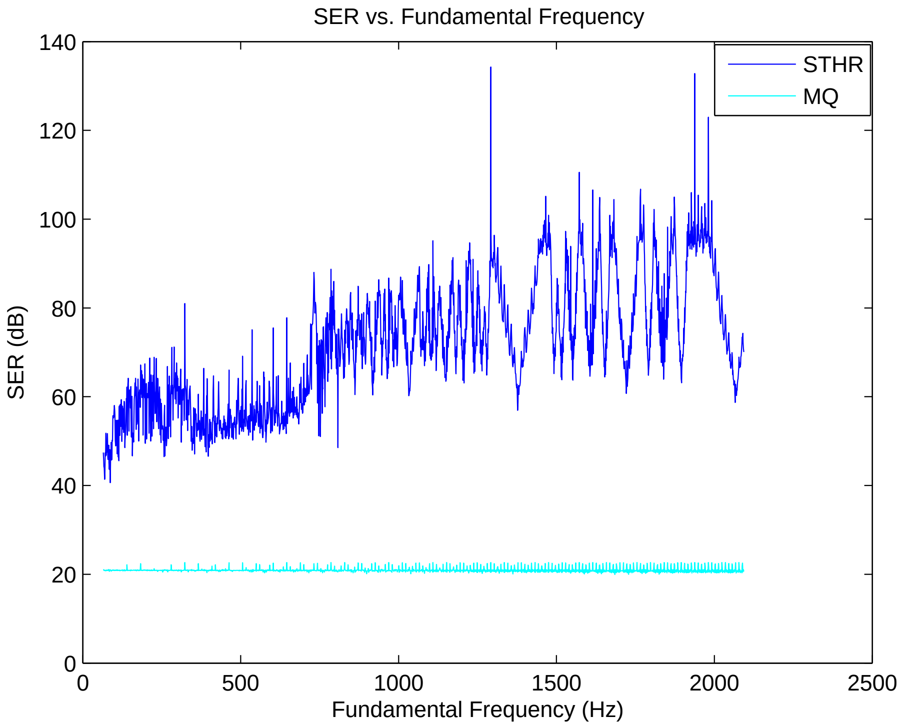

Figure 3.5: SER performance of STHR and MQ analysis synthesis on bandlimited pulse trains with varying fundamental frequency.

The bandlimited pulse trains are analyzed using STHR sinusoidal analysis. A 16-band filter bank and 46 ms frame size are used. The signals are resynthesized from the sinusoidal parameters and evaluated against the original signal using the spectral error ratio (SER) measure found in Equation 2.36. In addition, the signals are analyzed and resynthesized from STFT-based MQ analysis. Once gain, a frame size of 46 ms is used for direct comparison.

Figure 3.5 shows the SER of both STHR and MQ analysis as a function of fundamental frequency. It is evident that while the pulse train signal of Equation 3.6 perfectly matches the underlying model used by ESPRIT, and therefore should be perfectly analyzed and resynthesized, the various processing surrounding ESPRIT in STHR analysis prevents perfect reconstruction. The causes that prevent perfect reconstruction are easily explainable. Recall that the filters used in the analysis are non-ideal. Although the use of complex filters greatly reduces in-band aliasing, there is nevertheless a presence of aliased components in each subband signal. In effect, the signal-to-noise in each subband is no longer infinite. With this manner of noise present in the signal, estimates of the sinusoidal parameters are prone to some manner of error. Because low tones generate dense line spectra, synthetic tones with lower fundamental frequencies have a lower effective signal-to-noise ratio. As a result, there is, on average, increasing reconstruction performance with increasing fundamental frequency. The increased performance at higher fundamental frequencies is also due to the fact that these spectra are more sparse, and therefore the accumulated error is lessened. In addition, the zeros of the filters are evenly spaced around the unit circle. Some of the harmonics of the synthesized signals may or may not have harmonics that align themselves with zeros of the filters producing varying signal-to-noise ratios. This explains the spiked performance response seen in Figure 3.5. In some cases, well aligned zeros have the potential to increase effective signal-to-noise ratio. Moreover, the compensation of the magnitude response of each subband filter is not perfect. Recall that response effects are compensated for through the use of a lookup table.

Despite the fact that STHR sinusoidal analysis does not produce a perfect reconstruction of the signal, it manages to produce a very accurate representation. Surprisingly, STFT-based sinusoidal analysis does not provide as accurate a representation. Figure 3.6 shows the true line-spectrum of a pulse train signal ($f_0$ = 220 Hz), along with the estimates for a single frame of the signal for both STHR and MQ analysis. In these examples, quadratic interpolation of the log-magnitude spectra are used to refine frequency and amplitude estimates of the sinusoidal partials. However, this manner of interpolation does not produce as accurate a result as the estimates produced by STHR analysis. Moreover, cross talk due to the DFT of the analysis window used also influences the sinusoidal estimates in STFT-based analysis.

The analysis of synthetic signals provides insight into how the various components of the STHR sinusoidal analysis interact. However, these signals are largely overly simplistic, and it is expected that the analysis system will perform well on them. Therefore, the ability of the system to accurately represent real-world signals becomes the focus of subsequent experiments.

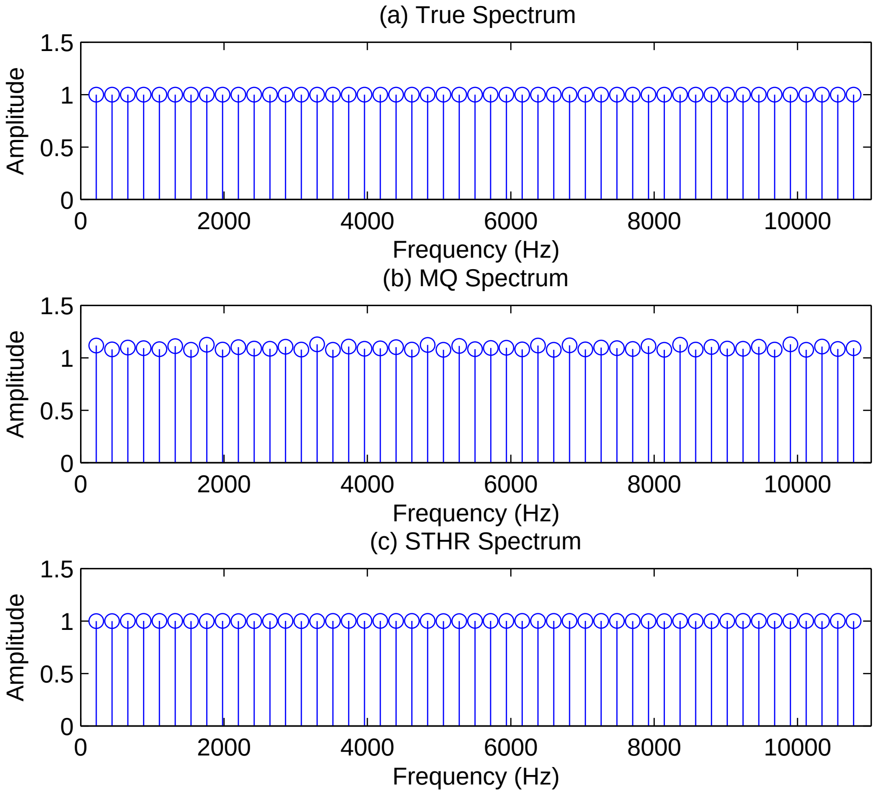

Figure 3.6: Line spectrum of a bandlimited pulse train ($f_0 = 220$ Hz)(a) and its estimate for a single frame of (b) MQ and (c) STHR sinusoidal analysis.

### 3.2.2 Regularization parameter optimization

Prior to performing a large scale evaluation of how well the STHR sinusoidal analysis system can represent real-world signals, recall that the choice of a regularization parameter remains an open problem. Because optimization of the regularization parameter for each least-squares subproblem is intractable, the value of the regularization parameter is chosen such that it maximizes the global SER on a small dataset. A selection of 100 real-world musical tones and 100 two-tone mixtures is used to evaluate a set of 20 possible regularization parameters between $\alpha = 0$ (no regularization) and $\alpha = 1$ (equal weighting between least-squares residual and solution norm). Although this is a relatively small dataset, the aim of this experiment is to provide insights on the interaction between regularization parameter and representation accuracy.

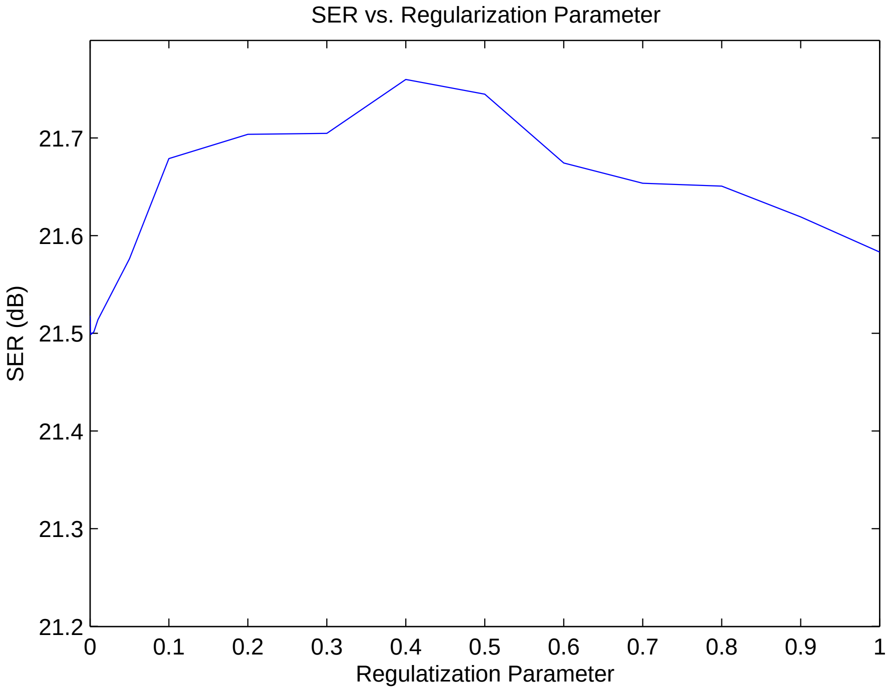

Figure 3.7: SER performance as a function of regularization parameter.

Figure 3.7 shows SER performance as a function of regularization parameter. Peak performance is attained at a value of $\alpha = 0.4$. This represents a relatively strong weighting of the solution norm minimization. It is observed, however, that there is not a great deal of variance of SER due to the regularization parameter. First, the harmonic "explosion" problem is somewhat of a rare occurrence. In terms of global SER, such misrepresentations contribute large errors only at single frames. Moreover, the workings of the tracker serve to largely ignore such gross errors. In the case where these "explosions" do occur over longer time spans, and are thusly not handled by the tracker, regularization can contribute to some performance gain. The good representation performance with large regularization parameters also provides hints that in general, the solution that minimizes the least-squares residual is also one with small norm. With no large performance degradation due to a large regularization parameter, a conservative system can use a relatively large parameter such as $\alpha = 0.4$ to aid in handling the occasional numerical instability in the estimation of harmonic amplitudes. A regularization parameter of $\alpha = 0.4$ is chosen for the STHR sinusoidal analysis system and used throughout the remaining experiments in this work.

### 3.2.3 Real-world musical tones and tone mixtures

A dataset of 2000 real world musical tones and two tone mixtures is used to evaluate the STHR sinusoidal analysis system's ability to accurately represent pitched musical sources. The tones are derived from the RWC Musical Instrument Database [108]. Examples are drawn from brass (trumpet, trombone, tuba, and French horn), woodwind (oboe, flute, bassoon, clarinet, alto sax, tenor sax, baritone sax), and string (violin, viola, and cello) families for a total of 15 unique instrument types. The dataset for each instrument comprises recordings of three unique instruments with varying articulation methods, pitches, and dynamics. Half the dataset used in this evaluation consists of single instrument tones, and half of two-tone mixtures. Each tone or tone mixture is analyzed using both STFT-based MQ analysis and STHR sinusoidal analysis. Sounds are resynthesized from the extracted sinusoidal tracks and evaluated against the originals in terms of SER performance.

Table 3.1 shows performance results of MQ and STHR analysis against the aforementioned dataset of musical tones and tone mixtures. It is evident that the STHR sinusoidal analysis method generates a valid sinusoidal model for real-world sounds, largely equivalent to traditional sinusoidal analysis. Average performances are similar between the two competing techniques, although STHR does show a larger standard deviation in its performance (it occasionally produces representations that can be either far superior or significantly worse that STFT-based methods). Example harmonic tracks of a trombone tone for both an MQ and STHR sinusoidal analysis are shown in Figure 3.8, reiterating that STHR analysis is capable of producing a valid sinusoidal model.

It is important to emphasize that the simple analysis and resynthesis of musical tones gives little additional insight into the functionality of STHR

Table 3.1: SER performance (average, standard deviation, minimum, and maximum) of MQ and STHR sinusoidal analysis on real-world sounds.

| **SER** | **MQ** | **STHR** |
|---|---|---|
| **Avg.** | 20.68 | 23.97 |
| **Std.** | 0.768 | 4.25 |
| **Min** | 18.26 | 14.04 |
| **Max** | 22.58 | 33.58 |

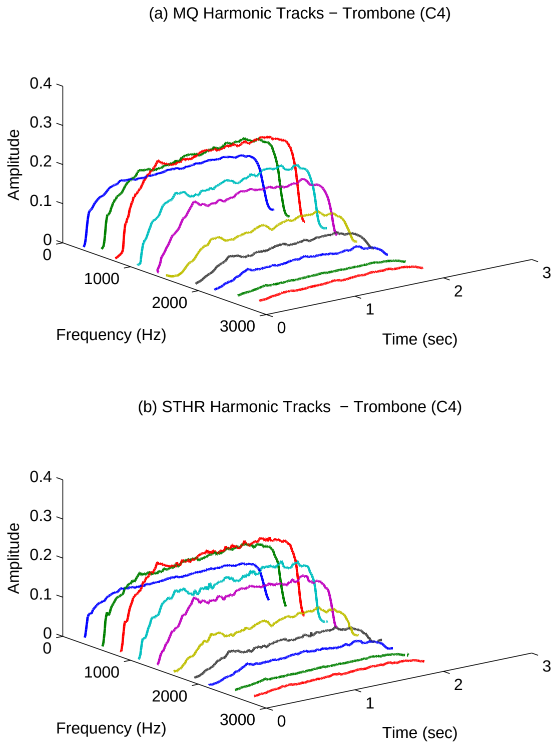

Figure 3.8: Harmonic tracks for a monophonic C4-trombone tone derived from (a) MQ and (b) STHR sinusoidal analyses.

analysis other than that it is capable of producing valid sinusoidal models. An ultimate goal of the analysis system is to resolve and track closely spaced harmonics. A resynthesis of a sound mixture does little to capture whether two closely spaced individual tracks are actually resolved (or are merged into a single track as is often the case with STFT-based methods). It is also unclear whether harmonics are tracked properly in the case that closely spaced sinusoids are resolved at the frame level by matrix pencil estimation. Because the tracker is tied to continuity in amplitude, there exists the possibility that, if two closely spaced harmonics are also close in amplitude, tracks can incorrectly switch from tracking the partial of one source to another. In an additive resynthesis of such a mixture, these behaviors will not be clearly evident in the overall spectrogram from which the SER measure is derived. Therefore, to truly evaluate tracking performance, experiments can be carried out in the context of source separation. In such an application, errors due to mistracking or track fusion will become more apparent. Such experiments are reserved for Chapter 5. It is generally true, however, that the resynthesis of mixed signals produces poorer performance than the analysis and synthesis of monophonic tones.

### 3.2.4 Harmonic collision detection

As an intermediate evaluation of the potential for STHR sinusoidal analysis to resolve closely spaced partials, a harmonic collision detection experiment is performed at the frame level. In this experiment, 1500 two-tone mixtures are produced, drawn from the same set of instruments as those presented in Section 3.2.3. The mixtures are produced from one-second segments of the individual source tones. The segments are extracted from the one second of audio immediately following the peak amplitude value of the tone (i.e., the segments are meant to represent the steady-state portion following the attack).

Prior to mixing, an STFT is performed on each source segment. Peak picking is performed on each frame of the resulting spectrograms to estimate the locations of harmonics for each source at each frame. In addition, an equivalent analysis is performed on the mixture. The produced line spectrum of the mixture is compared to the individual source spectra. Cases where a single peak is observed in the mixture and where prominent peaks exist at the same locations in both individual sources are marked as harmonic collisions. Observed harmonic peaks in the mixture that can only be attributed to a single source are marked as uncollided harmonics. This markup of each frame serves as ground-truth for the harmonic collision detection experiment. The mixtures are only produced for three musical intervals, namely perfect octaves, fourths, and fifths (500 of each interval). These intervals produce relatively even balances among uncollided and collided harmonics.

STHR sinusoidal analysis is performed on the one-second two-tone mixtures. For every frame, the ground truth is compared to the active harmonic tracks of the STHR analysis. There exist two distinct outcomes for each of the harmonic types (collided and uncollided). For uncollided harmonics, if only a single prominent track exists at that location in the STHR analysis, a true negative (TN) is counted. If the STHR analysis reports multiple tracks at the location of an uncollided harmonic, a false positive (FP) is counted.

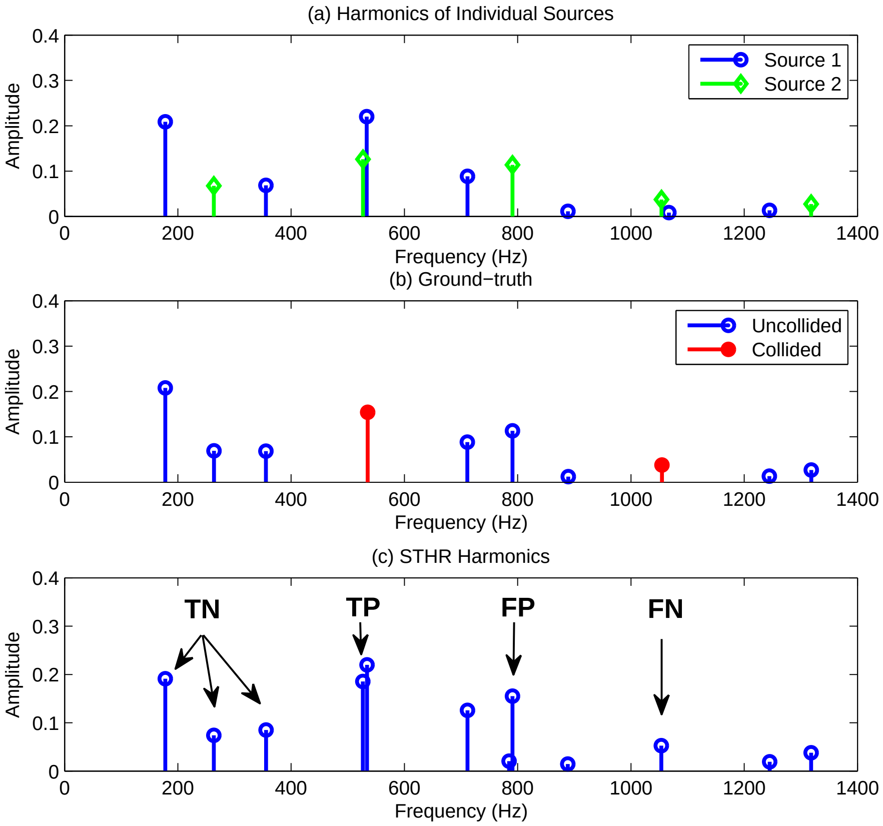

Figure 3.9: Example harmonic detection collision experiment for a perfect fifth mixture. Individual source harmonics are shown in (a). The marked up ground-truth derived from the mixture spectrum is shown in (b). The STHR analysis with different outcome types for each type of harmonic is shown in (c).

For collided harmonics, if two prominent tracks in the STHR analysis exist, a true positive (TP) is counted. If only a single track exists at a collision, a false negative (FN) is counted. Figure 3.9 shows an example of a mixture of a tuba and French horn at a perfect fifth interval for a single frame. The bottom plot shows the STHR analysis with the four possible outcomes (TP, FN, FP, TN) highlighted.

Table 3.2 shows the true positive, false negative, false positive, and true negative rates over the 1500 example dataset. STHR sinusoidal analysis correctly detects collided harmonics 75.5% of the time, and misreports uncollided harmonics as collided 16.3% of the time. Overall, these detection rates show promise as collisions are correctly identified for roughly 3/4 of all

Table 3.2: True postive rate (TPR), false negative rate (FNR), false positive rate (FPR), and true negative rate (TNR) for a harmonic collision detection experiment.

| | **Collision Present** | **Collision not Present** |
|---|---|---|
| **Collision Detected** | TPR = 75.5% | FPR = 24.5% |
| **Collision not Detected** | FNR = 16.3% | TNR = 83.7% |

occurrences.

Some reasons for the failure cases (false negatives and false positives) can be explained by a deeper understanding of the analysis system and the nature of real-world tones. While the direct matrix pencil method can resolve arbitrarily closely spaced harmonics when no noise is present and when the signal fits the model perfectly, these two criteria are not met with real-world sounds. There is always either some form of noise or noise residual accompanying a real source's harmonics, and there are deviations from the model within each frame.

In the presence of noise, resolution of arbitrarily closely spaced harmonics is not possible. If two harmonics are indeed very closely spaced in a noisy signal, it is common for only a single pole to be extracted. In general, these false negatives occur for lower harmonics. Consider a unison in which the fundamental frequencies differ by 0.1 Hz. As harmonic number increases, there is a larger separation between overlapping harmonics. By the tenth harmonic, there is a 1 Hz separation, and by the twentieth harmonic there is a 2 Hz separation. An observation of the STHR analyses of the mixtures shows that false negatives do indeed occur most often for the first one or two collisions.

Many of the false positives (reporting a collision when none exists) are attributed to a deviation from the model that has, heretofore, been largely unaddressed. Because the dataset contains tones with varying articulation styles, instruments such as violin and oboe produce sounds containing vibrato (periodic frequency modulation). Therefore, within a single frame, the frequencies of sinusoidal partials are not stationary. As shown by Badeau, a stationary sinusoidal model does indeed support tones with frequency modulation [24]. However, the means by which non-stationary frequencies are supported produce a somewhat unfortunate consequence in terms of collision detection. Non-stationarity of frequency can be considered a form of frequency modulation. It has long been known that the frequency modulation of audio tones produces line spectra where additional partials appear above and below the carrier frequency at multiples of the modulation frequency [109]. The strength of these generated partials is governed by Bessel functions that depend on the strength of the modulation (modulation index). Therefore, a partial undergoing frequency modulation may produce a group of poles. Within a single short time-frame, frequency modulation more resembles a frequency chirp. Nevertheless, multiple poles are often extracted. The propensity to extract multiple poles occurs most often for higher partials, as the effective modulation index is increased. A peak modulation of 5 Hz at the fundamental will correspond to a 50 Hz peak modulation at the tenth harmonic. Therefore for vibrato tones, multiple poles are often extracted for each harmonic, with higher harmonics having a greater tendency to produces multiple pole estimates.

## 3.3 Summary and Discussion

This chapter presented a short-time high-resolution sinusoidal analysis system. The analysis entails filtering a signal into sub-bands through the use of a complex filter bank. Each sub-band is downsampled, and segmented into overlapping frames. Direct matrix pencil sinusoidal analysis is performed on each frame of each sub-band to extract local estimates of sinusoidal partials. Regularization is performed on the least-squares estimation of sinusoidal amplitudes in each frame to reduce effects of numerical instability. A regularization parameter was determined empirically using a small dataset. A sinusoidal tracker that ensures continuity in both frequency and amplitude is used to link sinusoidal estimates between frames to generate sinusoidal tracks.

The STHR analysis system was evaluated against a range of input sounds to test its ability to provide an accurate sinusoidal representation of musical sounds. In the analysis and resynthesis of synthetic sounds, it was made evident that the STHR system does not provide a perfect reconstruction of tones that perfectly fit the underlying model used in matrix pencil estimation. Nonetheless, the resulting representations were still very accurate. For real-world musical tones and mixtures, the analysis system showed it is capable of producing valid sinusoidal representations largely equivalent to existing Fourier transform based techniques.

A final experiment tested the STHR analysis system's ability to resolve and detect harmonic collisions. The harmonic collision detection experiment highlighted some of the shortcomings of the STHR sinusoidal analysis system. Those deficiencies that are largely attributed to background noise and deviation from a fixed-frequency model are not easily addressed and persist as the major weaknesses of the system. Nevertheless, the analysis method still maintains a moderately good performance at detecting and resolving overlapping partials. The focus for the remaining chapters in this work becomes: Can the STHR sinusoidal analysis system provide benefits for a range of musical signal processing applications?

# CHAPTER 4

# MULTIPLE FUNDAMENTAL FREQUENCY ESTIMATION

This chapter presents a method for estimating the fundamental frequencies of multiple sources in music mixtures based on high-resolution sinusoidal analysis. An overview of multiple fundamental frequency estimation (multi-$f_0$) techniques was provided in Chapter 2. The overview highlighted that harmonic collisions play an important role in multi-$f_0$ estimation. Existing methods for handling harmonic collisions in this context usually rely on allowing multiple fundamental frequency hypotheses to share the energy of an observed harmonic if the multiple hypotheses indicate that the harmonic is collided. In iterative-cancellation estimation techniques, harmonic contributions are not fully canceled when a single $f_0$ estimate is made. This allows for the possibility that an observed partial can also be attributed to another source whose $f_0$ will be estimated in subsequent steps.

The ultimate goal of this chapter is to test whether high-resolution sinusoidal analysis can provide any benefit in frame-level estimation of the fundamental frequencies of mixed musical sources. Specifically, if collided harmonics can be correctly resolved by STHR analysis, there is strong additional evidence to support that a specific harmonic location should be attributed to more than one source. The method proposed for estimating multiple fundamental frequencies in this chapter is rather naïve. The primary justification for choosing a simplistic estimation method is to test the merits of the STHR sinusoidal analysis system in its own right. If a rather complex system is used to produce estimates from the STHR analysis, it is difficult to attribute which aspect of the overall system (harmonic retrieval vs. fundamental frequency estimation) is responsible for any performance gains. In other words, does the system perform well because of the underlying high-resolution sinusoidal representation, or does the system perform well due to robustness of the $f_0$-estimation method? By restricting the fundamental frequency estimation portion of the overall system to work only on basic principles, the potential of STHR sinusoidal analysis is better evaluated.

This chapter is organized as follows. Section 4.1 introduces a cancel and iterate approach to multi-$f_0$ estimation. The presented method operates on estimates of harmonic frequencies and amplitude present in a signal in short time-frames. It is therefore applicable to harmonics retrieved using either STHR sinusoidal analysis or by peak picking of the spectrogram. Section 4.2 presents two experiments. In the first experiment, only two-tone mixtures are considered. The mixtures are produced for 13 intervals from perfect unison to perfect octave (0 to 12 semitones). The second experiment considers higher orders of polyphony and tests performance on two-, three-, and four-source mixtures. The proposed multi-$f_0$ method is tested using both STHR and STFT derived harmonics. In addition, the performances are compared to a current state-of-the-art system.

## 4.1 An Iterative-Cancellation Method for Multiple Fundamental Frequency Estimation

The short-time high-resolution sinusoidal analysis system presented in Chapter 3 generates estimates of the parameters of sinusoidal partials present in a signal. It was shown in Section 3.2.4 that in approximately 75% of cases where two harmonics effectively overlap, the STHR analysis system correctly identified that multiple sinusoids were present. A multiple fundamental frequency estimation system can potentially leverage this information to better estimate the fundamental frequencies of multiple sources present in a signal. The iterative-cancellation techniques for multi-$f_0$ analysis presented in Section 2.4.2 are appealing in terms of both their efficiency and conceptual simplicity. Assume that an analysis system has the ability to perfectly resolve harmonic collisions, and an iterative-cancellation technique is used to estimate what fundamental frequencies are present at a given time based on this hypothetical analysis system. In such a case, *full* cancellation of the harmonics that can be attributed to a predominant $f_0$ estimate is a valid approach. Because only a single harmonic at a given location can be attributed to a single source, in the case of multiple closely spaced harmonics, only one will be canceled. The residual will therefore maintain the other harmonics at that general collided location, to be used in later iterations. This behavior is

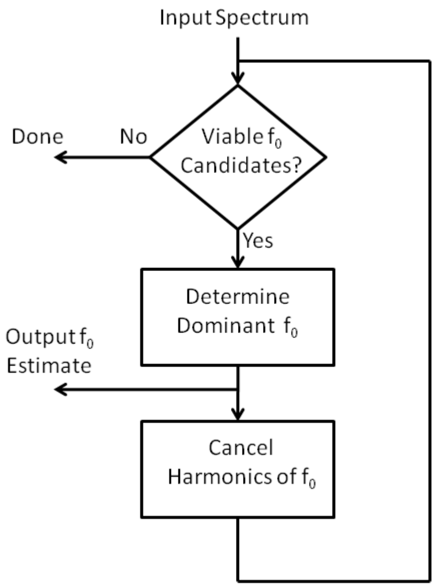

Figure 4.1: Block diagram of a cancel and iterate multiple fundamental frequency estimation procedure.

the principle by which the multi-$f_0$ system presented in this section operates.

Figure 4.1 shows a block diagram of an iterative-cancellation multi-$f_0$ estimation technique. The input spectrum is surveyed to see whether harmonics of significant amplitude are present in the signal to produce a viable fundamental frequency estimate. This serves as the stopping criterion for the iterative algorithm. If there are enough significant partials present in the spectrum, the predominant $f_0$ is estimated and harmonics of that $f_0$ are canceled from the spectrum. The procedure is then repeated on the residual spectrum. After enough iterations, there will not be enough evidence left in the residual to support a $f_0$ hypothesis, and the stopping criterion is met.

### 4.1.1 A fundamental frequency salience function

To evaluate a set of $f_0$ hypotheses, some manner of scoring or weighting them must be devised. The term *salience* is often used as a generic description of such scoring methods to encompass the broad range of approaches that can achieve such ends, ranging from probabilities (in probabilistic frameworks) to simple scoring functions, measures, and metrics. An $f_0$ estimate is made when a hypothesis produces a (relatively) high salience. Therefore, the first step in the design of a multiple fundamental frequency estimator is the choice of an appropriate salience function.

The chief characteristic of $f_0$ salience functions is that they measure the periodic, and thus harmonic, nature of pitched sources. As mentioned in Section 2.4, regardless of the details of how periodicity or harmonicity is measured, the measurement can often be conceptualized as some manner of harmonic sieve or comb that operates in the frequency domain. In other words, harmonics that reside at the expected harmonic locations of a $f_0$ hypothesis are either selected or accentuated through the calculation of the salience function. Therefore, a simple salience function can be calculated by directly implementing such a comb or sieve in the frequency domain.

The proposed salience function operates directly on a frequency domain representation in the following manner. First note that the type of frequency domain representation used is that of a line spectrum (i.e. a group of harmonic frequencies and amplitudes) as derived from harmonic retrieval techniques such as the STHR sinusoidal analysis or peak picking of the spectrogram (as done in MQ analysis). In the case of STHR analysis, it is critical that the sinusoidal tracking procedure is used prior to multi-$f_0$ estimation, as opposed to the pole estimates directly generated by the matrix pencil method in each sub-band. Recall that the model order in each sub-band of STHR analysis is usually overestimated, resulting in spurious signal poles. Sinusoidal tracking serves to separate and prune spurious poles from those that correspond to actual partials. Denote each of $M$ currently active sinusoids in a given analysis frame $S_m$, $m = 1$, ..., $M$, and the functions $f$ and $A$ the frequency and amplitude of the argument, respectively. All observed harmonics form a set denoted $H_{all}$ with $M$ members. For each expected harmonic location of a $f_0$ hypothesis, the observed sinusoid that is closest to it in frequency, and within a quarter tone (3% of frequency), is selected. The selected harmonics form a subset of $H_{all}$, containing only partials that are selected due to a $f_0$ hypothesis. Denote this subset $H_{f_0}$ and each member $S_k$ (the partial selected to serve as the $k^{th}$ harmonic of the $f_0$ hypothesis). Therefore, for all harmonic locations $kf_0$ (the $k^{th}$ harmonic of the hypothesis), the set of selected partials $H_{f_0}$ with elements $S_k$ is defined as

$$\begin{aligned} S_k &= \underset{S_m}{\operatorname{argmin}} |f(S_m) - kf_0| \quad \text{for k = 1, ...,K} \\ H_{f_0} &= \{S_k : |f(S_k) - kf_0| < 0.03kf_0\} \end{aligned} \tag{4.1}$$

The total number of harmonics, $K$, is chosen such that $Kf_0$ is less than the half sample rate. The subset $H_{f_0} \subseteq H_{all}$ has $K$ members.

The rudimentary selection procedure of Equation 4.1 has a relatively large tolerance window of a quarter tone. To ensure that the selected sinusoidal partial is indeed a viable harmonic, its amplitude is weighted by a (unnormalized) Gaussian function centered on the hypothetical harmonic location. This weighting of partial amplitude enforces a strong harmonicity on potential $f_0$ estimates. If the sinusoidal partial is relatively close to the hypothetical harmonic location, its amplitude is largely unaffected. However, if the selected partial does not align well with the hypothetical harmonic location, it is greatly penalized by the weighting function. The weighted, selected harmonics are then summed to produce a salience score for the $f_0$ hypothesis. The salience function of a fundamental frequency hypothesis, denoted $SAL(f_0)$, is calculated as

$$SAL(f_0) = \sum_{S_k \in H_{f_0}} \sqrt{A(S_k)} e^{-\frac{(f(S_k) - kf_0)^2}{2\sigma_k^2}} \tag{4.2}$$

where the harmonic, $S_k \in H_{f_0}$, selected as the best match to a hypothetical harmonic location, $kf_0$, is defined as before in Equation 4.1. The square root of the harmonic amplitudes prior to summation serves to compress the spectrum. The choice to compress the spectrum stems from the fact that many instruments produce tones whose harmonics decay with harmonic number. Compressing the spectrum accentuates higher harmonics so that they play a role in the salience score. Summation of linear-scale and log-scale amplitudes were also tested, but found to perform worse. The direct summing of harmonic amplitudes follows closely from a technique proposed by Klapuri [110].

The choice of a Gaussian weighting function to enforce strong harmonicity is somewhat arbitrary. Any function that monotonically decreases from its center and effectively vanishes at distances deemed too far from its center

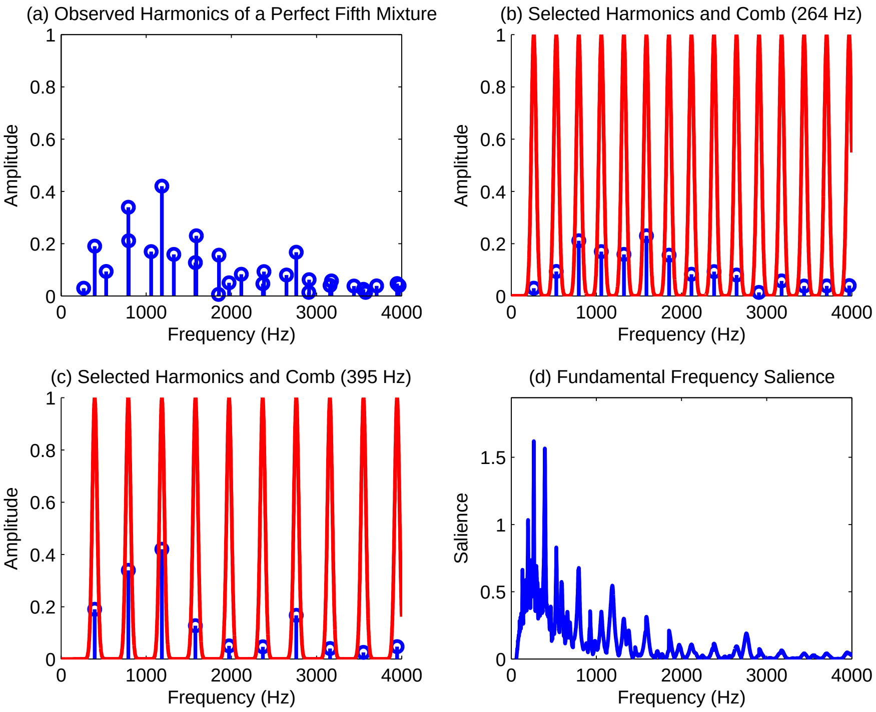

Figure 4.2: Example salience calculation of a perfect fifth mixture. The mixture spectrum is shown in (a). The selected harmonics and weighting comb of the two true fundamental frequencies are shown in (b) and (c). The salience function is shown in (d).

is suitable. The choice of the variance parameter, $\sigma_k^2$, controls the effective width of each "tooth" of the harmonic comb. Because frequency deviations from a fundamental become more extreme with increasing harmonic number (due to the multiplicative effect of harmonic number and fundamental frequency), $\sigma_k^2$ could be made to increase with increasing harmonic number, $k$. In addition, inharmonicities in source spectra could be better accounted for if $\sigma_k^2$ is made a function of $k$ in such a manner. However, for the sake of simplicity, the width of each component Gaussian is made to be fixed-width. The standard deviation is set to $\sigma_k = 15$ Hz. Therefore, this salience scoring method is best suited to spectra that do not show a large degree of inharmonicity.

An example of salience calculation is presented in Figure 4.2. The source mixture contains two tones a perfect fifth apart at fundamental frequencies of 264 Hz and 394 Hz. The harmonic selection procedure and weighting combs are displayed for each of these two fundamentals. The resultant salience function, calculated on a 1 Hz search grid, is also displayed. In this particular example, the top two saliences match the true fundamental frequencies.

### 4.1.2 Cancel and iterate procedure

The salience function described in the previous section details a method of scoring and estimating the dominant fundamental frequency present in an observed line spectrum. In order to estimate the fundamental frequencies of other sources, the contributions of the dominant source's harmonics must be eliminated. The need for elimination stems from the fact that a dominant $f_0$ will also produce high saliences for its octaves and sub-octaves that can dominate the saliences of other sources. The proposed system relies on full cancellation of harmonics that are attributed to the highest salience $f_0$ estimated in the current step. Full cancellation is an aggressive strategy. However, the primary intent for this system is to operate off of harmonic estimates derived from short-time high-resolution sinusoidal analysis. If closely spaced harmonics are resolved, full cancellation of one harmonic will leave its close-by neighbors intact.

The first consideration in a fundamental frequency estimation procedure is the granularity at which $f_0$ hypotheses are tested. For instance, $f_0$ hypotheses could be tested on a 1 Hz resolution grid. However, doing so requires the evaluation of thousands of salience functions to evaluate the entire pitch range musical instruments are capable of. To greatly reduce the search space in evaluating $f_0$ hypotheses, candidates are drawn directly from the observed spectrum. Each observed harmonic serves as a potential candidate and $f_0$ hypothesis. Naturally, such a process does not account for the full scope of human pitch perception where a missing fundamental in a harmonic spectrum will still yield a percept of a tone at that fundamental. However, acoustic instruments rarely produce spectra with altogether missing fundamentals (greatly suppressed fundamentals in comparison with other harmonics are common, however). The cancel and iterate approach is carried out as follows.

**Step 1.** Find the predominant fundamental frequency in the current input spectrum. Denote the set of the $M$ currently active, observed harmonics $H_{all}$. Once again, denote the frequency of the sinusoid $S_m \in H_{all}$ as $f(S_m)$ and its amplitude $A(S_m)$. For $m = 1, ..., M$ compute

$SAL(f(S_m))$ as in Equation 4.2. The dominant $f_0$ is chosen as the $S_m$ that maximizes $SAL(f(S_m))$. Denoting this predominant $f_0$ as $f_0^{max}$, it is calculated as $f_0^{max} = \underset{f(S_m)}{\operatorname{argmax}} \left[SAL(f(S_m))\right]$

**Step 2.** Calculate the stopping criterion. Calculate the root mean square (RMS) of the harmonic amplitudes selected and assigned to $f_0^{max}$. These harmonics are the subset $H_{f_0^{max}} \subseteq H_{all}$ defined in Equation 4.1. Calculate the RMS amplitude of predominant estimate $A_{RMS}^{f_0^{max}}$ as $A_{RMS}^{f_0^{max}} = \sqrt{\frac{1}{K}\sum_{S_k \in H_{f_0^{max}}} A(S_k)^2}$. If this is the first iteration of the algorithm, set $A_{RMS}^1 = A_{RMS}^{f_0^{max}}$. If $A_{RMS}^{f_0^{max}} > 0.01A_{RMS}^1$, return $f_0^{max}$ as an estimated $f_0$ and continue to Step 3. If the extracted $f_0^{max}$ does not meet this criterion (40 dB less in RMS amplitude than the first, predominant $f_0$ estimate), stop.

**Step 3.** Cancel the selected harmonics attributed to $f_0^{max}$. The set of all $M$ harmonics is $H_{all}$. The subset of $K$ harmonics attributed to $f_0^{max}$ is $H_{f_0^{max}}$. Remove this subset from the set of all harmonics as $H_{all} \leftarrow H_{all} - H_{f_0^{max}}$. There are $M \leftarrow M - K$ residual harmonics. Go to Step 1.

Figure 4.3 provides an example of the cancel and iterate approach on the same perfect fifth mixture used to demonstrate salience scoring. Each source's set of harmonics, $H_{f_0}$, is extracted and canceled to produce a residual spectrum at each iteration of the algorithm. After two iterations of the algorithm, any extracted $f_0$ candidate does not meet the stopping criterion, and the procedure is stopped.

An important behavior to note about the harmonic selection procedure (used for salience scoring and cancellation) is that it selects harmonics based solely on frequency proximity. In cases where two partials are closely spaced (i.e., overlapped), it is possible for the selection procedure to select the wrong partial (i.e., the harmonic of another source). Therefore, the true harmonic of another source may be eliminated. However, because only a single partial is eliminated, there still exists an observed partial at that harmonic location. In other words, there exists the distinct possibility that the harmonic extraction procedure can cause sources to "swap" harmonics.

As stated earlier, the presented multi-$f_0$ estimation system operates under the assumption that all partials in a musical mixture are perfectly resolved.

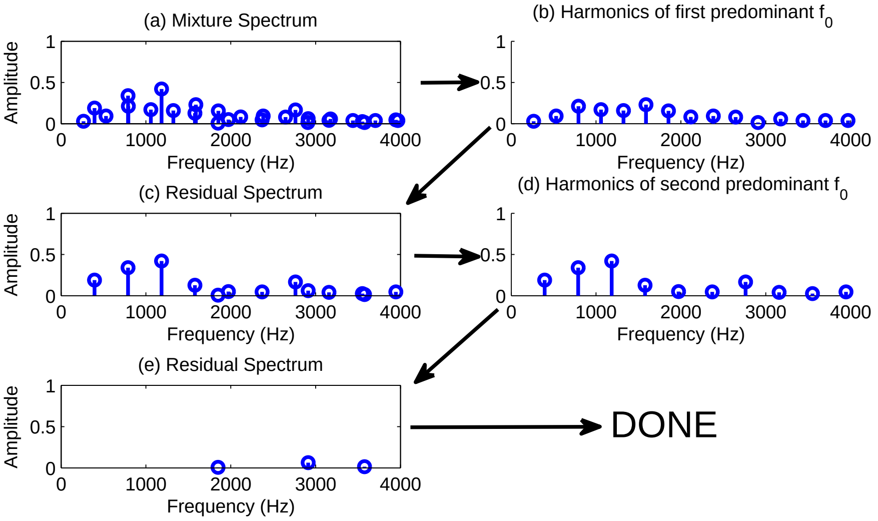

Figure 4.3: Cancel and iterate approach in a perfect fifth mixture (262 and 392 Hz). The mixture spectrum is shown in (a). The extracted spectrum of the first predominant $f_0$ is shown in (b). The residual after the spectrum of (b) is canceled is shown in (c). The spectrum of the second predominant $f_0$ is shown in (d). The residual after the spectrum of (d) is canceled is shown in (e).

The full cancellation of partials attributed to a predominant $f_0$ estimate assumes that if a collision exists at a harmonic location, there will be close-by harmonics left in the residual. If collisions are not resolved, the cancellation will remove evidence to support subsequent $f_0$ hypotheses. Furthermore, because the $f_0$ hypotheses are drawn from the set of currently available partials, there is the distinct possibility that a potential hypothesis is removed altogether in a previous cancellation. The removal of a hypothesis due to unresolved collisions represents a worst possible failure case. While the proposed technique can easily be adapted to be more robust to such failures, it provides a good opportunity to evaluate the ability of short-time high-resolution sinusoidal analysis to resolve overlapping partials.

## 4.2 Experiments and Evaluation

The proposed multiple fundamental frequency estimation system is evaluated against a set of musical tone mixtures with varying degrees of polyphony. The first set of experiments involves testing the system's ability to correctly estimate the fundamental frequencies present in two-tone mixtures. These two-tone mixtures are constructed by mixing instrument tones at predetermined musical intervals. To compare the representations produced by short-time high-resolution sinusoidal analysis to the ones that would be similarly generated from the STFT, the fundamental frequency estimation system is tested with these two different harmonic retrieval front-ends. Moreover, the multi-$f_0$ estimation system of Yeh [40] is used as a state-of-the-art benchmark for comparison. The system of Yeh is the top performing system in the 2008 and 2009 iterations of the Music Information Retrieval Evaluation eXchange (MIREX). In these evaluations, the 2009 version of Yeh's frame-level estimation algorithm is tested. The three methods (proposed system with STHR front-end, proposed system with STFT front-end, and Yeh) are also evaluated against a dataset of arbitrary two-, three-, and four-tone mixtures. For the remainder of this chapter and presentation of the results, these multi-$f_0$ systems will be referred to as STHR, STFT, and Yeh. Section 4.2.1 presents the results of the known-interval two-tone evaluation. The test of broader ranges of polyphony is found in Section 4.2.2.

### 4.2.1 Evaluation of two-tone mixtures at known intervals

A dataset of one-second mixtures is used to evaluate the multi-$f_0$ estimation systems. As in Section 3.2.3, the individual tones are drawn from 15 instrument types derived from the RWC music instrument database. One-second segments are extracted from the one second of audio immediately following the peak amplitude value in each tone. Mixtures are produced by mixing one-second segments such that the two fundamental frequencies of the individual tones correspond to 13 base intervals (zero to twelve semitone separations). The fundamental frequencies of each individual tone are verified (to ensure that the supplied RWC metadata is correct) by using a monophonic pitch detection algorithm [111]. Each interval has 1000 examples for a total of 13000 mixtures. Note that no compensation for differing amplitudes is performed on the individual tone segments prior to mixing. This simulates real world cases where mixtures are not guaranteed to be mixed at 0 dB ratios. Pitches are estimated for every frame of the one-second mixture. All algorithms are

Table 4.1: Multiple fundamental frequency estimation accuracies of algorithms for two-tone mixtures at known musical intervals.

| **Interval** | **STFT** | **STHR** | **Yeh** |
|---|---|---|---|
| **P1** | 0.491 | 0.629 | 0.618 |
| **m2** | 0.492 | 0.905 | 0.847 |
| **M2** | 0.617 | 0.937 | 0.813 |
| **m3** | 0.841 | 0.944 | 0.891 |
| **M3** | 0.885 | 0.943 | 0.923 |
| **P4** | 0.887 | 0.942 | 0.883 |
| **d5** | 0.882 | 0.938 | 0.921 |
| **P5** | 0.883 | 0.941 | 0.845 |
| **m6** | 0.883 | 0.948 | 0.905 |
| **M6** | 0.891 | 0.945 | 0.913 |
| **m7** | 0.891 | 0.941 | 0.910 |
| **M7** | 0.811 | 0.942 | 0.903 |
| **P8** | 0.549 | 0.800 | 0.856 |
| **Avg.** | 0.770 | 0.904 | 0.864 |

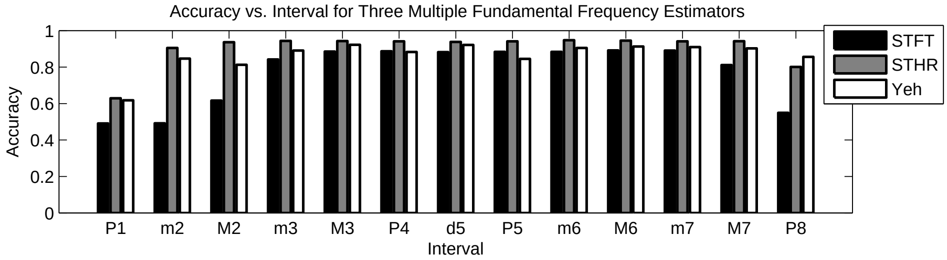

Figure 4.4: Multiple fundamental frequency estimation accuracies of algorithms for two-tone mixtures at known musical intervals.

set to use a 93 ms frame size with 87.5% overlap. The 93 ms frame size is chosen to equate with settings commonly used by Yeh. All algorithms are set to detect a maximum polyphony of two. However, as all algorithms have stopping criteria, they can report a single $f_0$.

Table 4.1 and Figure 4.4 show the frame-level fundamental frequency estimation accuracies for the three algorithms and each of the 13 tested intervals. The ground-truth fundamental frequencies are quantized to the nearest semitone in the equal-tempered scale. Likewise, algorithm outputs are also quantized to the nearest note. Accuracy measures the proportion of fundamental frequency estimates that perfectly match the ground-truth. It is true that there can potentially be some pitch deviation within a note. However, because the segments are drawn from steady-state portions of the source signals, and the average fundamental frequencies of each source segment are verified using a monophonic pitch detector, the number of mis-annotated frames is assumed to be small.

The proposed multi-$f_0$ estimation algorithm using the short-time high-resolution sinusoidal analysis front-end produces the best average performance (0.904 accuracy). The algorithm of Yeh has an average accuracy of 0.864. It is important to note that this is a significantly higher error rate than reported by Yeh in [40]. The degraded performance can be attributed to a slightly different experimental setup. The underlying difference is that the error rates reported by Yeh correspond to frame-level estimations where each source is mixed at a 0 dB ratio to the other for every single frame. In this experimental procedure, it is not uncommon for one source to be significantly stronger than the other as within a one-second span. Some of the source tones decay in amplitude at a faster rate than others. These measured accuracy rates are not inconsistent with Yeh's performance in MIREX 2009 for musical passages with two instruments. Furthermore, it is possible that the algorithm submitted to MIREX 2009 by Yeh (and used in this evaluation) is tuned to work on the types of musical mixtures used in the MIREX evaluations. Finally, the proposed system with a STFT-based front-end achieves an average accuracy of 0.770. This poor performance is expected, as the aggressive cancellation strategy of the multi-$f_0$ system does not interact well when overlapping partials exist.

Significance testing is performed on the three systems to measure whether or not the system performances differ in a statistically significant way. The average performance of each system is measured for each audio mixture. Each mixture is treated as a separate sample, for a total of 13000 samples. The three different algorithms form three separate groups. To test statistical significance, a one-way analysis of variance (ANOVA) test is run against the 13000 samples. The one-way ANOVA test indicates that there is a statistically significant difference in the performance of the algorithms ($p < 0.001$). A subsequent Tukey-Kramer Honestly Significant Difference test (TK-HSD) [112] ($\alpha = 0.05$) shows that all algorithms perform significantly different from one another. The TK-HSD comparison plot for the one-way ANOVA test is shown in Figure 4.5. The comparison intervals shown in the figure do not

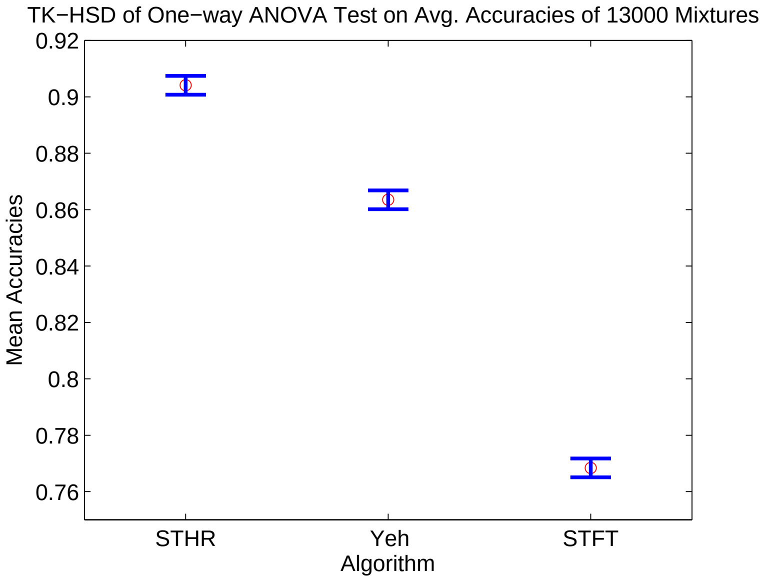

Figure 4.5: TK-HSD comparison using one-way ANOVA of algorithms over 13000 two-tone mixtures.

overlap (disjoint), indicating that the algorithms are significantly different.

An additional statistical test is also performed directly on Table 4.1. For this test, Friedman's ANOVA test is chosen [113]. Friedman's test is a nonparametric test meaning that the samples do not need to be drawn from some known distribution. With a sample size of only 13, a nonparametric test is preferred because the normality requirement of standard ANOVA is likely not met. The test operates based on performance rankings for each sample. In this case, there are 13 samples, one for each interval. Once again, Friedman's test indicates that there exists a statistically significant difference among the algorithms ($p < 0.001$). The subsequent TK-HSD test ($\alpha = 0.05$), however, indicates that only the STHR-based method is significantly different from the STFT-based method and Yeh's method. The TK-HSD result for the Friedman's test is shown in Figure 4.6.

Examination of the average performances of the algorithms for each musical interval provides some interesting insights. As expected, perfect unisons (P1) and octaves (P8) present the most difficult mixtures. The STHR-based algorithm and Yeh's algorithm perform similarly for unison mixtures. Yeh's algorithm produces the best estimates for octave mixtures highlighting the robustness of the algorithm to difficult mixtures. The failures of the STHR-based system in these cases are largely attributed to false negatives of re-

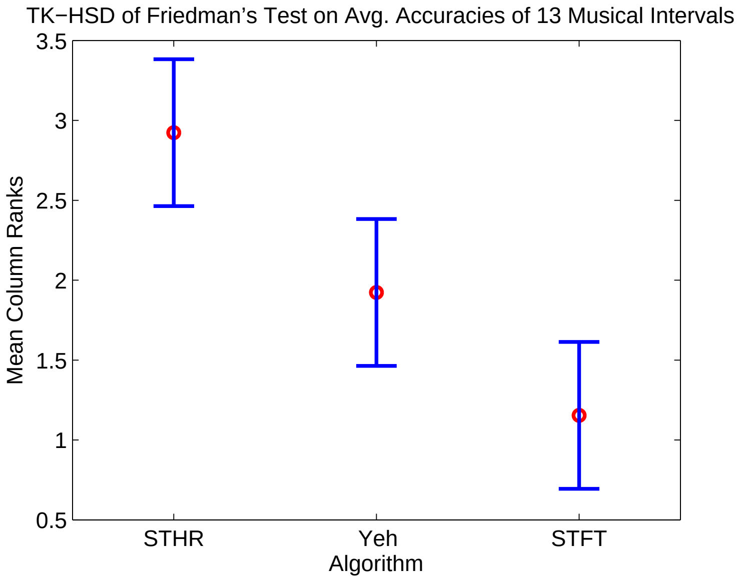

Figure 4.6: TK-HSD comparison using Friedman's test of algorithms over average performance on 13 musical intervals.

ported harmonic collisions in the first few harmonics. In other words, harmonic collisions are not always resolved. When the collisions are not resolved, potential $f_0$ candidates are removed due to the full-cancellation strategy. For unison and octave mixtures, the STFT-based method produces accuracies on the order of 0.5. This is due to the fact that the cancellation of the predominant $f_0$ leaves no substantial residual for subsequent steps. Therefore, a comparison of the STFT and STHR methods indicates that there are performance gains attributed to STHR analysis. The minor second (m2) interval (one semitone) also presents an interesting case in terms of the performance of the STFT-based approach. For low pitches, a one semitone difference is a relatively small separation in frequency. Therefore, low harmonics effectively overlap in this case producing a high error rate in the STFT-based estimations. However, for STHR analysis, these separations are large enough that false negatives rarely occur in resolving these overlaps. Yeh's algorithm shows slight degradations in performance for perfect fourth (P4) and perfect fifth (P5) intervals. These two intervals, aside from unisons and octaves, have the highest proportion of overlapping partials.

Table 4.2: Multiple fundamental frequency estimation accuracies of algorithms for two-, three-, and four-tone mixtures.

| **Polyphony** | **STFT** | **STHR** | **Yeh** |
|---|---|---|---|
| **2** | 0.848 | 0.869 | 0.868 |
| **3** | 0.761 | 0.830 | 0.871 |
| **4** | 0.662 | 0.780 | 0.840 |

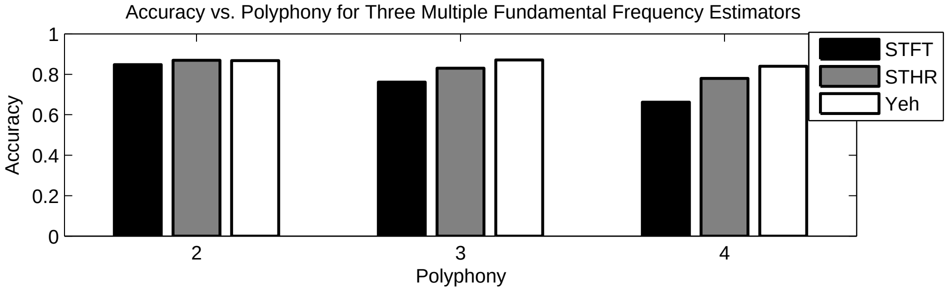

Figure 4.7: Multiple fundamental frequency estimation accuracies of algorithms for two-, three-, and four-tone mixtures.

### 4.2.2 Evaluation of two-, three-, and four-tone mixtures

The multiple fundamental frequency estimation systems are further evaluated to test their performances against varying degrees of polyphony. Mixtures containing two, three, and four sources are constructed. Unlike the mixtures produced in Section 4.2.1, the mixtures for this evaluation are produced by randomly drawing tones from the dataset of 15 musical instruments. Therefore, the mixtures are not constrained to contain any predetermined musical intervals. The two-tone mixtures in this case are not restricted to be in the same general pitch register allowing, for example, a mixture of a low tuba tone with a high flute tone. A total of 3000 mixtures are produced with 1000 examples for each level of polyphony. The source segments are extracted from the one second following the peak amplitude point of each component tone, as before. The maximum polyphony present (two, three, or four) is supplied to each of the systems. The same parameter settings (e.g. 93 ms frame size) as Section 4.2.1 are used.

Table 4.2 and Figure 4.7 show the frame-level fundamental frequency estimation accuracies for the three algorithms and each of the three tested degrees of polyphony. For random two-tone mixtures, all algorithms show

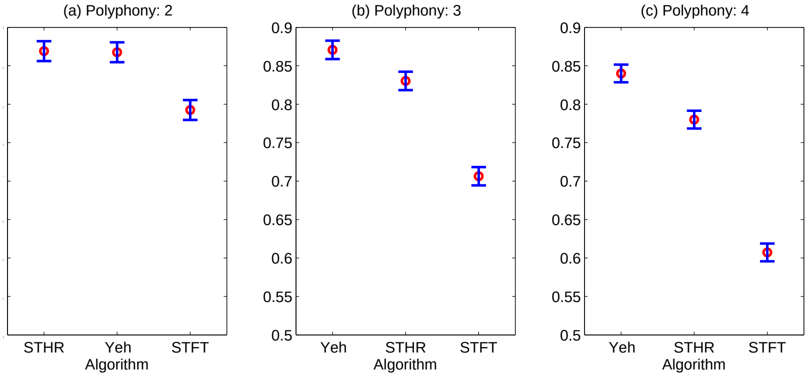

Figure 4.8: TK-HSD comparison using one-way ANOVA of algorithms over 1000 two-tone (a), three-tone (b), and four-tone (c) mixtures. Each individual plot is sorted by average performance.

similar performances. When the mixtures are produced randomly, and the component tones are possibly at vastly different pitch registers, the STFT-based system does not show as poor a performance as before for two-tone mixtures (its previous poor average performance can largely be attributed to unison and octave cases). With increasing polyphony, the system of Yeh degrades in performance more gracefully than the other systems. The STFT-based system is most strongly affected with increased polyphony. This degradation demonstrates that the proposed multi-$f_0$ estimation method is too aggressive with its source cancellation policy, especially for harmonics retrieved from STFT analysis. The STHR-based method does not show as strong a performance decline as compared to the STFT-based method with increasing polyphony. The better performance the STHR front-end suggests that it produces a better representation than STFT-based harmonic retrieval in this application.

Significance testing is performed to measure whether statistically significant differences exist between the algorithms for each degree of polyphony. A one-way ANOVA test is performed using the average performance on each of the 1000 mixtures (for a given degree of polyphony) as a sample. Figure 4.8 shows the subsequent TK-HSD comparisons for each of the three polyphonies. For the random two-tone mixtures, the STHR-based system and Yeh's system do not differ significantly (overlapping comparison intervals). For three and four-tone mixtures, all systems perform significantly differently, with Yeh's system consistently performing best, followed by the STHR-based method and finally the STFT-based method.

## 4.3 Summary and Discussion

This chapter presented a multiple fundamental frequency estimation system. The system produces estimates of predominant fundamental frequencies at the analysis-frame level. The estimates are based on the observed harmonic frequencies and amplitudes produced by a harmonic retrieval or sinusoidal analysis system within a given analysis frame. Therefore, the system is suitable for use with any analysis system that produces this form of line spectrum. The system is designed to function under the assumption that the analysis front-end is able to produce a perfect line spectrum, with all closely spaced partials perfectly resolved. Under this assumption, a full-cancellation iterative procedure is adopted. The fundamental frequency hypotheses are drawn directly from the observed spectrum, with each observed partial serving as a potential $f_0$ hypothesis. When a predominant fundamental frequency estimate is made, all harmonics that can be attributed to it are completely removed from the spectrum. The procedure is repeated on the residual spectrum until any $f_0$ estimate that is made does not produce a harmonic spectrum that has sufficient energy. In other words, the procedure is stopped when the residual contains no partials of significance.

The underlying assumption that an analysis front-end produces a perfectly resolved line spectrum is intentional. This naïvety is included to test the ability of the short-time high-resolution sinusoidal analysis system of Chapter 3 to resolve harmonic collisions. The evaluation of this system can be considered an extension of the experiment conducted in Section 3.2.4 where a collision detection is performed. In this case, the analysis system's capability to resolve harmonic collisions is tested in the context of a useful music signal processing application.

Two experiments were conducted and presented in this chapter to evaluate the STHR-based multiple fundamental frequency estimator. To provide baselines, the proposed system was also tested with a STFT-based front-end that produces the same type of line spectra that STHR sinusoidal analysis produces. In addition, the evaluation of a state-of-the-art system was included for comparison. In the first experiment, systems were evaluated against two-tone mixtures created at known musical intervals. The STHR analysis-based system produced the best overall performance, far outperforming the equivalent system using a STFT-based front-end. In the second experiment, higher degrees of polyphony were tested. For polyphonies greater than two, the state-of-the-art system outperformed those presented here. This is an expected result, as the fundamental frequency estimator presented here operates under flawed assumptions. As the level of difficulty increases, the weaknesses of this multi-$f_0$ estimator become more apparent. However, all things being equal in terms of $f_0$ estimation, the STHR-based front-end shows a less severe performance degradation than the same system with a STFT-based front-end. This fact provides a strong indication that there is merit and potential to STHR sinusoidal analysis and its ability to produce better and more accurate signal representations. Also, the fact that the STHR-based method performs the best of all algorithms in the case of difficult two-tone mixtures (drawn from the same octave) implies that high-resolution analysis is a powerful technique.

# CHAPTER 5

# SOURCE SEPARATION USING COMPUTATIONAL AUDITORY SCENE ANALYSIS

This chapter presents approaches to audio source separation based on short-time high-resolution sinusoidal analyses. In previous chapters, STHR sinusoidal analysis was demonstrated to have the ability to extract the parameters of closely spaced sinusoids. This behavior is most strongly evident in the harmonic collision detection experiment of Section 3.2.4 and the multiple fundamental frequency estimation experiments of Chapter 4. However, the astute observer would notice that in these previous experiments, what is truly being evaluated is the ability of the STHR analysis to determine the existence of multiple sinusoids at a given harmonic location, and not the accuracy of the estimates of the sinusoidal parameters. In the case of the multiple fundamental frequency estimation system, for example, it is largely unimportant if the parameter estimates of a resolved partial are a few hertz off in frequency or a few decibels off in amplitude. The harmonic detection collision experiment is also indifferent to the quality of the parameter estimates. While the analysis/synthesis of single and mixed tones found in Section 3.2.3 showed that STHR analysis produces good reproductions of mixed signals, it is once again unclear what proportion of the overlapped partials are truly resolved. While the accuracy of parameter estimates can be deduced by direct comparison of the representation of a mixed signal to those of the source signals, such an evaluation can also be performed in the context of a potentially useful application, namely, source separation. Source separation encompasses a broad range of techniques. The technique adopted here is computational auditory scene analysis (CASA).

The basic principles of CASA were introduced in Section 2.5.2. With a sinusoidal representation (as produced by sinusoidal tracking methods), CASA aims to group sinusoidal partial tracks into the constituent sources of a musical sound mixture. The grouping is achieved through a variety of cues including, but not limited to, harmonicity and common fate of partial onsets, amplitudes, and frequencies. The harmonics constituting a group can then be additively synthesized to produce a source estimate. This is the basic principle of the CASA system presented here. Assuming that harmonics are grouped correctly, the accuracy of extracted sinusoidal parameters can be evaluated by comparing an estimated source to its corresponding unmixed original. Note that the correctness of the partial grouping is also largely dependent on the accuracy of the parameter estimates. This is due to the fact that for closely spaced partials, amplitude and frequency trajectories of the sinusoidal tracks serve as the main cues for grouping. Closely spaced partials naturally have a high harmonic concordance to both sources that share that harmonic location. Therefore, errors in both frequency and amplitude can cause trajectories that do not match well with others, leading to grouping errors. Because the CASA based separations are dependent on the accuracies for parameters of for grouping, and the signal-to-error ratio of the resulting separation to the original is also directly dependent on parameter accuracy, source separation serves to evaluate the accuracy of STHR sinusoidal analysis.

This chapter is organized as follows. Section 5.1 introduces the previously outlined CASA system. In addition, an alternative interpretation of CASA based on the sinusoidal tracks is also presented. The alternative system operates under the premise that the end-goal for CASA is the extraction of the ideal binary mask (IBM) for a source. This alternative system therefore uses the partial groupings to build STFT separation masks for a source. Some baseline masks are also covered to serve as a basis for comparison. Section 5.2 presents experimental results for these systems on audio mixtures.

## 5.1 CASA-Based Separations from Sinusoidal Tracks and Baseline Systems

This section introduces two CASA source separation systems derived from short-time high-resolution sinusoidal analyses. A method that groups sinusoidal tracks and produces source estimates is presented in Section 5.1.1. The synthesis of groups of harmonic tracks using additive synthesis is discussed in Section 5.1.2. An alternative system that builds binary time-frequency masks derived from these groupings is presented in Section 5.1.2. Finally, Section 5.1.3 introduces baseline systems for comparison including ideal binary masks, harmonic masks, and their variants.

### 5.1.1 Grouping of sinusoidal tracks

The grouping of sinusoidal tracks first involves measuring their similarity to one another. Tracks that have high similarity are grouped together. Similarity scores among sinusoidal tracks can be generated in a variety of ways. A common approach is to combine the scores produced by measuring similarities across a variety of facets (e.g., harmonicity, common fate, etc.) to produce a summary similarity score. Such an approach requires that proper weights for each of the individual measures be determined to produce the best performing summary similarity measure. Because such weighting schemes can only be determined empirically and are not guaranteed to generalize well, a multi-step approach is adopted here instead. Tracks are first grouped based on a single cue, where groups are allowed to share tracks that are mutually similar to them. Tie-breaking procedures are then performed on shared tracks to uniquely assign them.

For pitched musical instrument tones, perhaps the most powerful of grouping cues is harmonicity or harmonic concordance. A measure of harmonic concordance is found in Equation 2.33. In the case of musical mixtures, most musical intervals produce spectra where some harmonic locations are not interfered with by other sources and some are. For partials that are not overlapped, the harmonic concordance measure is effective for uniquely assigning these tracks to groups. This unique assignment stems from the fact that these unshared harmonics have a harmonic relationship *only* with the partials belonging to their parent source. Partials that are overlapped will have a high harmonic concordance with the harmonics of multiple sources. These overlapping partials are initially assigned to all groups that have a high harmonic concordance with them. The uniquely assigned (not overlapped) harmonics of each group can then be used as a reference to assign the closely spaced partials to individual groups. A tie-breaking procedure is performed by measuring common frequency and amplitude modulation of an overlapped partial to the (strongest) uniquely assigned partials of each group.

The step-by-step harmonic grouping procedure proceeds as follows. Denote the set of all observed time-varying sinusoids (tracks) $H_{all}$, with a total of $M$ individual sinusoids each denoted $S_m^t$ for $m = 1, ..., M$, where $t$ is the time (frame) index. The frequency of track $S_m$ at time $t$ is $f(S_m^t)$ and its amplitude is $A(S_m^t)$. The goal is to form disjoint subsets of $H_{all}$ whose members constitute the sinusoidal tracks for an estimated source. For the sake of notational simplicity, assume that only two sources are present. Denote the target groups (subsets) $G_1$ and $G_2$ with $G_1 \subseteq H_{all}$, $G_2 \subseteq H_{all}$, and $G_1 \cap G_2 = \oslash$.

**Step 1.** Form the groups $G_1$ and $G_2$ from the set of all observed harmonics $H_{all}$ based on harmonic concordance. Form an $M \times M$ distance matrix, $D_h$, of each track $S_m \in H_{all}$ to all others using the harmonic concordance distance of Equation 2.33. Denote this harmonic distance between two frequencies $f_i$ and $f_j$ as $d_h(f_i, f_j)$. The frequency of a given sinusoidal track, $S_m$ is expressed as $f(S_m)$. Therefore the harmonic distance between two sinusoidal tracks, $S_i$ and $S_j$ is $d_h(f(S_i), f(S_j))$ Because the frequencies of each track are time varying, harmonic concordance is measured on a frame-by-frame basis and integrated over the time region that tracks simultaneously exist to produce the harmonic distance between time-varying sinusoidal tracks. The frequency of a track $S_m$ at time-step $t$ is expressed as $S_m^t$. The first and last frames that two partials overlap are $t_1$ and $t_2$, respectively. Thus, each element of the distance matrix $D_h(i, j)$ is

$$D_h(i,j) = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} d_h\left(f(S_i^t), f(S_j^t)\right) \quad \text{for i,j = 1, ...,M} \quad (5.1)$$

The assignment of tracks to groups involves thresholding the above distance matrix. However, reference tracks from which to form the target groupings must be established. To do so, tracks that have no significant close neighbors in frequency are identified. Such tracks are deemed to be uncollided harmonics. The uncollided harmonics that have significant duration (due to proper and stable tracking) serve as reference candidates. The candidates form a set of tracks, $C$. The mean amplitudes of the reference candidates are then measured. Candidates with relatively high amplitude are selected to serve as references because it is expected that the extracted sinusoidal parameters are most accurate for significant partials. The candidate with highest mean amplitude is considered first. Assume that this highest mean amplitude track has index $k$. An initial grouping $G_1$ is formed by thresholding the row of the distance matrix corresponding to track $S_k$. That is, all track indices $j$ are extracted such that $D(k, j) < 0.03$ (quarter-tone tolerance). This operation forms the group $G_1$. The elements of $G_1$ that are also members of the candidate set $C$ are removed from $C$. That is, $C \leftarrow C - C \cap G_1$. The highest mean amplitude track of the residual candidates $C$ is then used to form the subsequent group $G_2$. Such a procedure can be iterated while groups with sufficient cardinality (number of elements) are formed to produce groups in mixtures of more than two sources.

**Step 2.** For sinusoidal tracks that are initially assigned to more than one group, disambiguate and uniquely assign them to individual groups. The initial groupings $G_1$ and $G_2$ are bound to share partials if the sound mixtures contain sources played at consonant musical intervals. These overlapping partials, $G_1 \cap G_2$, must be uniquely assigned to a group. The disambiguation is performed by measuring the common frequency modulation of Equation 2.30 of collided harmonics to uncollided ones. While common amplitude modulation can be a powerful cue, it is not used in this system for the following reasons: First, the estimation of amplitudes during STHR sinusoidal analysis is a numerically sensitive procedure. Therefore, the most egregious errors for parameter estimates are expected to be in the partial amplitudes. Second, because harmonics are tied to a fundamental frequency, their frequency trajectories are more likely to match well. Some instrument families produce harmonics whose amplitude envelopes closely match. However, common fate of harmonic amplitude envelopes is not universally true. For example, the harmonics of brass instruments often have similarly shaped amplitude envelopes, while instruments such as flute often have harmonic amplitude envelopes that behave more chaotically, even during steady-state portions of the tone. Thus, denote the set of shared (overlapping) partials, $O$, where $O = G_1 \cap G_2$. Non-overlapped partials for group $G_1$ can be expressed as $G_1 - O$, and for group $G_2$ as $G_2 - O$. The frequency modulation similarity between two tracks $S_i$ and $S_j$ is expressed as $s_f\left(f(S_i), f(S_j)\right)$ as in Equation 2.30. Therefore, the group assigned to some shared partial, $S_i \in O$, is chosen to be the one that maximizes a total frequency modulation similarity score. If there are $L$ potential groups, $G_l$, the track $S_i$ is assigned to the group that satisfies

$$\underset{G_l}{\operatorname{argmax}} \sum_{S_j \in G_l - O} \frac{\bar{A}(S_j)}{\sum_{S_j \in G_l - O} \bar{A}(S_j)} s_f\left(f(S_i), f(S_j)\right) \quad \text{for l = 1, ...,L} \tag{5.2}$$

where $\bar{A}(S_j)$ denotes the mean amplitude of track $S_j$. Weighting each individual similarity score, $s_f\left(f(S_i^t), f(S_j^t)\right)$, by the mean amplitude of the track ensures that more weight is placed on stronger partials.

The preceding procedure serves to uniquely assign tracks into groups of harmonics that are estimated to be attributed to individual sources. This information can be used to subsequently synthesize the source estimates.

### 5.1.2 STHR-CASA with an sinusoidal additive synthesis engine

The sinusoidal tracks of a group of harmonics can be synthesized by performing sinusoidal additive synthesis. For mixed signals, the previous section outlined a procedure to form harmonic groups that pertain to estimates of individual sources. Therefore, sinusoidal additive synthesis is used on each extracted group to synthesize a separated source. A sinusoidal oscillator is used for each harmonic track of a group. The time-varying amplitudes and frequencies of each harmonic track are used to drive each oscillator. The outputs of the oscillators are then summed to produce the final synthesized signal.

### 5.1.3 STHR-CASA with a binary mask synthesis engine

An alternative interpretation of the goal of auditory scene analysis is the extraction of an ideal binary mask. With this end-goal in mind, a binary time-frequency mask for a source estimate can be constructed based on the sinusoidal tracks that constitute the source estimate. Binary masks operate on time-frequency representations of the signal. For the sake of simplicity, and without loss of generality, assume that the time-frequency representation used to operate on a mixed signal is a short-time Fourier transform. For a group of harmonic tracks that constitute a source estimate, the time-varying frequencies of the partials are easily translated to DFT bins. Therefore, the time frequency points that correspond to the active tracks of an estimated source in a given frame are assigned a value of "one" in a binary mask. Because a single DFT bin is rather narrow, the direct neighbors of the bin that most closely matches the track frequency at a given time are also assigned a value of "one." In the case of overlapped partials, a "one" is assigned to the mask corresponding to the group with the stronger partial at the shared location. This effect mimics ideal binary masks which assign a time-frequency point to a source that has local dominance. These masks (one for each set of grouped harmonics) are then applied to the STFT of the mixed signal. The masked time-frequency representation can then be synthesized using overlap-add STFT synthesis to produce a synthesis of a source.

### 5.1.4 Baseline masks

As stated previously, if the end-goal of CASA is the extraction of ideal binary masks, then a source synthesis derived from an IBM serves as a performance goal. However, the extraction of an IBM requires perfect knowledge of the source signals. A mask that requires less (but still some) prior knowledge of the source signals is a harmonic mask (HM). A harmonic mask is generated from the known fundamental frequency trajectories of a source. Time-frequency points that correspond to harmonic locations of the known fundamental are allowed to pass in a harmonic mask. Note that for harmonic masks, time-frequency points at harmonic collisions are shared among the individual source masks. The harmonic masks used here have a mask width of five bins for each harmonic. That is, the bin that most closely matches the expected harmonic location, and its two neighbors on either side, are given a value of "one." As a final baseline system, an ideal harmonic binary mask is also used for a basis of comparison. Because the sinusoidal model that STHR sinusoidal analysis operates under ignores the noise components of signals, the IHBM is a restriction of the IBM to harmonic locations. Recall that the IBM allocates all time-frequency points based on local dominance, including time-frequency regions that do not correspond to harmonic locations. It therefore also effectively separates the noise components of each source signal. Thus the IHBM can be viewed as a performance goal of the underlying sinusoidal model used to produce separations. The IHBM can be constructed by simply taking the Hadamard (element-wise) product of the ideal binary mask and harmonic mask of a source. Denoting an ideal binary mask, $M_{IBM}$, and a harmonic mask, $M_{HM}$, the ideal harmonic binary mask is therefore $M_{IHBM} = M_{IBM} \circ M_{HM}$.

## 5.2 Experiments and Evaluation

A dataset of two-tone mixtures is used to evaluate the CASA-based source separation systems. As in previous experiments, the individual tones are drawn from 15 instrument types derived from the RWC music instrument database. The mixtures are restricted to perfect fifths. Perfect fifths provide a good balance between collided and uncollided harmonics. While the separation of octaves and unisons would be a desirable goal, such mixtures are too difficult in this context. All partials of such mixtures are harmonically concordant. Furthermore, although the mixtures of two tones may seem like trivial examples by which to evaluate the systems, these mixtures actually represent a very challenging case from a CASA perspective. Most of the articulations of the different instrument types do not have a significant amount of modulation. Furthermore, the simple mixing of individual tones represents a case where two tones have a common onset (and usually a common offset as well). Therefore, for octave or especially unison mixtures there is not sufficient information to produce accurate separations. As a concession, the next most difficult case, the perfect fifth, is used for evaluation. In total, 2000 perfect fifth mixtures are produced. The basic parameters of the STHR sinusoidal system are 46 ms frame size and 87.5% frame overlap. Both variants of STHR-based CASA are evaluated, namely the system that uses additive synthesis as its separation engine and the system that uses binary masks as its separation system. These systems are referred to as STHR-AS and STHR-BM, respectively. Moreover, the separations produced by the ideal binary mask (IBM), the ideal harmonic binary mask (IHBM), and the harmonic mask (HM) are also evaluated for comparison purposes.

Table 5.1: Average spectral to signal error ratios (SER) for five CASA-based source separation systems and masks.

| | **IBM** | **IHBM** | **HM** | **STHR-AS** | **STHR-BM** |
|---|---|---|---|---|---|
| **SER (dB)** | 17.34 | 17.34 | 12.44 | 13.01 | 14.63 |

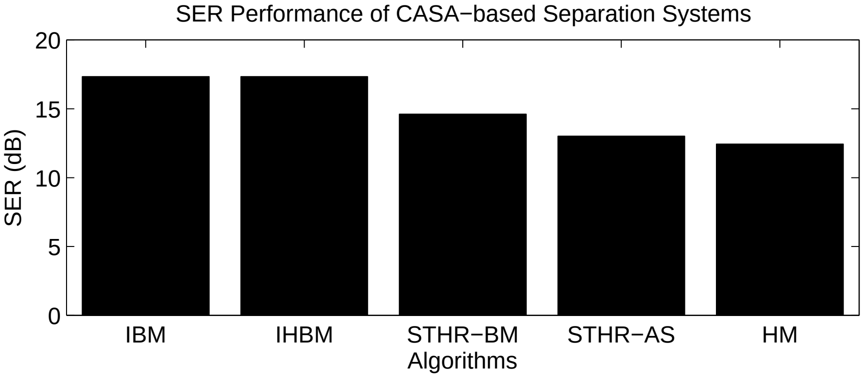

Figure 5.1: Average spectral to signal error ratios (SER) for five CASA-based source separation systems and masks.

The performances are measured using the spectral signal-to-error ratio (SER) measure of Equation 2.36

Table 5.1 and Figure 5.1 show the performances of each of the systems on the set of 2000 mixtures. Restricting an ideal binary mask to pass only harmonics has no net effect on performance. This is not unsurprising because the individual tones are recorded in a relatively noise-free environment and because many instruments do not produce a significant noise component. The purely harmonic mask (which allows all harmonic collisions to pass to all sources) is the worst performer. Both STHR-based systems fall between these baselines in terms of performance.

As with the multiple fundamental frequency estimation experiments, a significance test is performed to measure whether the systems differ in a statistically significant way. A one-way ANOVA test indicates that the systems do indeed differ ($p < 0.001$). The follow-up Tukey-Kramer Honestly Significantly Different test is performed. The TK-HSD plot is shown in Figure 5.2. The *post hoc* test indicates that IBM and IHBM separation masks do not differ significantly. All other systems have significant differences from one to another. However, STHR-AS has a minimal performance gain over a simple harmonic mask. While statistically significant, this performance gain is not

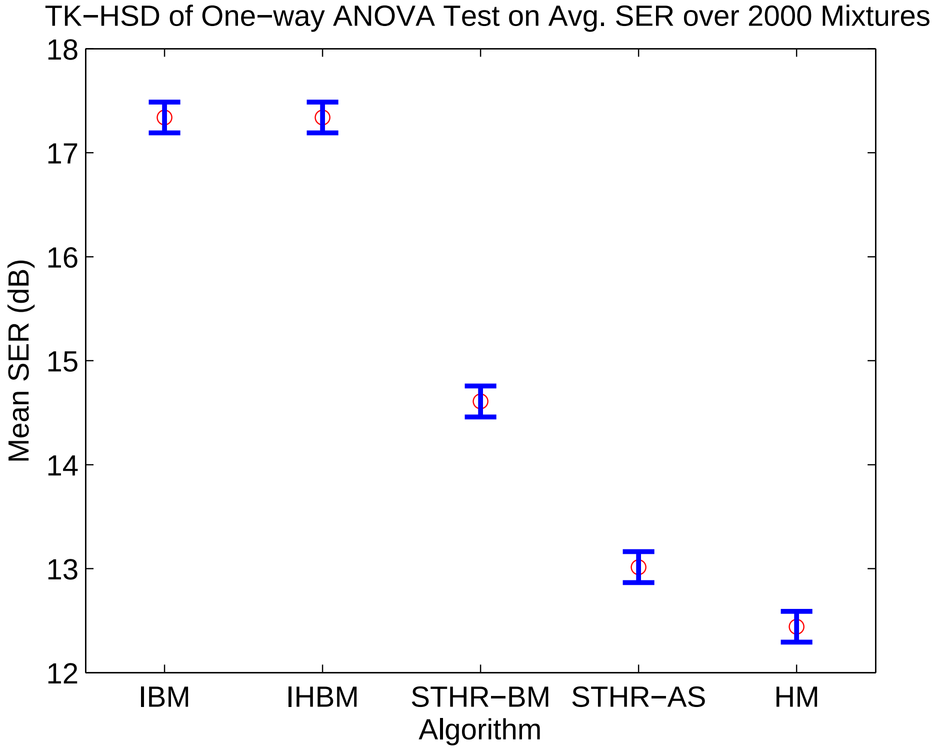

Figure 5.2: TK-HSD comparison using one-way ANOVA of separation systems.

of practical significance. The binary masks derived via the STHR sinusoidal analyses do seem to produce some benefit to separation performance.

The failures (or rather, lack of benefit) of the additive synthesis based separation system are a discouraging but not wholly unexpected result. Put simply, the shortcomings of STHR sinusoidal analysis can be attributed to three main factors. First, the amplitude and frequency estimates derived from ESPRIT's parameter estimation are not perfect for closely spaced partials. In the resulting separations, errors in the estimated harmonic amplitudes for tracks contribute to the total overall error. Perhaps the most significant of the shortcomings is that the tracking of closely spaced partials is prone to failure. This largely stems from poor parameter estimates in some frames of an analysis. These poor estimates can misguide sinusoidal tracks and eventually cause them to break. The end result is that instead of a well behaved track with long duration, a series of shorter-lived tracks are often produced in its place. With the additive resynthesis, this track-breaking behavior can cause severe artifacts. While there exist methods to

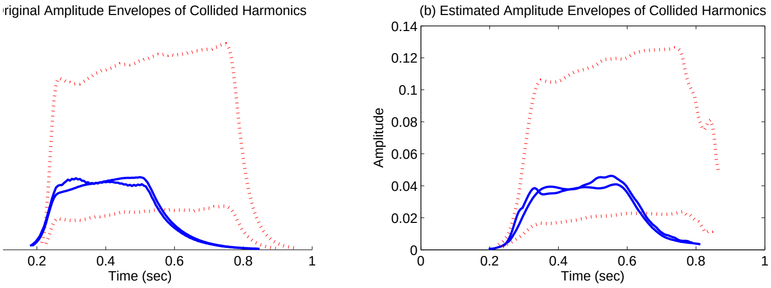

Figure 5.3: An example of good performance in resolving collided harmonics. The amplitude envelopes of the first two collided harmonics of a mixture are shown for two sources. One source's harmonics are drawn with solid lines and the other's with dotted lines. The original harmonic amplitude envelopes are shown in (a). The resolved and grouped harmonic amplitude envelopes are shown in (b).

address such behaviors in the synthesis step, such concessions were not made in order to evaluate the performance of STHR-separation as is. Finally, as with all bottom-up systems, the rather rudimentary grouping procedures used here are prone to error propagation. Poor parameter estimates can lead to wrongly assigned tracks and incorrect groupings. This is especially true of broken tracks because they have short durations. Overlapped tracks are compared to reference tracks only over the span of time they both occupy. Short tracks produce less evidence upon which to base grouping decisions, and can be susceptible to grouping errors.

Figure 5.3 provides an example of where the resolution and tracking of overlapping partials produces reasonably accurate estimates. In this perfect fifth mixture, the amplitude envelopes of the first two collided harmonics of each source are shown. One source's harmonic amplitude envelopes are drawn with solid lines, and the other source's harmonic amplitudes with dotted lines. The left plot shows the amplitude envelopes extracted from the monophonic source tones. The right subplot shows the grouped source estimates from the mixture. In this case, the STHR sinusoidal analysis accurately resolved the harmonics.

Figure 5.4 provides an example of where the STHR sinusoidal analysis system produces estimates that are not quite as accurate as the previous example case. Occasional (inaccurate) dips and peaks in the amplitude envelopes are evident. This case also shows an example of a broken track in the decay portion of one of the harmonics. The attack and steady state portion

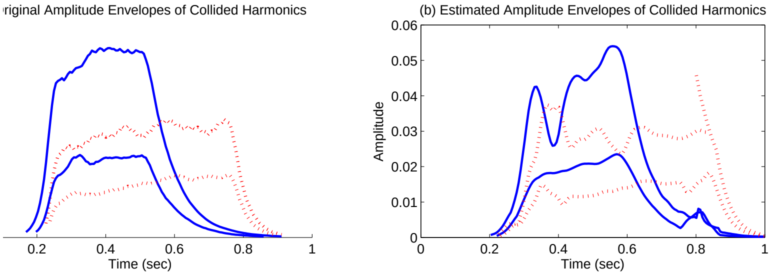

Figure 5.4: An example of satisfactory performance in resolving collided harmonics. The amplitude envelopes of the first two collided harmonics of a mixture are shown for two sources. One source's harmonics are drawn with solid lines and the other's with dotted lines. The original harmonic amplitude envelopes are shown in (a). The resolved and grouped harmonic amplitude envelopes are shown in (b).

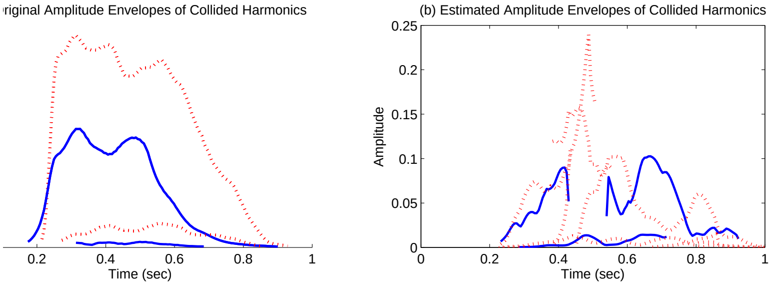

Figure 5.5: An example of poor performance in resolving collided harmonics. The amplitude envelopes of the first two collided harmonics of a mixture are shown for two sources. One source's harmonics are drawn with solid lines and the other's with dotted lines. The original harmonic amplitude envelopes are shown in (a). The resolved and grouped harmonic amplitude envelopes are shown in (b).

of this harmonic are captured in one track. However, near the beginning of the decay portion of the harmonic, a separate track forms that encompasses the decay. In this case, the system was fortunate enough to correctly group this short track.

Figure 5.5 provides an example of a near catastrophic failure in terms of parameter estimate accuracy and the successful tracking of partials. A series of shorter-lived tracks with wildly behaving amplitude envelopes are formed. Cases such as these drive average performance down, and negate any net benefit drawn from the previous two cases. The sinusoidal additive synthesis of the sinusoidal tracks shown in this figure produces more artifacts than benefits.

Despite the potential for occasional failure, if short-time high-resolution sinusoidal analysis is viewed as a means to identify harmonic collisions, and *approximate* their parameters, performance gains can be made. Although the actual amplitude estimates of partials of closely-spaced partials may be prone to some error, it is often the case that the dominant partial will still produce a dominant amplitude estimate. In this sense, STHR-based sinusoidal analysis can be used to derive a binary mask that more closely resembles the ideal binary mask. It is this effect that allows a binary mask derived from STHR analysis to produce a performance gain.

## 5.3 Summary and Discussion

This chapter presented methods for CASA-based source separations derived from short-time high-resolution sinusoidal analyses. The partial tracks generated by the analysis system undergo a grouping procedure to form representations of source estimates. The groupings are formed based on shared harmonic relationships and the frequency trajectories of the tracks. Two methods of generating separated source signals from these groups were presented. One method relies on simple additive synthesis of the partials belonging to a group. The other method aims to derive binary masks for each source from the groupings based on local dominance of its tracks over the tracks of other groups.

To evaluate the performances of the systems, 2000 perfect fifth mixtures were produced. Separated source estimates were then synthesized and compared to the original component tones. In addition, a series of baseline systems including ideal binary masks and harmonic masks were evaluated. The additive synthesis-driven separation system showed no real performance gain over a simple harmonic mask. Errors in parameter estimates, difficulties in tracking closely spaced partials, and potential misgroupings of partials all contribute to errors in this form of separation. When all of these errors are accumulated, there is a zero net gain in (average) performance. However, the separation system that forms binary masks from STHR analyses does increase separation performance. In this case, errors in the estimated amplitudes of partial tracks do not contribute *directly* to the synthesis of a separated source. Therefore, STHR sinusoidal analysis can be used to some benefit in the context of CASA-based source separation.

# CHAPTER 6

# CONCLUSIONS

## 6.1 Summary and Conclusions

This thesis has explored the potential of signal subspace-based sinusoidal parameter estimation techniques to resolve the parameters of sinusoids that are closely spaced in frequency (i.e., harmonic collisions) in music signals. Before directly examining and establishing the techniques needed to evaluate this potential (if any), this work first aimed to quantify how prevalent harmonic collisions are, or can be, in music. Such a quantification establishes whether or not these harmonic collisions are frequent enough to warrant such attention in the first place. The prevalence of these collided harmonics was explored by analyzing symbolic music information of classical music pieces. In this small dataset, it was found that, on average, approximately 50% of a single source's nontrivial time-frequency points were interfered with by other sources. Such a proportion serves as a strong indicator that harmonic collisions do indeed play a large role in music mixtures.

Chapter 3 established a sinusoidal analysis system built upon signal subspace-based sinusoidal parameter estimators, namely the direct matrix pencil, or ESPRIT, method. To analyze time-varying signals, estimates are performed on short time-frames of the signal. Because of the computational complexity of the sinusoidal estimator, the signal is also divided into sub-bands in order to reduce computational cost. Therefore, for each frame of each sub-band of a signal, ESPRIT estimation is performed to extract the sinusoidal parameters therein. A sinusoidal tracker was presented that produces linkages of the sinusoidal estimates at each time step (frame) to produce sinusoidal tracks. The performance of the analysis system was evaluated to see whether or not it is capable of producing a valid sinusoidal representation. To do so, signals were synthesized from their representations and compared to the originals.

It was found that the presented analysis system does indeed produce a sinusoidal representation that performs well in comparison to existing techniques, namely those that derive sinusoidal tracks from short-time Fourier transforms (STFT). Finally, the analysis system was evaluated on a set of two-tone mixtures to test its ability to identify if multiple collided sinusoids are present in music mixtures that are otherwise ambiguous through direct inspection of the spectrum. It was found that in cases where harmonic collisions existed, approximately 75% of the time the parameters of two nontrivial and non-spurious sinusoids were estimated. However, due to limitations of the underlying sinusoidal model used by ESPRIT, the system also produced false estimates of multiple sinusoids being present in cases where there were not roughly 16% of the time. These collision false positives are largely attributed to the time-varying frequencies of real-world musical source harmonics.

Chapter 4 presented a multiple fundamental frequency estimation technique suited to the sinusoidal representation that the short-time high-resolution (STHR) sinusoidal analysis system produces. The fundamental frequency estimation system adopts a cancel-and-iterate strategy to estimate the fundamental frequencies of multiple sources. A method of scoring a fundamental frequency hypothesis was presented. The predominant fundamental frequency is chosen such that it is the one with a maximum score, or salience. This system uses an aggressive strategy of fully canceling the harmonics of an estimated predominant fundamental frequency in the observed sinusoidal spectrum. Furthermore, to reduce the search space of fundamental frequency hypotheses, the hypotheses are drawn directly from the observed spectrum. These design decisions were made based on the principle that if the STHR sinusoidal analysis perfectly resolves harmonic collisions, no significant effect would be made on the performance of the system even with such aggressive policies. Although these decisions can be considered somewhat naïve, or at the very least, overly optimistic, they serve to provide an analysis of the potential of STHR sinusoidal analysis in this application context.

The experimental evaluations of the multiple fundamental frequency estimation system validated the potential benefits of STHR sinusoidal analysis in this application domain. To form a basis of comparison, the proposed multiple fundamental frequency estimation technique was also used with a sinusoidal analysis front-end that derives its sinusoidal parameter estimates from the STFT spectrum. In addition, a current state-of-the-art system was used for comparison. The proposed multi-$f_0$ system with a STHR sinusoidal analysis front-end outperformed all others for two-tone musical mixtures at known musical intervals. The proposed system did, however, show a more rapid performance for increasing degrees of polyphony than the state-of-the-art system. Nevertheless, the system that used STHR-derived sinusoidal estimates far outperformed the same system using STFT-derived sinusoidal estimates.

Chapter 5 presented a music source separation method inspired by computational auditory scene analysis (CASA). The system uses grouping cues to group observed sinusoidal tracks into estimates of sources. Two subsequent synthesis methods were then presented. In one method, a simple additive synthesis of the harmonic groups is performed. If harmonics are correctly resolved and correctly grouped, such a system should produce a source synthesis with very little interference from other sources. The other synthesis method instead uses the sinusoidal representation to derive a time-frequency binary mask. This derived mask is then applied to the corresponding time-frequency representation of the mixture signal. In this case, there will still be interference from other sources in the separated signal. Time-frequency points are allocated based on local dominance of a given source. Therefore, this approach has the potential to more closely resemble an ideal binary mask if no information of the individual sources is known *a priori*.

The sinusoidal additive synthesis-based separation scheme showed no practical average performance gain over a simple harmonic mask applied to the mixed signal. The harmonic mask is one that simply allows all harmonics of a known fundamental to pass, and therefore makes no concessions for harmonic collisions. The lack of average performance gain can be attributed to a propagation of failures in the building and grouping of sinusoidal tracks. If poor estimates of sinusoidal parameters are made, even within a short span of time, sinusoidal tracks are prone to breaking. Instead of a well-behaved long-duration sinusoidal track, such harmonics are represented as a series of shorter tracks. In the resynthesis, this effect causes noticeable artifacts that largely cause any benefit of resolving closely-spaced partials to be negated. Moreover, the least-squares estimation of closely spaced harmonic amplitudes was established to be a sensitive procedure. Therefore the amplitude estimates of partials are prone to some error even if correctly tracked, also contributing to overall errors. The binary mask derived from the STHR sinusoidal analysis did produce some performance gains over a harmonic mask, however.

The series of experiments that were explored throughout this thesis, and discussed here, allow for some of the following conclusions to be drawn. The short-time high-resolution sinusoidal analysis system presented in this work does have the potential to increase performance in music signal processing applications. The most important attribute of the analysis system is that in many cases where collided harmonics exist in music mixtures, the system correctly identifies the fact that multiple sinusoids are present. With this knowledge, strong additional evidence can be used by music processing applications to deal with the fact that a given time-frequency region carries the contribution of more than one source. If the end goal of a STHR sinusoidal analysis is perfectly resolved and tracked sinusoids with near-perfect parameter estimates, the current system has a series of short comings. Many of these short comings simply stem from the fact that subspace-based parameter estimators are not perfect, especially in the sense that any real-world signal is corrupted with some amount of noise, and deviations from a fixed-frequency model constitute a form of model noise. Because sinusoidal tracks are built from frame-based estimates, this bottom-up procedure is error prone if the estimates themselves have errors. However, even though parameter estimates are not wholly accurate for partials that are closely spaced in frequency, the evidence they provide can indeed be of some benefit.

## 6.2 Future Research Directions

The work presented in this thesis allows for ongoing research in this field. One of two obvious research directions is to improve the currently proposed analysis system. Another direction is to improve the example applications presented in this work, and to develop more applications that can potentially leverage the additional information STHR sinusoidal analysis can currently provide.

It has already been discussed that the current STHR sinusoidal analysis system suffers from low-level parameter estimates that are sometimes error-prone. The propagation of these errors has a strong influence on the system as a whole. Therefore, a more robust tracking procedure could be of great benefit. The tracking method in this thesis works on relatively simple principles, and increasing its performance could be of great benefit. In terms of frame-level parameter estimates, there is little that can be done if the pole estimation portion of ESPRIT produces poor estimates of the frequencies and damping factors of the poles. However, there is the potential to make gains if estimation of amplitudes is one of the main culprits in less-than-ideal performance. In Chapter 3, the numerically sensitive least-squares procedure was addressed by using the most basic form of regularization (minimizing the $l^2$-norm of the solution). Because parameters are estimated over overlapping analysis frames, the extracted parameters of any given frame should be similar to the parameters of the frame preceding it. Therefore the extracted parameters of a preceding frame can serve as an estimate of the parameters in the current frame. Regularization can be used in this case to minimize not only the solution norm, but also the norm of the difference between a solution and its estimate. Such a system would enforce a sort of temporal smoothness on sinusoidal amplitude envelopes, and perhaps make tracks less prone to breakages. Furthermore, the Tikhonov matrix used in regularization need not be restricted to the identity matrix. If, for instance, a first-order difference operator is used, the favored solution will be smooth. A smooth solution in this case corresponds to one with a smooth spectral envelope. Placing these additional constraints on the regularization are just an example of a possible direction to further improve the system. In fact, examining regularization methods for least-squares estimation of harmonic amplitudes has ramifications in any case where the amplitudes of partials that are closely spaced in frequency need to be resolved, and good estimates of the partial frequencies exist. The frequency estimates of partials could, for example, be deduced if source fundamental frequencies are known.

The exploration of benefits of STHR sinusoidal analysis were made most evident in the case of multiple fundamental frequency estimation. This application area seems to be one that could be fruitful. The multi-$f_0$ system used in this work has on many occasions been said to work in a very basic way. Most current state-of-the-art methods operate on a more solid foundation to produce far more robust estimators. The STHR analysis system presented in this work, used as a front-end in combination with the fundamental frequency estimation procedure back-ends of more robust systems, seems natural. While the CASA-inspired separations introduced in this thesis were only moderately fruitful, once again, STHR sinusoidal analysis could potentially aid as a front-end in other source separation methodologies. The fact that STHR sinusoidal analysis can identify corrupted time-frequency regions has potential applications in polyphonic musical instrument identification as well. If instrument classifiers operate on harmonic amplitudes, as they often do, STHR analysis at the very least can help guide a classifier by indicating which harmonic locations should be treated as missing data. It is not unlikely that the parameters estimated by the system in its current state (even if they have occasional errors) can be used directly by such classifiers for some benefit. In summary, this thesis has shown there is some merit to subspace-based sinusoidal analysis of music signals. As such, a large number of music signal processing applications can be rethought to include the additional information such analysis systems can provide. As is the nature of all research, the line of thought presented in this thesis need not end here.

# REFERENCES

[1] M. Vetterli and J. Kovačević, *Wavelets and subband coding.* Upper Saddle River, NJ: Prentice Hall PTR, 1995.

[2] J. Brown and M. Puckette, “An efficient algorithm for the calculation of a constant q transform,” *Journal of the Acoustic Society of America*, vol. 92, pp. 2698–2698, 1992.

[3] J. Allen and L. Rabiner, “A unified approach to short-time Fourier analysis and synthesis,” *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[4] S. Qian and D. Chen, *Joint time-frequency analysis: methods and applications.* Upper Saddle River, NJ: PTR Prentice Hall, 1996.

[5] F. Harris, “On the use of windows for harmonic analysis with the discrete fourier transform,” *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[6] J. Flanagan, “Phase vocoder,” *The Bell System Technical Journal*, vol. 45, pp. 1493–1509, 1966.

[7] R. McAulay and T. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[8] J. Smith and X. Serra, “PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation,” in *Proceedings of the International Computer Music Conference*, 1987, pp. 290–297.

[9] M. Dolson, “The phase vocoder: A tutorial,” *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[10] J. Beauchamp, “Unix workstation software for analysis, graphics, modification, and synthesis of musical sounds,” Audio Engineering Society Preprint No. 3479, 1993.

[11] P. Depalle, G. Garcia, and X. Rodet, "Tracking of partials for additive sound synthesis using hidden Markov models," in *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing*, 1993, pp. 225–228.

[12] C. Kereliuk and P. Depalle, "Improved hidden Markov model partial tracking through time-frequency analysis," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2008, pp. 111–116.

[13] M. Lagrange, S. Marchand, M. Raspaud, and J. Rault, "Enhanced partial tracking using linear prediction," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2003, pp. 141–146.

[14] J. D. Markel and A. H. Gray, *Linear prediction of speech signals*. Berlin, Heidelberg, New York: Springer-Verlag, 1976.

[15] V. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophysical Journal International*, vol. 33, no. 3, pp. 347–366, 1973.

[16] R. Kumaresan and D. Tufts, "Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 6, pp. 833–840, 1982.

[17] D. Tufts and R. Kumaresan, "Estimation of frequencies of multiple sinusoids: Making linear prediction perform like maximum likelihood," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 975–989, 1982.

[18] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[19] R. Schmidt, "A signal subspace approach to multiple emitter location and spectral estimation," Ph.D. dissertation, Stanford University, 1982.

[20] J. Laroche, "The use of the matrix pencil method for the spectrum analysis of musical signals," *The Journal of the Acoustical Society of America*, vol. 94, p. 1958, 1993.

[21] A. Gunnarsson and I. Gu, "Music signal synthesis using sinusoid models and sliding-window ESPRIT," in *IEEE International Conference on Multimedia*, 2006, pp. 1389–1392.

[22] O. Izmirli, "Non-harmonic Sinusoidal Modeling Synthesis Using Short-time High-resolution Parameter Analysis," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2000, pp. 297–299.

[23] R. Badeau, R. Boyer, and B. David, "EDS parametric modeling and tracking of audio signals," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, vol. 2, 2002, pp. 139–144.

[24] R. Badeau, "High resolution methods for estimating and tracking modulated sinusoids. Application to music signals," Ph.D. dissertation, Telecom ParisTech, 2005.

[25] O. Gillet and G. Richard, "Drum track transcription of polyphonic music using noise subspace projection," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2005, pp. 92–99.

[26] Y. Hua and T. Sarkar, "Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 5, pp. 814–824, 1990.

[27] Y. Hua and T. Sarkar, "Matrix pencil and system poles," *Signal Processing*, vol. 21, no. 2, pp. 195–198, 1990.

[28] T. Sarkar and O. Pereira, "Using the matrix pencil method to estimate the parameters of a sum of complex exponentials," *IEEE Magazine on Antennas and Propagation*, vol. 37, no. 1, pp. 48–55, 1995.

[29] J. Downie, K. West, A. Ehmann, and E. Vincent, "The 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005): Preliminary overview," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2005, pp. 320–323.

[30] A. Klapuri, "Number theoretical means of resolving a mixture of several harmonic sounds," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 1998, p. 400.

[31] J. Brown and B. Zhang, "Musical frequency tracking using the methods of conventional and narrowed autocorrelation," *The Journal of the Acoustical Society of America*, vol. 89, p. 2346, 1991.

[32] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 22, no. 5, pp. 353–362, 1974.

[33] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, p. 1917, 2002.

[34] M. Lahat, R. Niederjohn, and D. Krubsack, “A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 6, pp. 741–750, 1987.

[35] A. Noll, “Short-time spectrum and cepstrum techniques for vocal-pitch detection,” *The Journal of the Acoustical Society of America*, vol. 36, p. 296, 1964.

[36] A. Noll, “Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate,” in *Proceedings of the Symposium on Computer Processing Communications*, vol. 779, 1969.

[37] A. Klapuri, “Automatic music transcription as we know it today,” *Journal of New Music Research*, vol. 33, no. 3, pp. 269–282, 2004.

[38] P. Martin, “Comparison of pitch detection by cepstrum and spectral comb analysis,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 7, 1982, pp. 180–183.

[39] G. Poliner, D. Ellis, A. Ehmann, E. Gómez, S. Streich, and B. Ong, “Melody transcription from music audio: Approaches and evaluation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1247–1256, 2007.

[40] Y. Chunghsin, “Multiple fundamental frequency estimation of polyphonic recordings,” Ph.D. dissertation, University Paris 6, 2008.

[41] T. Parsons, “Separation of speech from interfering speech by means of harmonic selection,” *The Journal of the Acoustical Society of America*, vol. 60, p. 911, 1976.

[42] A. Klapuri, “Multiple fundamental frequency estimation based on harmonicity and spectral smoothness,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 804–816, 2003.

[43] A. Klapuri, “Multipitch analysis of polyphonic music and speech signals using an auditory model,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 255–266, 2008.

[44] R. Maher and J. Beauchamp, “Fundamental frequency estimation of musical signals using a two-way mismatch procedure,” *Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2254–2263, 1994.

[45] A. de Cheveigné, “Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing,” *The Journal of the Acoustical Society of America*, vol. 93, p. 3271, 1993.

[46] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[47] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 982–994, 2007.

[48] G. Poliner and D. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 154–154, 2007.

[49] J. Downie, "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008.

[50] M. Bay, A. Ehmann, and J. Downie, "Evaluation of multiple-f0 estimation and tracking systems," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2009, pp. 315–320.

[51] A. Hyvarinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, no. 4, pp. 94–128, 1999.

[52] M. Helen and T. Virtanen, "Separation of drums from polyphonic music using nonnegative matrix factorization and support vector machine," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, vol. 2005, 2005.

[53] P. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[54] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proceedings of the 5th International Conference on Independent Component Analysis, LNCS*, vol. 3195, 2004, pp. 494–499.

[55] M. Schmidt and R. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Ninth International Conference on Spoken Language Processing*, 2006, pp. 2614–2617.

[56] M. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis," in *Proceedings of the International Computer Music Conference (ICMC)*, 2000, pp. 154–161.

[57] S. Dubnov, “Extracting sound objects by independent subspace analysis,” in *Proceedings of the Audio Engineering Society Conference*, vol. 17, 2002.

[58] D. FitzGerald, E. Coyle, and B. Lawlor, “Sub-band independent subspace analysis for drum transcription,” in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2002, pp. 65–69.

[59] C. Uhle, C. Dittmar, and T. Sporer, “Extraction of drum tracks from polyphonic music using independent subspace analysis,” in *Fourth International Symposium on Independent Component Analysis and Blind Signal Separation*, 2003, pp. 1–4.

[60] T. Virtanen, “Sound Source Separation in Monaural Music Signals,” Ph.D. dissertation, Tampere University, 2006.

[61] S. Mallat and Z. Zhang, “Matching pursuit in a time-frequency dictionary,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[62] R. Gribonval, “Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture,” in *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing*, vol. 3, pp. 3057–3060.

[63] M. Bay and J. Beauchamp, “Harmonic Source Separation Using Prestored Spectra,” in *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Source Separation*, 2006, pp. 561–568.

[64] A. Cemgil, “Bayesian Music Transcription,” Ph.D. dissertation, Radboud University Nijmegen, 2004.

[65] M. Davy and S. Godsill, “Bayesian harmonic models for musical signal analysis,” *Bayesian Statistics*, vol. 7, pp. 105–124, 2003.

[66] C. Févotte and S. Godsill, “A bayesian approach for blind separation of sparse sources,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2174–2188, 2006.

[67] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, “Application of the Bayesian probability network to music scene analysis,” *Computational Auditory Scene Analysis*, pp. 115–137, 1998.

[68] E. Vincent, “Musical source separation using time-frequency source priors,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 91–98, 2006.

[69] M. Every and J. Szymanski, “Separation of synchronous pitched notes by spectral filtering of harmonics,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1845–1856, 2006.

[70] A. Klapuri, “Multipitch estimation and sound separation by the spectral smoothness principle,” in *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing*, vol. 5, 2001, pp. 3381–3384.

[71] T. Virtanen and A. Klapuri, “Separation of harmonic sound sources using sinusoidal modeling,” in *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing*, vol. 2, 2000, pp. 765–768.

[72] H. Viste and G. Evangelista, “A method for separation of overlapping partials based on similarity of temporal envelopes in multi-channel mixtures,” *IEEE Transactions on Audio, Speech, and Language Process*, vol. 14, no. 3, pp. 1051–1061, 2006.

[73] T. Tolonen, “Methods for Separation of Harmonic Sound Sources using Sinusoidal Modeling,” Audio Engineering Society Preprint No. 4958, 1999.

[74] J. Woodruff, Y. Li, and D. Wang, “Resolving Overlapping Harmonics for Monaural Musical Sound Separation Using Pitch and Common Amplitude Modulation,” in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2008, pp. 538–543.

[75] D. E. Dudgeon, R. M. Mersereau, A. K. Krishnamurthy, D. C. Flint, J. Wiley, and D. Pitt, “Multidimensional digital signal processing,” *IEEE Communications Magazine*, vol. 23, no. 1, 1985.

[76] C. Avendano, “Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 55–58.

[77] C. Avendano and J. Jot, “Frequency domain techniques for stereo to multichannel upmix,” in *Proceedings of Audio Engineering Society Conference on Vitual, Synthetic and Entertainment Audio*, 2002, pp. 121–130.

[78] D. Barry, B. Lawlor, and E. Coyle, “Sound source separation: Azimuth discrimination and resynthesis,” in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2004, pp. 240–244.

[79] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures," in *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing*, vol. 5, 2000, pp. 2985–2988.

[80] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *Proceedings of the 6th International Conference on Digital Audio Effects*, 2003, pp. 209–213.

[81] S. Rickard and Z. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 529–532.

[82] H. Viste and G. Evangelista, "An extension for source separation techniques avoiding beats," in *Proceedings of the 5th International Conference on Digital Audio Effects*, 2002, pp. 71–75.

[83] A. Bregman, *Auditory scene analysis.* Cambridge, MA: MIT Press, 1990.

[84] L. Smith, "Sound segmentation using onsets and offsets," *Journal of New Music Research*, vol. 23, no. 1, pp. 11–23, 1994.

[85] M. Abe and S. Ando, "Auditory scene analysis based on time-frequency integration of shared FM and AM," in *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing*, vol. 4, 1998, pp. 2421–2424.

[86] G. J. Brown, "Computational auditory scene analysis: A representational approach," Ph.D. dissertation, University of Sheffield, Department of Computer Science, 1992.

[87] K. Kashino and H. Tanaka, "A Sound Source Separation System with the Ability of Automatic Tone Modeling," in *Proceedings of the International Computer Music Conference*, 1993, pp. 248–248.

[88] G. Brown and M. Cooke, "Perceptual grouping of musical sounds: A computational model," *Journal of New Music Research*, vol. 23, no. 2, pp. 107–132, 1994.

[89] D. Mellinger, "Event formation and separation in musical sound," Ph.D. dissertation, Stanford University, 1992.

[90] D. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 1996.

[91] G. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.

[92] M. Every and L. Litwic, "Fusing Grouping Cues for Partials Using Artificial Neural Networks," in *Proceedings of Audio Engineering Society Conference on Intelligent Audio Environments*, 2007.

[93] J. Moorer, "Signal processing aspects of computer music: A survey," *Proceedings of the IEEE*, vol. 65, no. 8, pp. 1108–1137, 1977.

[94] D. Brungart, P. Chang, B. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *The Journal of the Acoustical Society of America*, vol. 120, p. 4007, 2006.

[95] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Springer, 2005, pp. 181–197.

[96] D. Schobben, K. Torkkola, and P. Smaragdis, "Evaluation of blind signal separation methods," in *Proceedings of the International Workshop on Independent Component Analysis and Blind Source Separation*, 1999, pp. 261–266.

[97] J. Beerends, A. Hekstra, A. Rix, and M. Hollier, "Perceptual evaluation of speech quality (PESQ): The New ITU standard for end-to-end speech quality assessment part II-psychoacoustic model," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 765–778, 2002.

[98] J. Beerends and J. Stemerdink, "A perceptual audio quality measure based on a psychoacoustic sound representation," *Journal of the Audio Engineering Society*, vol. 40, pp. 963–963, 1992.

[99] J. Paulus and T. Virtanen, "Drum transcription with non-negative spectrogram factorisation," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2005, pp. 4–8.

[100] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[101] C. Févotte, R. Gribonval, and E. Vincent, "BSS EVAL Toolbox User Guide," IRISA, Rennes, France, Tech. Rep. 1706, 2005.

[102] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," *Lecture Notes in Computer Science*, vol. 4666, p. 552, 2007.

[103] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation*, 2009, pp. 734–741.

[104] A. Tkacenko and P. Vaidyanathan, "The role of filter banks in sinusoidal frequency estimation," *Journal of the Franklin Institute*, vol. 338, no. 5, pp. 517–548, 2001.

[105] J. P. C. L. da Costa, A. Thakre, F. Roemer, and M. Haardt, "Comparison of model order selection techniques for high-resolution parameter estimation algorithms," in *Proceedings of the 54th International Scientific Colloquium (IWK)*, 2009.

[106] R. Badeau, B. David, and G. Richard, "Selecting the modeling order for the esprit high resolution method: an alternative approach," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP)*, vol. 2, 2004, pp. 1025–1028.

[107] A. Tikhonov and V. Arsenin, *Solutions of ill-posed problems.* Washington, DC: Winston, 1977.

[108] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Popular, Classical, and Jazz Music Databases," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, vol. 2, 2002, pp. 287–288.

[109] J. Chowning, "The synthesis of complex audio spectra by means of frequency modulation," *Journal of the Audio Engineering Society*, vol. 21, no. 7, pp. 526–534, 1973.

[110] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2006, pp. 216–221.

[111] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, vol. 17, 1993, pp. 97–110.

[112] J. Tukey, "Quick and dirty methods in statistics. part ii. simple analyses for standard designs," in *Proceedings of the 5th Annual Convention of the American Society for Quality Control*, 1951, pp. 189–197.

[113] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.