Title

Feature preservation and negated music in a phase vocoder sound representation

Author

Apel, Theodore R.

Publication Date

2008

UNIVERSITY OF CALIFORNIA, SAN DIEGO

FEATURE PRESERVATION AND NEGATED MUSIC IN A PHASE VOCODER SOUND REPRESENTATION

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in

Music

by

Theodore R. Apel

Committee in charge:

Professor Miller Puckette, Chair
Professor Diana Deutsch
Professor Shlomo Dubnov
Professor F. Richard Moore
Professor Shahrokh Yadegari

2008

The dissertation of Theodore R. Apel is approved, and it is acceptable in quality and form for publication on microfilm or on the internet:

Chair

University of California, San Diego

2008

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My path to completing this work has been enriched by numerous people. My advisor, Miller Puckette, was always available to patiently discuss this work. I am particularly grateful for his serious consideration of my occasionally harebrained notions. Shlomo Dubnov has shared his insights with me since before he arrived at UCSD, and I am grateful for his support of the different directions of my research and his reminders to keep the musical goals at the forefront. I appreciate F. Richard Moore's rigorous yet joyful thoughts about this work. I am particularly indebted to his close reading of the manuscript of the dissertation. Diana Deutsch is an ideal model of enthusiasm for research into sonic phenomena. I thank her for welcoming me into her graduate level psychology classes and her continuing support of my work. Many thanks to Shahrokh Yadegari for his friendship during our time at CRCA, and his many words of encouragement on the dissertation.

My fellow graduate students in the computer music program at UCSD, Adam Findley, Joachim Goßmann, Cristyn Magnus, and Pei Xiang, contributed to my progress with our lively discussions. Many people at the Center for Research in Computing in the Arts also contributed to my work including Judy Brown, Harry Castle, Marek Chołoniewski, Mark Dolson, Tom Erbe, etoy, Kerry Hagan, Trevor Henthorn, Carol Hobson, Jarek Kapuscinski, Lev Manovich, Todd Margolis, Andrew May, Mike McCormick, Kazushi Mukaiyama, Ed Osborn, Gascia Ouzounian, Mark Pitt, Ron Robboy, Eliza Slavet, Carolyn Staggs, Brett Stalbaum, John Stevens, Liz Sisco and Robert Wannamaker.

The two directors of CRCA, Harold Cohen and Sheldon Brown, allowed me time to pursue this work, and the reduced fee enrollment program for regular status employees allowed me to afford graduate study at UCSD. Other people in the music department at UCSD provided encouragement: Gerald Balzano, Charles Curtis, Robert Erickson, Stephanie Ferneyhough, Steve Ford, Bonnie Harkins, Ed Harkins, János Négyesy, Peter Otto, Diana Platero, Ron Quillin, and Rand Steiger.

I am indebted to Anthony Burr for his friendship and his sage advice about the graduate program at UCSD. Email discussions with Mark Kahrs on the (ill-fated) Teager Energy chapter were much appreciated.

Jon Appleton's generous friendship and support continues to be an important part of my life. I find myself repeating many of Larry Polansky's words of wisdom to my own students. My short friendship with Dmitri Pokrovsky will always be treasured. Kathryn Alexander, Jamshed Bharucha, and Christian Wolff also contributed to my intellectual and musical development. My fellow students in the electro-acoustic music program at Dartmouth College, Michael Casey, Ko Umezaki, Ray Guillette, and John Puterbaugh remain close friends, and particular thanks go to John Puterbaugh for discussing ideas in this dissertation in its generative stages, commenting on the manuscript, and introducing me to the work of Stan Link.

Alma Zook fostered my undergraduate interest in both science and music, and Tom Flaherty welcomed me to the Pomona College music department. I am grateful to all my friends who quietly demand a doctorate for their continued friendship: Franklin Bruno, Dave Carl, Kathleen Keys, Tara Nummedal, Creighton Rosenthal, Seth Rockman, and Brian Wescott. And one, Brian Boyle, who does not. Discussions with Dave Carl about negation in the arts were helpful in formulating the ideas of the last chapter. Lee Ray introduced me to computer music by bringing me into the Center for Music Experiment at UCSD and opening my ears to the sounds that were possible. Lee's unfailing support, including his astute comments on this dissertation are well appreciated.

Judy Reilly provided a home away from home during my many trips to San Diego to complete this work as well as practical advice on making progress on the dissertation when no end appeared in sight. My parents, Joan and Dick Apel are unwavering in their support of my endeavors. My father's close proofreading of the manuscript was particularly important. My sister Karin Rocco and her children Taylor and Ella have given me a much needed change of perspective during the

course of this project.

Finally, I dedicate this work with love to Janice Neri, to whom I owe everything for her support, encouragement, and love.

VITA

1967	Born, Pasadena, California
1990	B.A., Physics Pomona College, Claremont, California
1993	M.A., Electro-Acoustic Music Dartmouth College, Hanover, New Hampshire
1994-1996	Audio Programmer and Sound Designer Annex Recording Studios, Menlo Park, California
1996-1997	Audio Programmer Lifelike Productions, Sausalito, California
1998–2003	Technical Director Center for Research in Computing and the Arts University of California, San Diego, California
2004–present	Adjunct Professor Department of Music, Boise State University, Boise, Idaho
2008	Ph.D., Music University of California, San Diego, California

PUBLICATIONS

Shlomo Dubnov and Ted Apel. 2004. *Audio Segmentation by Singular Value Clustering*, Proceedings of the International Computer Music Conference, November, Miami, Florida. pp. 454-457.

Shahrokh Yadegari, F. Richard Moore, Anthony Burr, Harry Castle, and Ted Apel. 2002. *Real-Time Implementation of a General Model for Spatial Processing of Sounds*. Proceedings of the International Computer Music Conference, Goteborg, Sweden. pp. 244-247.

Miller Puckette and Theodore Apel. 1998. *Real-time audio analysis tools for Pd and MSP*. Proceedings of the International Computer Music Conference, University of Michigan, Ann Arbor, Michigan. pp. 109-112.

Theodore R. Apel. 1993. *Transformation of Audio Signals by use of the McAulay-Quatieri Sinusoidal Model of Sound*. M.A. thesis, Dartmouth College. department of Music.

Theodore R. Apel. 1990. *Thin Layer Liquid Acoustical Filters*. B.A. thesis, Pomona College. department of Physics.

ABSTRACT OF THE DISSERTATION

FEATURE PRESERVATION AND NEGATED MUSIC IN A PHASE VOCODER SOUND REPRESENTATION

by

Theodore R. Apel
Doctor of Philosophy in Music

University of California, San Diego, 2008

Professor Miller Puckette, Chair

This dissertation presents two extensions to the phase vocoder method of sound analysis and synthesis, as well as an examination of the author's spectral subtraction audio works based on the phase vocoder. The phase vocoder technique has proved to be an effective method of "time stretching" musical sounds, however, some musical features of the original sounds are not maintained during the transformation. We consider here maintaining two of these aspects: noise levels, and sub-audio (vibrato and tremolo) characteristics within the phase vocoder. The noise levels are maintained by determining the "sinusoidality" of each spectral component, separating the spectral energy into sinusoidal and noise energy based on these sinusoidality measurements, and modulating the noise based spectrum before re-synthesizing with the traditional IFFT method. The vibrato and tremolo (sub-audio modulation) rates of a sound are maintained by modeling each channel of the phase vocoder in a higher order spectral domain, removing the modulations

in this domain, and re-imposing them after time-dilation. Finally, several of the author's audio works involve use of the phase vocoder representation of sound to subtract spectral components from a long term average of a time varying sound. These works will be considered along with their aesthetic motivations and technical implementations.

Chapter 1

Introduction

Very gradually slow down a recorded sound to many times its original length without changing its pitch or timbre at all.

Steve Reich [76]

1.1 Motivations

The phase vocoder technique is well known for its ability to “time stretch” a sound. That is, to increase the duration of a sound without changing the component frequencies of the sound. Yet the phase vocoder has proven less useful for more exotic modifications of sound due to limitations in its underlying Fourier based spectral model. Sinusoidal modeling methods, such as McAulay-Quatieri sinusoidal modeling, have been developed which overcome these limitations of the phase vocoder with much success. However, the increased computational load, complexity, and auditory artifacts of sinusoidal modeling motivate the extensions to the phase vocoder method presented here.

We introduce two extensions to the phase vocoder method of time stretching of musical sound to maintain the noise level and retain the original vibrato and tremolo characteristics. Our noise method is based on encoding the noise level of each spectral channel and re-imposing this level during synthesis. Our vibrato and tremolo method is based on a second order representation of sound in which each time-evolving spectral channel is spectrally modeled to determine the sub-audio characteristics of a sound. Our methods allow the time-stretching of musical sounds while maintaining the noise and vibrato/tremolo characteristics of the original sound without the need for more complex sinusoidal modeling.

In addition we use the phase vocoder sound representation to explore an aesthetic concept I call “negated music” in a series of sound installations and audio works. The phase vocoder is shown to be ideally suited for this type of application because of its potential for real-time implementation.

1.1.1 Advantages and Limitations of the Phase Vocoder

The phase vocoder technique of time-stretching a sound is widely admired for its quality and detail. In comparison to the time-domain “brassage” techniques that preceded it, the phase vocoder technique at first appeared to solve the problem of time-stretching [16]. As the technique became more widely used, due to the increased computational power of personal computers, its particular characteristics and limitations become well known, i. e. the coherence of phase is not maintained after reconstruction, giving the phase vocoder time-stretched sound a reverberant quality. Methods of minimizing this systemic problem with the technique are discussed in section 2.3.1.

In addition to the phase incoherence characteristic, the phase vocoder’s representation of amplitude spectra and phase or instantaneous frequency spectra presents limitations on the type of meaningful manipulation that can be achieved. The phase vocoder’s parameters of amplitude and frequency have led many users to devise manipulation techniques based erroneously on changing individual spectral components independently from their neighboring channels. That is, treating the energy of a spectral component as a sinusoidal track that can be freely manipulated instead of being part of an indivisible array of Fourier transform based spectral energy. This type of spectral manipulation typically results in unexpected results and sonic artifacts. For example, silencing a given phase vocoder channel will result in an incomplete silencing of the spectral component in that channel as its energy will be partially contained in the adjacent channels. Despite these characteristics, this type of interpretation of the phase vocoder representation is advocated consciously by many. Most notably, Trevor Wishart suggests and catalogs numerous sonic effects that exploit these types of manipulation [90]. Wishart, for example, suggests a technique he calls "spectral tracing" in which the spectral channels with the highest amplitude are retained while silencing the adjacent channels that also contain energy from the same analysis components.

Sinusoidal Modeling

The phase vocoder remains one of the central tools of computer music composition despite newer methods of spectral analysis and synthesis such as the additive synthesis sinusoidal model of McAulay and Quatieri [58], the additive plus noise model (SMS) of Xavier Serra [84], and the enhanced bandwidth model of Haken, and Fitz [27]. These sinusoidal modeling systems analyze sound as a series of sinusoidal peaks connected through time. Sound is represented as a series of "tracks" each of which represent a single sinusoidal component of the analyzed sound. After sinusoidal analysis, these tracks can be freely manipulated in a much more perceptually salient manner than those of the phase vocoder. For example, a track might represent the fundamental frequency of a sound that can be changed in amplitude or frequency separately from the other spectral components with predictable results [3]. In addition to these advantages of sinusoidal modeling systems, extensions to the systems allow the noise components of a sound to be manipulated in an intuitive manner.

Without performing any spectral transformations, great care in choosing several analysis parameters of sinusoidal modeling systems is necessary in order to achieve results that are as convincing as those of the phase vocoder. Sounds with large noise components are modeled by the McAulay and Quatieri system as many short randomly spaced sinusoidal tracks. These tracks produce the characteristic "noodling" sound of sinusoidal modeling systems. In addition, the noise portion of an SMS analysis is sometimes perceived as separate from the sinusoidal portion after synthesis.

Compared to the phase vocoder method, the higher level representations of sinusoidal modeling have many more possibilities for mis-analyzing a sound, particularly when a sound is noisy. Side lobes of spectral peaks can be mis-identified as sinusoidal tracks, peaks can be connected to the wrong peak in subsequent spectral frames, and significant spectral energy can be lost when peaks are not correctly identified. In addition, the higher level analysis procedure can take several times longer than a phase vocoder analysis.

Although working with sinusoidal modeling systems provides significant advantages for the type of research undertaken here, we restrict ourselves to extensions of the phase vocoder representation of sound due to the above characteristics of sinusoidal modeling. In fact, our noise analysis synthesis method and our vibrato retention system are both inspired by techniques developed with sinusoidal modeling systems that we are "retrofitting" to the phase vocoder system. We will see that there are particular difficulties encountered by these phase vocoder analysis limitations.

1.2 Overview of the Dissertation

This dissertion is organized as follows. Chapter 2 presents the background of the phase vocoder, the well known phase vocoder technique itself with its application to time-stretching musical sounds, and a brief discussion of techniques for reducing the artifacts created from phase incoherence. Chapter 3 presents our sinusoidality analysis and noise synthesis method of retaining noise characteristics of sounds during time-stretching. The chapter surveys existing methods of sinusoidality analysis, devises new sinusoidality methods, and quantifies a comparison between them. Example sounds from the system are presented. Chapter 4 presents our vibrato/tremolo (sub-audio modulation) extraction and re-introduction method based on a second order spectral analysis. Example sounds from the technique are presented. Chapter 5 introduces the author's idea of negated music with examples of sound installations and audio works that employ this artistic technique. The phase vocoder based technique used for these works is presented, and artistic examples involving negation are presented in order to contextualize these works.

Chapter 2

The Phase Vocoder

> *...composers/programmers are notorious for "hacking" pvoc frames in ways that would no doubt horrify dsp engineers, but which almost always produce musically interesting results.*
>
> Richard Dobson [14]

All of the sound manipulation techniques introduced in this dissertation are based on the Phase Vocoder technique of sound analysis and synthesis. We here review the relevant background in Stort-Time Fourier Analysis and Phase Vocoder based musical manipulations. The presentation here is not intended to be exhaustive or complete, as more rigorous and complete presentations are abundant in the literature [61, 68, 33, 16, 60, 6, 70, 12].

2.1 Phase Vocoder Background

In 1967 the composer Steve Reich wrote a conceptual musical composition entitled "Slow Motion Sound," in which a recorded sound was to be slowed down without changing its pitch or timbre [76]. At the time, Reich's piece was considered conceptual because a sonic realization of it could not be produced.

However, presumably unbeknownst to Reich, a computer based method of realizing this composition had been developed at Bell Labs the prior year. This method, called “the phase vocoder,” was introduced by Flanagan and Golden in their paper “Phase Vocoder” [29]. Flanagan and Golden’s work was an attempt to produce a compressed representation of speech signals by breaking the signal into frequency channels that could be encoded with less data then the original signal. The fundamental difference between the new phase vocoder and the well-established “channel vocoder” of Homer Dudley [22], was the calculation of the phase spectra in addition to the magnitude spectra.

Apart from the computational constraints of 1966, the phase vocoder was lacking several aspects that would allow it to serve as a practical tool for musical manipulation. In 1976, Portnoff improved the phase vocoder in several ways; his implementation explicitly employed the fast Fourier transform (FFT), used an appropriate windowing function, and pointed out the importance of window size [66, 67, 68].

In 1978, Andy Moorer suggested that the phase vocoder could be used specifically for musical applications and implemented a phase vocoder with musical purposes in mind [61]. As will be seen below, Moorer’s phase vocoder made an important change from Portnoff’s implementation in the method of calculating the phase spectrum values. Moorer replaced the inexact phase differentiation method with a much simpler phase difference calculation.

In addition to introducing a “tracking” phase vocoder, Mark Dolson’s 1982 dissertation entitled, “A Tracking Phase Vocoder and its Use in the Analysis of Ensemble Sounds,” showed that the Moorer phase difference method was “crucial to the effective use of the phase vocoder” and that the the old phase-derivative method was “inevitably a source of error” [15].

By the late 1980’s the phase vocoder was being used regularly in musical composition. Specifically, Mark Dolson’s phase vocoder in the CARL computer music software system, by F. Richard Moore, Mark Dolson, et al. [16, 60] and later the `pvanal` and `pvoc` programs by Dan Ellis in the csound language by Barry Vercoe [7] were in widespread use.

By the early 1990s the phase vocoder became ubiquitous in computer music composition. The phase vocoder in Tom Erbe's "Soundhack" program adapted the CARL phase vocoder to the simple GUI of the Macintosh computer, allowing many composers the opportunity to use the tool on their personal computers [25]. Another important advancement in the phase vocoder during the late 1990s was Richard Dobson's extensions to the csound phase vocoder, which allowed individual frequency bins to be manipulated in the csound environment [7]. More recently the speeds of personal computers have advanced enough to execute the phase vocoder analysis and synthesis in real time. Interestingly, the most notable use of the phase vocoder, time-streching, is not an application that is well suited to a real-time environment. That is, the interesting characteristics of real-time computation are lost in the inherently non-real-time operation of time-stretching.

2.2 The Phase Vocoder

2.2.1 Fourier Transform

One way of looking at Fourier analysis is to conceive of the analysis as comparing a periodic signal to sine and cosine waves of various frequencies. This comparison is achieved mathematically by multiplying the periodic signal by sine and cosine waves at various frequencies and, for each analysis frequency, accumulating the resultant values. This summation gives a measure of "how much" of that sinusoidal component is contained in the periodic signal. This procedure can be carried out in a computer using the discrete Fourier transform (DFT) shown here,

$$\mathrm{DFT}[x(n)] = X(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)e^{-i\omega kn}, \quad 0 \geq k \geq N-1 \tag{2.1}$$

where $x(n)$ is a digital signal, $X(k)$ is the complex valued spectrum, N is the number of samples in the transformed signal, and $\omega = 2\pi/N$. Here we use Euler's relation $e^{i\theta} = \cos\theta + i\sin\theta$ to simplify the addition of the sine and cosine terms. That is, one complex multiplication by $e^{-i\omega kn}$ gives us both the sine and cosine components at that frequency. As can be seen from equation 2.1, a signal $x(n)$ is multiplied by sine and cosine waves at frequencies k, over each time point in the analysis range. For each frequency k, results of these multiplications are added together to produce a measure of how much this frequency is contained in the original signal. The resulting spectrum $X(k)$ is composed of complex numbers with the real part of the number storing the co-sinusoidal magnitude measurement and the imaginary component containing the sinusoidal magnitude measurement.

In order to increase the computational efficency of the discrete Fourier transform, the fast Fourier transform (FFT) was developed in 1965 by Cooley and Tukey [11]. The FFT achieves an increase in computational speed by taking advantage of redundancies in the DFT. The matrix multiplication shown above can be factored into a product of very sparse matrices that reduces the number of necessary multiplications. The FFT is most efficient when the number of samples being transformed is a power of two, but lesser increases in efficiency can also be found with other lengths.

2.2.2 Windowing

In order to analyze the frequency content of a sound that is changing in time, a series of FFT's may be taken. In this section I will look at how this is achieved through overlapped "windows" of sound. Because the Fourier transform expects the signal to be a periodic signal, an arbitrary short section of sound will not necessarly fulfill this requirement because the discontinuities between the beginning and end of the segment create FFT energy in many other bands based on the size of the segment. In order to reduce this problem, methods of smoothing the end points of the analysis period have been devised. A windowing signal is multiplied by the analysis signal section in order to smooth out and connect the two ends of the analyzed sound. This process, of course, modifies the analyzed signal, but it allows arbitrary length signals to be analyzed such that the spectrum contains energy in the expected bands. The shape of this windowing function has been extensively studied and will not be discussed here except to note that the Hamming and Hanning windows are typical for use with the phase vocoder, the Hanning window reaching zero at the ends, and the Hamming window not quite reaching zero.

In addition to modifying the frequency content of a section of sound, the windowing process creates an amplitude modulation in the re-synthsized sound. This amplitude modulation can be abated by overlapping the analysis periods. During resynthesis, the overlapped sections are added together. The more the windowed analyses are overlapped, the less amplitude modulation will result. Typically, an overlap factor of two, four, or eight is used.

2.2.3 STFT Frames

At this point we can consider a representation of sound consisting of a series of windowed and overlapped FFT analyses. We call each of these analyses a short-time Fourier transform frame (STFT). This representation of sound can be converted back to a time-domain bit stream by inverse Fourier transforming each windowed spectrum, dividing it by its window function, and adding the appropriate overlapped samples. This process does not yet allow us to time-stretch, frequency shift, or otherwise manipulate a sound. As will be seen, a conversion to polar coordinates and a careful consideration of the phases of each bin will allow us to achieve these modifications.

2.2.4 Instantaneous Frequency

Only one step remains in order to time-stretch a sound with the phase vocoder. We can think of the time-stretching problem as first calculating new Fourier transform frames of data *between* analyzed Fourier frames and then inverse Fourier transforming the sound with these new frames interspersed between the analysis frames.

In order to calculate these frames we first convert the complex valued Fourier transform data into polar coordinates from rectangular coordinates. These polar coordinates are the magnitude and phase of the spectral components. The magnitude is calculated by

$$|X(k)| = \sqrt{\mathrm{Re}(X(k))^2 + \mathrm{Im}(X(k))^2}. \tag{2.2}$$

The phase $\theta(k)$ of $X(k)$ is calculated by

$$\theta(\mathrm{k}) = -\arctan\left(\frac{\mathrm{Im}(X(k))}{\mathrm{Re}(X(k))}\right), \tag{2.3}$$

where $\theta(k)$ are the principle values of the phase bounded by π and $-\pi$. This bounding will be discussed below.

From this representation, we can calculate a new spectral frame for some arbitrary time position n_p between two spectra $X_{(n-1)}(k)$ and $X_n(k)$. The new amplitude spectrum $|X_{n_p}|$ is calculated by interpolating magnitudes between $|X_{(n-1)}(k)|$ and $|X_n(k)|$.

$$|X_{n_p}| = (1-p)(|X_{(n-1)}(k)|) + p(|X_n(k)|), \tag{2.4}$$

where p is the position between frames expressed as a number between 0 and 1.

It is here, finally, that we come to the essence of the phase vocoder. The new phase spectrum cannot simply be interpolated between the neighboring values because this would in effect change the rate at which the phase advances. New phase values must be created that maintain the rate of phase advancement for each channel. So, the new phase spectrum is calculated by looking at the rate of change of the phase in the region, the "instantaneous frequency," and advancing the phase an amount that would be advanced at that rate. The instantaneous frequency is calculated by taking the difference between frames n and $(n-1)$.

$$\Delta\theta_n(k) = \theta_n(k) - \theta_{(n-1)}(k) \tag{2.5}$$

This instantaneous frequency is an approximation of the rate of change of the phase. In the original phase vocoder, the instantaneous frequency was calculated by a phase derivative

$$\dot{\theta}_n(k) = \frac{\mathrm{Re}(X_n(k))\mathrm{Im}(\dot{X}_n(k)) - \mathrm{Im}(X_n(k))\mathrm{Re}(\dot{X}_n(k))}{\mathrm{Re}(X_n(k))^2 + \mathrm{Im}(X_n(k))^2}, \tag{2.6}$$

where $\mathrm{Im}(\dot{X}_n(k))$ and $\mathrm{Re}(\dot{X}_n(k))$ are the derivatives of the real and imaginary components. This method was used by Flanagan and Christensen [28], but is no longer used because the original phase can be reconstructed exactly using the phase difference method [15].

Using the phase difference method, each spectral frame of the new phase value is calculated from the instantaneous frequency $\Delta\theta_n(k)$ by summing all the previous instantaneous frequencies. So, for the new position n,

$$\theta_n(k) = \theta_{(n-1)}(k) + \Delta\theta_n(k) \tag{2.7}$$

where $\theta_{(n-1)}(k)$ is the prior output of this equation.

An addition to this procedure must be made because of the bounding between $-\pi$ and π of the phase component when the FFT analysis is performed. This bounding creates incorrect jumps in the instantaneous phase calculation. It is necessary to create a continuous phase function from the bounded phase function before any of the above phase calculations are completed. This procedure is called phase unwrapping, and is performed by adding and subtracting 2π at the discontinuities in order to make the function continuous [60].

2.3 Phase Vocoder Extensions

2.3.1 Phase Coherence

There is no objective way of defining a "correct" method of time-stretching sound because there are many ways of defining auditory time-stretching. While the phase vocoder method is generally considered satisfactory compared to other methods such as granular methods, it does have notable characteristics that appear to be artifacts of the analysis method, and not characteristics of an ideal time-stretched audio method. As noted by Puckette [71], the phase vocoder analysis/synthesis process does not maintain relationships between the phases of each bin. Quatieri uses the term "horizontal coherence" to denote the phase vocoder's ability to create phase functions without discontinuities in time, and "vertical coherence" as the coherent relationships between phase components at different frequencies. It is this loss of vertical coherence that is responsible for the phase vocoders "reverberant" or "chorused" effect. The slight mistunings between components create sounds similar to sounds in a reverberant environment.

Independently from each other in 1995, Puckette proposed his "phase-locked vocoder" [71] and Quatieri, Dunn and Hanna [73] proposed their "instantaneous invariance" method of maintaining vertical coherence at distinct time instances for reducing transient smearing.

The instantaneous invariance method of Quatieri et al. works by phase-synchronizing (or phase-locking) phase components at the specific times when salient events occur in the sound. This method is well-suited to speech sounds where the signal is characterized by onsets of speech events, but requires the explicit computation of event time locations. Duxbury, Davies, and Sandler adapt this idea of phase coherence at particular transient times to musical signals [24]. Röbel suggests an alternate method of transient detection for use in such systems based on phase-synchronizing phase components around individual spectral peak onsets [77].

Puckette's phase-locking method, demonstrated below, reduces reverberation by increasing phase coherence around spectral peaks [71]. The method does not require peak-picking and avoids any heuristic algorithms in favor of performing the same operation on all spectral bins.

Laroche and Dolson [47, 46, 48] extended Puckette's method by explicitly finding amplitude peaks in the spectrum and using them as a guide to determining the dominant spectral regions. This method is called "rigid phase-locking." Laroche and Dolson report superior results to those of Puckette at the cost of a more complex algorithm.

Dorran, Lawlor, and Coyle have recently proposed a hybrid method of increasing vertical coherence in the phase vocoder [17]. Their system takes advantage of limitations of phase coherence perception to nudge phase values closer to a coherent state. They report significant improvement for voice signals over the Laroche and Dolson method.

Puckette's Phase-Locking

Puckette notes that a single sinusoid typically centers its energy around 3 or 4 channels in the phase vocoder analysis [71, 72]. A sinc function shaped main lobe (Dirichlet kernel) of spectral bins all contain significant energy from the single analyzed sinusoid, and because they represent a single sinusoid, these bins should ideally have the same phase. Since the phase spectra of a time-stretched sound from the phase vocoder are calculated with no regard to their vertical coherence, and the phase values are calculated by accumulating instantaneous frequency values, they shift and become incoherent with respect to one another. To account for this, instead of attempting to modify the phase vocoder analysis method, Puckette applies "a-posteriori constraints to the synthesis phases" [49]. The first step in Puckette's method is to calculate a new spectrum that consists of the average of each successive complex valued spectral bin k. We will call this new complex spectrum $X(l)$, where

$$X(l) = \frac{X(k-1) + X(k) + X(k+1)}{3} \tag{2.8}$$

for all $k, k = l$. Here, each successive spectral bin is in phase. If the time origin of the FFT is at the beginning of an FFT window, the phase will unwrap in opposite directions for even and odd bins and every other term in the above equation will need to be negated. In this equation, the term $X(k-1)$, $X(k)$, or $X(k+1)$ with the greatest magnitude will have the largest effect on the phases of the new spectrum $X(l)$. In other words, spectral peaks will tend to impart their phase on nearby bins.

Finally, the modified phases of $X(l)$ are combined with the original magnitude of $X(k)$ to produce the desired rectangular phase-locked spectrum,

$$X(m) = |X(k)|e^{i(\angle(X(l))}. \tag{2.9}$$

Several other modifications to the phase vocoder technique have been proposed to impove the analysis or expand the transformational ability of the phase vocoder. We briefly mention them here.

2.3.2 Other Phase Vocoder Extensions

In addition to the phase coherence technique mentioned above, Laroche and Dolson suggested that partitioning the amplitude spectra of the STFT around spectral peaks allows new types of sound transformations. All of the bins associated with a spectral peak could be transformed together relative to the other sets of spectral bins [48].

Marchand proposed a method of improving the instantaneous frequency estimates of the phase vocoder by analyzing the derivatives of the time-domain signal in addition to the traditional analysis [55]. Marchand suggests that improved time resolution can be achieved by employing a shorter analysis window without a loss of frequency resolution.

Ferreira devises a novel time-stretching algorithm for speech signals in which the STFT is interpolated along the frequency axis before conversion to the time domain [26]. A new sampling rate is used to compensate for the change of window length.

We have now presented the phase vocoder and significant extensions in a form that is optimal for many musical applications. In the next chapter we will extend this technique to retain noise characteristics of sound during time-stretching.

Chapter 3

Sinusoidality Analysis and Noise Synthesis in a Phase Vocoder Sound Representation

But even when a musical tone continues with uniform or variable intensity, it is mixed up, in the general methods of excitement, with certain noises, which express greater or less irregularities in the motion of the air.

Hermann Ludwig Ferdinand von Helmholtz [39]

3.1 Introduction

When a sound is lengthened with the phase vocoder, the noise aspects of the sound tend to become pitched. Under extreme time lengthening, all noisy aspects of the original sound are transformed into stable sinusoidal components. This behavior is consistent with the phase vocoder's modeling of short-time Fourier transform (STFT) energy as exclusively sinusoidal energy. The purpose of the research presented in this chapter is to extend the phase vocoder (PV) representation of monophonic sounds to allow for the original noise characteristics to be maintained during PV lengthening and other PV based sound manipulations. In this chapter, both the analysis and synthesis procedures of the phase vocoder are expanded in order to maintain the noise aspects of sound.

In order to derive the noise characteristics of sound in a PV representation, methods of analyzing the "sinusoidality" of STFT channels will be employed. STFT channels with high sinusoidality are composed of predominantly sinusoidal energy and should exhibit sinusoidal characteristics during synthesis. STFT channels with low sinusoidality have predominantly noise energy and should be synthesized with noise characteristics. In this chapter, (i) techniques for measuring the sinusoidality of STFT channels are surveyed, (ii) modifications to existing as well as new methods of sinusoidality analysis are presented, (iii) measurements of sinusoidality are compared, (iv) a new method of combining pitched and noisy components during PV synthesis is presented, and (v) resultant example sounds are shown.

3.2 Sinusoidality Analysis

The phase vocoder is built upon the Fourier transform. The Fourier transform analyzes a signal for sinusoidal components, that is, the output consists of coefficients to a sinusoidal basis function. The sinusoidal nature of Fourier analysis is ill-suited to the analysis of noisy sounds or sounds with significant noise components because these components are analyzed as many rapidly varying sinusoidal components. While this analysis can be used to reconstruct the original sound, temporal manipulation of this Fourier energy using phase vocoder techniques typically results in the noise characteristics of the sound taking on a pitched or sinusoidal character.

In order to alleviate this problem of noisy components of a sound being misinterpreted in phase vocoder manipulations, we will attempt to analyze the phase vocoder representation for its noise characteristics, and to create these noise levels during a phase vocoder synthesis. This analysis will consist of calculating the sinusoidality of each Fourier analysis channel and, using this measure, determining noise levels for each channel during synthesis.

The idea of determining the nature of STFT channel energy for the purpose of enhancing subsequent noise synthesis is relatively new for musical applications. Many speech analysis/synthesis methods use a voiced/unvoiced decision making algorithm as part of the analysis method. These methods have been extended to calculate the voicing coefficient of individual spectral channels [34]. As we will see, the idea of voicing coefficients in speech signal processing is very similar to the idea of the sinusoidality coefficients discussed here. The sinusoidality of an STFT channel represents the degree to which the energy of each spectral bin consists of sinusoidal based energy. A low sinusoidality measure indicates that the energy in that band is based on random or noise signals. Here we will follow the notation and conventions set by Peeters and Rodet [31] where $\Gamma(n, k)$ is the sinusoidality of spectral bin k at time frame n, and $\Gamma(n, k)$ varies between high sinusoidality of 1 and 0 for low sinusoidality or high-noise content.[1] All of the existing and new sinusoidality measurement algorithms presented here will be scaled to this range.

The sinusoidality analysis methods are based on a FFT window length of 1024 with a sampling rate of 44 100 Hz. This limits the frequency resolution between spectral bins to approximately 43 Hz. Many of the sinusoidality analysis techniques presented here rely on the energy of a single bin coming from a single spectral component. As polyphonic musical signals may have multiple spectral components within one spectral bin, the sinusoidality measures presented here will perform best with monophonic signals or signals that do not have more than one

[1] Dubnov defines a similar “noisality” in which high noisality denotes random energy in that band [21]. We will not use this term due to its redundant relationship to our sinusoidality definition.

spectral component in a bin.

In the next section existing sinusoidality measures are reviewed and modified and new sinusoidality measures are presented. Three sinusoidality measures based on the amplitude spectrum of the STFT are presented, followed by two sinusoidality measures based on the phase spectrum. Next, a method based on correlation of the complex STFT spectrum to the window function is presented, and then a method based on the changes in narrowband temporal envelopes. Finally, a new method based on the harmonic structure of the sound is presented. In addition, sinusoidality techniques that are not implemented as part of this study are briefly surveyed.

3.2.1 Power Spectrum Sinusoidality

The first three sets of sinusoidality coefficients are derived from the amplitude spectrum of the STFT, $|X_n(k)|$. Our first is a scaled version of the power spectrum. An estimate of the power spectrum (also known as the power density spectrum) can be calculated from the periodogram of a signal [38]. The power spectrum is simply the square of the amplitude spectra, $|X_n(k)|^2$. The idea behind this measure is that peaks in the power spectrum are composed of sinusoidal energy. If we make the assumption that sinusoidal components of a musical sound produce peaks in the power spectrum that are higher than peaks produced by noise components, we can use these peaks as indicators of sinusoidal energy. These sinusoidality coefficients can be calculated by simply squaring the amplitude spectrum and scaling this power spectrum by the highest amplitude in that frame:

$$\Gamma_p(n,k) = \frac{|X_n(k)|^2}{\underset{k}{\operatorname{argmax}}\left(|X_n(k)|^2\right)}. \tag{3.1}$$

Here $\Gamma_p(n,k)$ are the power spectrum sinusoidality coefficients for time frame n and spectral bin k, and $\max(|X_n(k)|^2)$ is the single highest amplitude bin of each power spectrum. Γ_p varies from 1 for the highest peak to 0 for no energy, here treated as noise energy.

Scaling the power spectrum in this manner produces sinusoidality coefficients that have at least one value that is set to the maximum value of one regardless of how much sinusoidal energy is present in the spectral frame. If no sinusoidal peaks are present, for example, noise energy is incorrectly scaled to a high sinusoidality value. Later, this problem is alleviated before synthesis by linearly scaling all of the sinusoidality values by an overall measure of tonality derived from the spectral flatness measure in section 3.4 below. However, it should be noted that this sinusoidality measure, and others that are normalized in this way are susceptible to this problem.

The assumption that sinusoidal energy peaks are higher than peaks from noise energy is not true for many musical sounds. For example, the attack portion of many musical instrument sounds typically contains power spectrum peaks of noise energy of higher magnitude than the peaks composed of sinusoidal energy. We will see this confirmed when we compare sinusoidality measures in section 3.3.

3.2.2 Power Persistence Sinusoidality

The next sinusoidality measure is a modification of the power spectrum technique. The new power persistence sinusoidality measure is calculated by multiplying each amplitude spectral bin by the two prior amplitude spectral bins for every channel, and normalizing the result between 0 and 1. Here we have:

$$\Gamma_{pp}(n,k) = \frac{\left(\left|X_{(n-2)}(k)\right|\right)\left(\left|X_{(n-1)}(k)\right|\right)\left(\left|X_{(n)}(k)\right|\right)}{\underset{k}{\operatorname{argmax}}\left((|X_{(n-2)}(k)|)(|X_{(n-1)}(k)|)(|X_{n}(k)|)\right)}. \tag{3.2}$$

Where $\Gamma_{pp}(n,k)$ are the power persistence spectrum sinusoidality coefficients. The idea here is that amplitude spectral components that remain prominent over three spectral frames will produce a high sinusoidality value for that channel as sinusoidal components will tend to be slowly changing relative to noise components. As shown in section 3.3, this technique appears to create an improved set of sinusoidality coefficients over the power method for sinusoidal signals whose amplitude maxima are lower than the noise power peaks.

3.2.3 Sigmund Sinusoidality

As part of his pitch tracking algorithm called “Sigmund”, Puckette, identifies spectral peaks of the amplitude spectrum [70]. Puckette zero pads the STFT spectrum by a factor of two, windows each time frame with a half cosine wave window, and then declares a spectral peak to exist in bin k when:

$$|X(k)| > 0.6\left(|X(k-2)| + |X(k+2)|\right). \tag{3.3}$$

This method of peak identification takes advantage of the difference in magnitude between a peak and the magnitude of the adjacent valleys in a cosine windowed and zero padded spectrum. We here adapt the sigmund method to create a sinusoidality measure for each spectral bin:

$$\Gamma_s(k) = \frac{\left(\frac{|X(k)|}{0.6(|X(k-2)|)(|X(k+2)|)}\right)}{\underset{k}{\operatorname{argmax}}\left(\frac{|X(k)|}{0.6(|X(k-2)|)(|X(k+2)|)}\right)}, \tag{3.4}$$

where $\Gamma_s(k)$ are the power persistence spectrum sinusoidality coefficients. The results of the sigmund sinusoidality method do not appear to be better than the power sinusoidality method. As sigmund is designed for identifying spectral peaks, it may be ill suited for determining sinusoidality coefficients near spectral peaks.

3.2.4 Charpentier Sinusoidality

The phase spectrum of the STFT can also be employed to create sinusoidality measures. As will be seen, by using consecutive frames of STFT data the instantaneous frequency spectra and the phase acceleration spectra can each be used to create sinusoidality measures. We will look at each of the possibilities in turn.

The phase vocoder method calculates an approximation of the instantaneous frequencies of spectral components from the difference in phase between consecutive spectral frames. These frequency values can be thought of as refinements to the nominal center frequency values of each bin. Charpentier devised a pitch detection algorithm that groups spectral bins based on their similar instantaneous frequencies [10]. Charpentier notes that a sinusoid will exhibit energy in at least three adjacent spectral bins, and that the instantaneous frequency of these bins will be correlated around the true frequency of that sinusoidal component. Figure 3.1 shows the amplitude spectrum of a synthetically generated harmonic sound and the corresponding phase difference spectrum both scaled between 0 and 1. It can be clearly seen that the phase difference values stabilize around the areas with sinusoidal energy. Dressler [18] formalizes Charpentier's phase difference

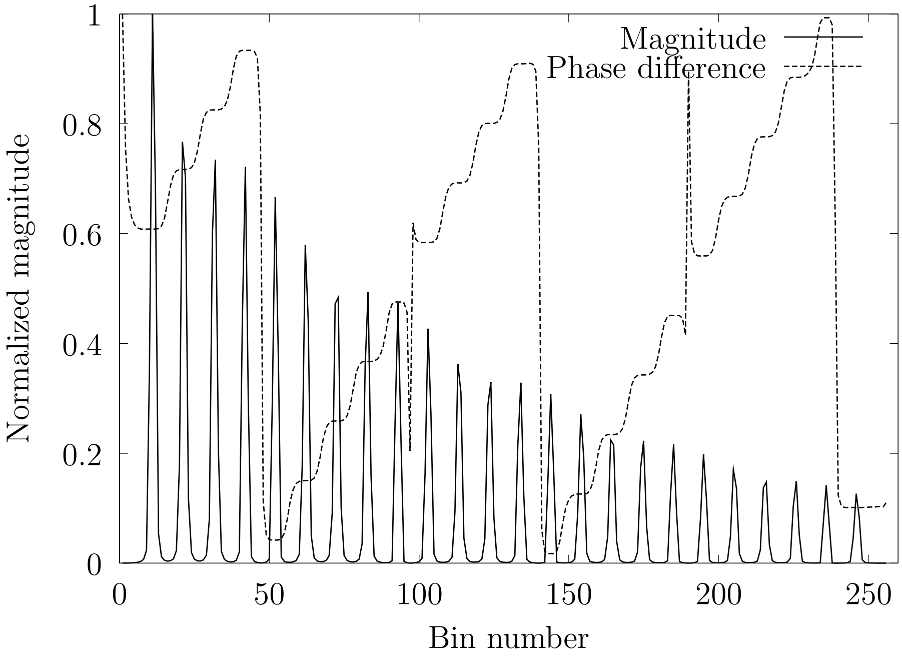

Figure 3.1: Amplitude spectrum and phase difference spectrum of a harmonic sound. Visible are the similar phase difference values around the sinusoidal energy peaks.

method as a decimal offset $\kappa(k)$ to the integer bin number k. We derive $\kappa(k)$ in terms of phase here. The center frequency $f(k)$ of each spectral bin k can be found by multiplying k by the sampling rate SR and divided by the hop size N.

$$f(k) = k\left(\frac{SR}{N}\right). \tag{3.5}$$

This offset $\kappa(k)$ is added to the integer bin number k in equation to calculate the instantaneous frequency in Hz.

$$f(k) = (k + \kappa(k))\left(\frac{SR}{N}\right). \tag{3.6}$$

Here $\kappa(k)$ is calculated from the offset between the measured phase $\theta_n(k)$ and the expected phase calculated from the prior phase of that bin, here called $\hat{\theta}_n(k)$. This offset in radians is:

$$\theta = (\theta_n(k) - \hat{\theta}_n(k)), \tag{3.7}$$

We can transform this offset into decimal units of bin by multiplying by $N/2\pi R$:

$$\kappa(k) = \left(\frac{N}{2\pi(SR)}\right)(\theta_n(k) - \hat{\theta}_n(k)). \tag{3.8}$$

Charpentier uses each value of $\kappa(k)$ to detect harmonics of a signal by declaring a bin k to contain a harmonic when $\kappa(k-1)$ and $\kappa(k+1)$ are "sufficiently close" to $\kappa(k)$. We will interpret Charpentier's idea by calculating the "spectral irregularity" of the instantaneous frequency values $\kappa(k)$.

Spectral Irregularity was devised by J. Krimphoff as a way of quantizing the overall smoothness of a spectrum [42]:

$$IRR = \sum_{k=2}^{N-1}\left| |X(k)| - \frac{|X(k-1)| + |X(k)| + |X(k+1)|}{3} \right|. \tag{3.9}$$

From this summation we can see that any three adjacent spectral bins with a constant slope contribute nothing to the overall irregularity summation. Only changes in slope over three frames contribute to the spectral irregularity. Our new measures will create a set of local irregularity coefficients by removing the summation over the entire spectrum and calculating irregularity solely for each spectral bin. In Krimphoff's formulation across spectral bins, this would be

$$IRR(k) = \left| |X(k)| - \frac{|X(k-1)| + |X(k)| + |X(k+1)|}{3} \right|, \tag{3.10}$$

for each bin [2]. In our formulation, however, we will look at the local irregularity of instantaneous frequency and not amplitude. That is:

$$IRR(k) = \left| \kappa(k) - \frac{\kappa(k-1) + \kappa(k) + \kappa(k+1)}{3} \right|. \tag{3.11}$$

Here we can see that three points that cluster around a particular instantaneous frequency value will produce a low irregularity coefficient for that bin, and conversely, large variations in instantaneous frequency between spectral bins will produce a larger irregularity coefficient.

We can put this measure into our sinusoidality framework by normalizing the results and subtracting from one.

$$\Gamma_c(k) = 1 - \frac{IRR(k)}{\underset{k}{\operatorname{argmax}}(IRR(k))}. \tag{3.12}$$

The Charpentier sinusoidality method devised here exhibits the lowest error on one of our test signals with a single sine tone and broadband noise. This result will be discussed in section 3.3 below.

3.2.5 Phase Acceleration Sinusoidality

Our second phase spectrum based sinusoidality measure also employes the phase difference method of instantaneous frequency computation. Settel and Lippe devised a “band-limited frequency dependent noise gate” as a method of separating stable STFT channels from non-stable channels [86]. Settel and Lippe suggest that, “pitched components in the input signal tend to be stable and can thus be independently boosted or attenuated.” Their method defines a threshold for changes in the

[2] Two sinusoidality measures were devised using this local amplitude irregularity both across bin number and across time. Neither method produced results better than the power spectrum sinusoidality measure, and the results are not presented here.

phase vocoder's instantaneous frequency values between spectral frames. Changes in instantaneous frequency below a given threshold are considered stable. Because sinusoidal components should have a relatively constant instantaneous frequency, these slowly changing components are typically sinusoidal and pitched. A mathematical presentation of the Settel and Lippe method is presented by Arfib, Keiler, and Zölzer [6]. Here we formulate this idea in terms of a variable sinusoidality instead of a threshold for each bin, and in terms of "phase acceleration" because the difference between instantaneous frequency values can be thought of as the acceleration of the phases of that frame.

As was shown in chapter 2, the instantaneous frequency is calculated for each channel from the phase difference between frames n and $(n-1)$ by,

$$\Delta\theta_n(k) = \theta_n(k) - \theta_{(n-1)}(k) \tag{3.13}$$

where $\theta(k)$ are the principle values of the phase and are bounded by π and $-\pi$. The phase acceleration, $\Delta\Delta\theta_n(k)$, is the difference between the instantaneous frequency and the prior instantaneous frequency,

$$\Delta\Delta\theta_n(k) = \Delta\theta_n(k) - \Delta\theta_{(n-1)}(k). \tag{3.14}$$

The above two equations are combined to show the phase acceleration in terms of phase:

$$\Delta\Delta\theta_n(k) = \big(\theta_n(k) - \theta_{(n-1)}(k)\big) - \big(\theta_{(n-1)}(k) - \theta_{(n-2)}(k)\big)\,, \tag{3.15}$$

so that,

$$\Delta\Delta\theta_n(k) = \theta_n(k) - 2\theta_{(n-1)}(k) + \theta_{(n-2)}(k). \tag{3.16}$$

We can normalize this phase acceleration to our nominal sinusoidality range of zero to one to create our final sinusoidality coefficients:

$$\Gamma_{pa}(n,k) = \left(1 - \left(\frac{\theta_n(k) - 2\theta_{(n-1)}(k) + \theta_{(n-2)}(k)}{2\pi}\right)\right)^p, \tag{3.17}$$

where $\Gamma_n(k)$ is the resultant sinusoidality spectrum for frame n. This spectrum shows values near 1 for stable channels and values tending toward 0 for unstable channels. This sinusoidality measure tends to produce sinusoidality coefficients near one for both stable and non-stable components. The variable p in the above equation is used to counteract this tendency. By raising each coefficient to a small integer power of itself, typically $p = 4$, this overall tendency is significantly abated.

Duxbury, Davies, and Sandler proposed a frequency dependent threshold to improve the Settel and Lippe method [23]. They note that for any given value of Γ_{pa}, low frequency channels tend to be selected as stable and high frequency channels selected as non-stable. They alleviate this tendency by choosing a different threshold value for each octave sub-band created with six constant-Q filters. Duxbury reports overcoming the frequency dependency tendency. However, this method requires setting six threshold values instead of one.

Here we propose a new method of reducing frequency dependency in terms of our sinusoidality spectrum $\Gamma_{pa}(k)$. In order to reduce the sinusoidality measure $\Gamma_{pa}(k)$ for high k, we will change the slope of the stability spectrum by multiplying each $\Gamma_{pa}(k)$ by a scaled version of the channel number k. Our new stability spectrum is:

$$\Gamma_{pa}(k) = \left(1 - \left(\frac{\theta_n(k) - 2\theta_{(n-1)}(k) + \theta_{(n-2)}(k)}{2\pi}\right)\right)^p (M) \left(\frac{k}{k_{max}}\right), \tag{3.18}$$

where M is a slope constant with a heuristically determined value of approximately -1.024. This correction to the stability spectrum avoids the constant-Q filter calculation and multiple threshold calculations of the Duxbury et al. method. Figure 3.2 shows the phase acceleration sinusoidality coefficients $\Gamma_{pa}(k)$ of a harmonic sound along with the corresponding power spectrum sinusoidality coefficients $\Gamma_p(k)$. The slope M is set to -1.024 and $p = 4$.

The phase acceleration sinusoidality measure performs better than the amplitude spectrum based sinusoidality measure for sounds with many sinusoidal components in broadband noise. In addition, it can be multiplied by the power

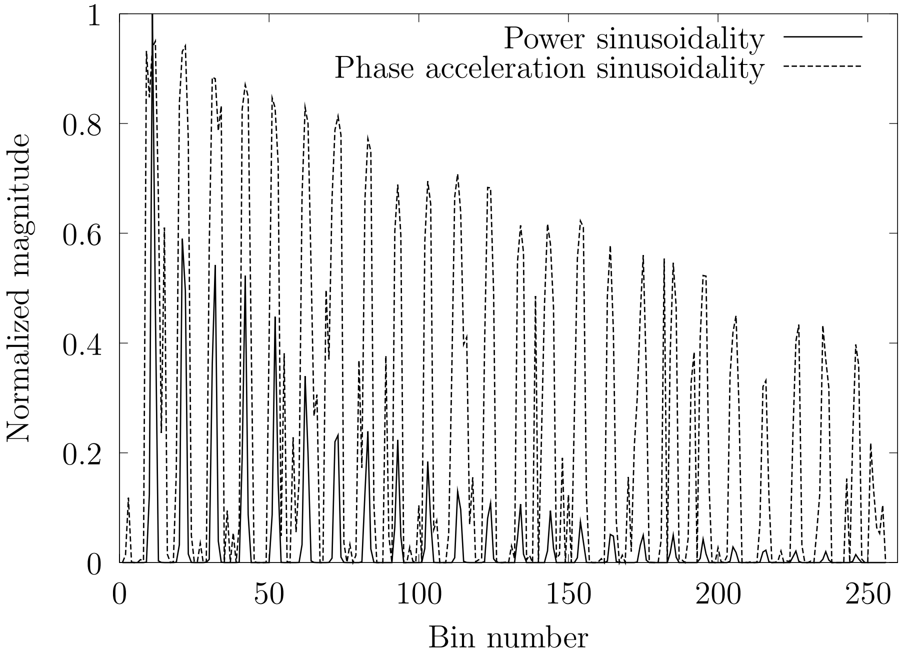

Figure 3.2: Phase acceleration sinusoidality coefficients and power spectrum sinusoidality coefficients of a harmonic sound.

persistence measure. As we will see in section 3.3, this combined measure performs well with sinusoidal signals that are rapidly changing in frequency.

3.2.6 Cross-Correlation Sinusoidality

As part of a speech analysis/synthesis system, Griffin and Lim proposed a method of labeling spectral components of a speech signal as voiced or unvoiced [34]. This method, based on correlating spectral peaks to the shape of a windowed sinusoid at target frequencies, was adapted as a sinusoidality measure by Rodet [78], and its characteristics were studied by Peeters [31]. As the shape and position of a sinusoid in the spectral domain should ideally match the shape of the windowing function of the transform translated to match its frequency, the correlation between these spectral shapes should be high for sinusoidal components.

This idea can be expressed in the frequency domain as:

$$\Gamma_{cc}(k) = \left| \sum_{k, |\omega - \omega_k| < W} X(\omega_k) H(\omega - \omega_k) \right| . \tag{3.19}$$

Here $X(\omega_k)$ is a STFT frame, $H(\omega)$ is the complex spectral window of the sinusoid, and W is the bandwidth of the sinusoidal spectral window.

Typically, due to its computational complexity the correlation is only performed within a small range W around spectral peaks found by other methods. In our case, as we have no peak picking method, this correlation is calculated for all bins k of the STFT. How the cross-correlation method comparisons with other sinusoidality measures will be presented in section 3.3.

3.2.7 Narrowband Variance Sinusoidality Measure

Hanna and Desainte-Catherine developed a method of detecting sinusoidal components based on the variance of narrowband temporal envelopes [36, 37]. They suggest that the temporal envelope will vary more for bands that contain noise than those with sinusoidal components.

Their method operates by creating a series of narrowband filters on the STFT. These bandpass filters are typically three to 5 bins wide. Each filtered STFT is converted back to the time-domain by an inverse FFT. The total variance of each of these filtered segments is then calculated and used to create a variance spectrum in which channels with high variance represent noisy components and channels with low variance represent sinusoidal components.

Hanna and Desainte-Catherine's algorithm features many more details of implementation that are not presented here or implemented in our version of the algorithm. Our implementation uses a constant phase spectrum for each narrowband calculation and uses a single frame length for the final variance calculations. Because of these simplifications our results should not be taken as comparable to the results presented by Hanna and Desainte-Catherine. Results of our implementation of this algorithm are presented in section 3.3.

3.2.8 Harmonic Sum Spectrum Sinusoidality

Many voicing decision algorithms for speech signals are based on the harmonicity of the components of speech [73]. The use of harmonic relationships to determine sinusoidality coefficients is also possible with sinusoidal modeling systems in which the frequency of individual spectral components is known explicitly [85]. The measure of the overall “harmonicity” of a STFT frame is a common feature, that typically requires the explicit calculation of a single fundamental frequency (f_0) by peak picking or other technique [80].

As our project here is based on the traditional PV parameters of amplitude, phase and/or instantaneous frequency, those methods that require higher order analysis such as spectral peak picking or fundamental frequency analysis are precluded. With these restrictions, we have created a new non-parametric sinusoidality analysis technique for STFT frames that relies on the harmonic relationships present in the sound, but without explicit computation of these frequencies.

Musical instrument sounds typically have a predominantly harmonic structure. The spectral components of a musical tone that are in a harmonic relationship are the predominantly sinusoidally based energy of a spectral frame. Conversely, we can consider spectral energy that is not in a harmonic relationship to be noise based. This is the basis of the sinusoidality coefficient measure presented here.

The harmonic product spectrum (HPS) technique is a method of finding the fundamental frequency energy of a STFT by combining the energy of the harmonics with each other at the location of the fundamental frequency [81, 75]. The harmonic product spectrum is the product of the power spectrum multiplied by successively compressed versions of the same spectrum. These spectra are downsampled along the frequency axis by consecutive integer amounts. The resultant product reinforces the amplitude of the fundamental frequency. A related spectrum, the harmonic sum spectrum (HSS), with similar output, sums these down-sampled spectra instead of multiplying them [2]. The harmonic sum spectrum is given by:

$$HSS(k) = \sum_{r=1}^{K} |X(k_r)|^2 \quad \text{where} \quad k_r = (r)(k). \tag{3.20}$$

Here K is the number of down-sampled spectra typically around six.

Conceptually, our proposed method is the reverse of the HSS. To the normalized power spectra we add each integer multiple up-sampled copy of the spectra. These up-sampled spectra are calculated by interpolating new spectra at each integer multiple of the original spectra. That is:

$$\Gamma_h(k) = \sum_{r=1}^{K} |X(k_r)|^r \quad \text{where} \quad k_r = k/r. \tag{3.21}$$

Spectral peaks in each of these up-sampled spectra are proportionately wider than those in the original power spectra. For example, the spectral peaks in the $r = 2$ spectra are twice as wide as the original. In order to slim these peaks back to an approximation of their original width, the up-sampled spectra are raised to increasing powers and normalized to their highest peak. In order to heuristically calculate this power term, a Hann window is interpolated to twice its length and raised to integer powers. These are normalized to the original height of the Hanning window. It was found that the window raised to the 4th power was slightly wider than the original window and was slightly narrower when raised to the 5th power. So, the $r = 2$ spectra would be taken to the 4th power in order to reduce the width of its peaks to approximate the original width. However, this 4th power was found to produce banks that are too narrow to use in practice. The first up-sampled spectrum is set to the 2nd power in our algorithm, and each subsequent up-sampled length is raised to the next power as can be seen in equation 3.21. This produces approximately equal width peaks for six to twelve up-sampled spectra K. Finally, the slope correction devised for the phase acceleration sinusoidality method is used to reduce the sinusoidality of higher harmonics.

Figure 3.3 shows a normalized power spectrum sinusoidality for a violin tone and the corresponding harmonic sum sinusoidality for this spectra. The harmonics higher than those of the power spectra are visible in the sinusoidality spectrum. This sinusoidality measure is applicable for musical instrument sounds with harmonic content. Indeed, we will see that it performs well on a synthetically generated harmonic sound in section 3.3.

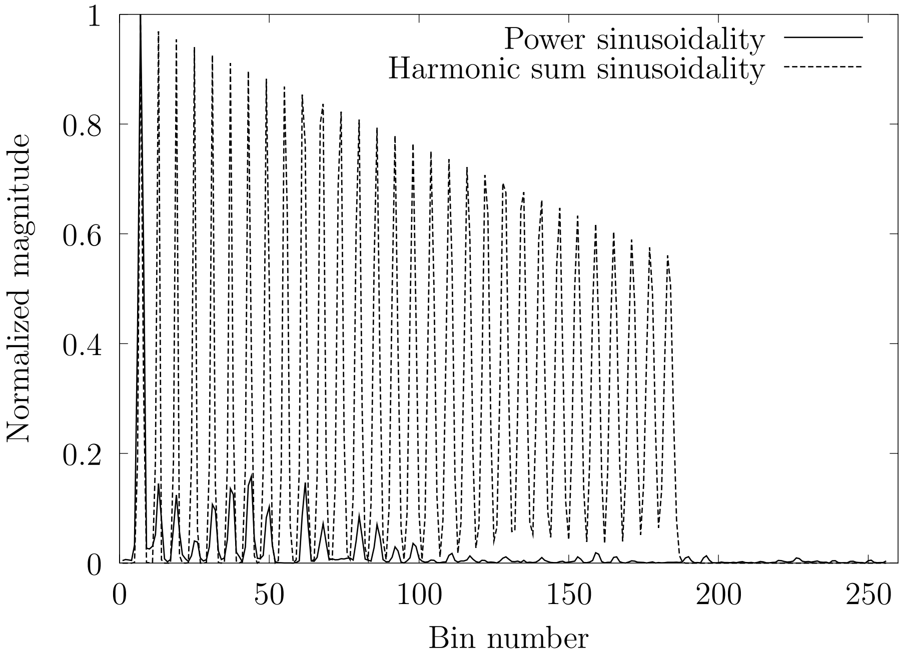

Figure 3.3: Power spectrum sinusoidality and harmonic sum sinusoidality for a a single frame of a violin tone.

3.2.9 Other Sinusoidality Methods

In addition to the methods presented and devised above that have been implemented for comparison and use with our noise synthesis system, other sinusoidality methods have been developed.

Zivanovic, Roebel, and Rodet propose four mutually complementary methods of classifying spectral peaks as sinusoidal and noise based [92]. The characteristics of all bins around a spectral peak are examined to characterize the peak. As these estimates are not computed for each spectral bin, but for each spectral peak of the STFT, they are not directly applicable for our application here [3].

Dubnov proposed a sinusoidality coefficient measure which compares the "bispectra" of the sound with an expected bispectra for a sinusoidal component [19]. This method can be considered an extension of the Lim cross-correlation method into higher order spectra. Dubnov reports results comparable to the Lim method. Notably here, Dubnov suggests that "It seems difficult to find the true definition of voicing [sinusoidality] and much emperical listening work is required in order to determine a good voicing estimator."

Dubnov proposes another sinusoidality coefficient measure as part of a novel sinusoidal analysis/synthesis system [21]. This method compares two parametric spectral estimates calculated not from the STFT, but from LPC filter coefficients. A comparison between these spectra gives an estimate of sinusoidality at any target frequency. As this method produces the sinusodality coefficients in terms of LPC coefficients, it would need to be adapted to produce coefficients in terms of STFT bin.

We have now completed presenting the sinusoidality coefficients analysis techniques. The three power spectrum based methods, the two phase spectrum based methods, the cross-correlation method, a simplified version of the narrow-band variance method, and the harmonic sum method have been implemented. We will now present our method of comparing these measures before discussing

[3] In these methods, the spectrum is partitioned into regions. This partitioning is carried out by the method proposed by Laroche and Dolson [49]. The minimum value between two maxima of the STFT amplitude spectra are used as the limit of each group of spectral bins. Of the four descriptors, the "normalized bandwidth descriptor" appears to be the most salient for sinusoidality detection as it "can be viewed as a measure of the noise energy in the neighborhood of a sinusoidal spectral peak." Although not attempted here, these four methods could be adapted to a bin by bin sinusoidality coefficient method by assigning the found sinusoidality measure to all the bins of each peak. However, this method still would be subject to the artifacts of spectral partitioning and peak picking.

their strengths for our phase vocoder noise synthesis system.

3.3 Sinusoidality Error Analysis

In this section we will devise a technique for quantifying the error present in the various sinusoidality analyses. We test the sinusoidality measures with synthetic sounds using this technique, and suggest which sinusoidality method is appropriate for differing signals based on these results. Our error analysis method compares the known power spectrum of sinusoidal components in a *synthetic* mixed sinusoid and noise signal to the power spectrum generated by multiplying our sinusoidality coefficients by the original combined power spectrum being tested.

Other studies of sinusoidality have quantified results in terms of detecting individual sinusoids in a power spectrum, as their purpose has typically been to improve the frequency detection of sinusoidal components in a sinusoidal modeling system [44, 37]. Our method is designed to give an overall measure of the accuracy of the sinusoidality measures not in terms of *detecting* sinusoids in a noisy signal, but of characterizing the energy of each spectral bin as part sinusoidal and part noisy. This more accurately reflects our use of sinusoidality in a phase vocoder context.

Each test sound is composed of two separate synthetically synthesized components, the sinusoidal component, and the noise component. A PV analysis is carried out on each of these signals creating the sinusoidal spectra $S_n(k)$ and the noise spectra $N_n(k)$. These two complex spectra are combined,

$$S(k) + N(k) = X(k), \tag{3.22}$$

to create our spectral frames, $X(k)$, for sinusoidal analysis. Next, each of the different sinusoidality coefficient measures is applied to on each $X(k)$ to create $\Gamma(k)$. Next, the estimated sinusoidal amplitude spectrum, $|\hat{S}(k)|$, is calculated by scaling the combined amplitude spectrum $|X(k)|$ by the coefficients:

$$|\hat{S}(k)| = \Gamma(k)|X(k)|. \tag{3.23}$$

Now we subtract this estimated sinusoidal amplitude spectrum from the true synthetic sinusoid amplitude spectrum,

$$\epsilon_{missed}(k) = \begin{cases} \Big(|S(k)| - |\hat{S}(k)|\Big) & \text{when } \Big(|S(k)| - |\hat{S}(k)|\Big) > 0, \\ 0 & \text{when } \Big(|S(k)| - |\hat{S}(k)|\Big) < 0, \end{cases} \tag{3.24}$$

where $\epsilon_{missed}(k)$ is the energy error in sinusoidality estimate due to missed energy for each spectral bin k. That is, $\epsilon_{missed}(k)$ is the energy that is in fact sinusoidal, but was not found as such by the sinusoidality measure. The total missed energy error for each frame $T\epsilon_{missed}$ is found by summing $\epsilon_{missed}(k)$ for all k. Finally, the missed energy ϵ_{missed} is scaled by the total original sinusoidal energy in order to express the error independent of the amount of sinusoidal energy present in the signal:

$$\epsilon_{missed} = \frac{T\epsilon_{missed}}{|(S(k)|}. \tag{3.25}$$

The "false" energy error is similarly found by subtracting the true synthetic sinusoid amplitude spectrum from the estimated sinusoidal amplitude spectrum,

$$\epsilon_{false}(k) = \begin{cases} \Big(|\hat{S}(k)| - |S(k)|\Big) & \text{when } \Big(|\hat{S}(k)| - |S(k)|\Big) > 0, \\ 0 & \text{when } \Big(|\hat{S}(k)| - |S(k)|\Big) < 0, \end{cases} \tag{3.26}$$

where $\epsilon_{false}(k)$ is the energy falsely attributed to sinusoidal energy. The total false scaled error ϵ_{false} is found in a similar fashion as above, however the false energy is scaled by the original *noise* energy:

$$\epsilon_{false} = \frac{T\epsilon_{false}}{|(N(k)|}. \tag{3.27}$$

The total error ϵ is found by adding the missed energy, $T\epsilon_{missed}$, and the false energy, $T\epsilon_{false}$, and scaling by the total energy of both spectra.

$$\epsilon = \frac{T\epsilon_{missed} + T\epsilon_{false}}{|(N(k)| + |(S(k)|}. \tag{3.28}$$

Finally, mean values μ and standard deviations σ are calculated for ϵ_{missed}, ϵ_{false}, and ϵ over all spectral frames n.

We have now seen how the synthetic test signals will be employed to evaluate our sinusoidality measures. Unless otherwise noted, in all of our tests, a sampling rate of 44 100 Hz, a FFT frame length of 1024, a frame hop size of 256 (overlap 4), and a Hanning window are used. Table 3.1 lists the names and descriptions of the evaluated sinusoidality measures. Six different synthetic sounds are used to test the sinusoidality measures; (i) a sine tone without noise, (ii) pure noise, (iii)a sine tone with noise, (iv) 32 unevenly spaced sine tones with noise, (v) a harmonic sound with noise, and (vi) a sinusoid with time-varying frequency (chirp signal) with noise. The results of these six tests are presented in table form in appendix A, and four of these tests are discussed here.

Table 3.1: Tested Sinusoidality coefficient measures.

Name		Method
Zeros	Γ_0	Zero for each sinusoidality coefficient.
Ones	Γ_1	One for each sinusoidality coefficient.
Power	Γ_p	Scaled power spectrum.
Power Persistence	Γ_{pp}	Product of three amplitude spectra.
Sigmund	Γ_s	Puckette amplitude difference method.
Charpentier	Γ_c	Irregularity of instantaneous frequency.
Phase Acceleration	Γ_{pa}	Instantaneous frequency difference method.
Phase Acc. × Pow. Per.	Γ_{papp}	Phase Acceleration times Power Persistence.
Cross-Correlation	Γ_{cc}	Griffin and Lim window correlation method.
Variance	Γ_v	Simplified Narrowband Variance method.
Harmonic Sum	Γ_{hs}	Normalized reverse harmonic sum method.

Graph 3.4 shows the sinusoidality error analysis for a single sine tone of 440 Hz with white noise added. The energy of the sine tone is equal to the total energy of the white noise, that is, the sum of the power spectrum of the sinusoidal signal is equal to the sum of the power spectrum of the noise signal. In this case, the sine tone was generated with approximately 20 dB less intensity than the white

noise. As can be seen in graph 3.4, a set of sinusoidality coefficients that are all zero, Γ_0, produces approximately the same total error as the coefficients set to all one, Γ_1. Graph 3.4 shows the lowest total error, 0.153, for the phase based Charpentier sinusoidality measure, Γ_c. The power sinusoidality measure Γ_p shows the second lowest total error.

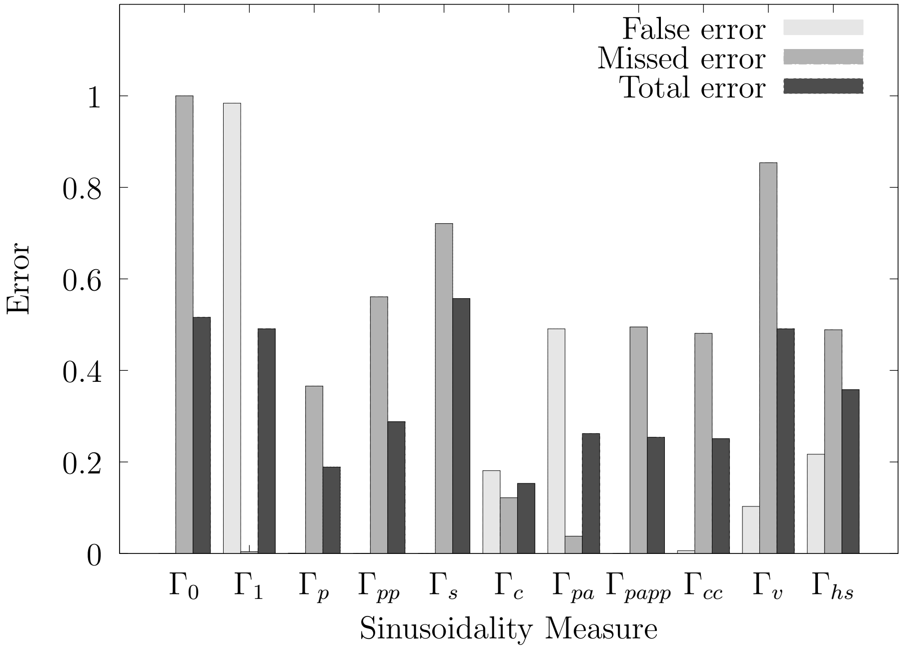

Figure 3.4: Sinusoidality error analysis for a sine tone of 440 Hz with white noise. The energy of the sine tone is equal to the total energy of the white noise.

Graph 3.5 shows the sinusoidality error analysis for 32 equal amplitude sine waves with a random distribution (not harmonic) of frequencies between 40 Hz and 10 000 Hz and white noise. As above, the total energy of the sine tones is equal to the total energy of the white noise. Here the sine tones were each generated with approximately 10 dB less intensity than the white noise. With this signal the power persistence method, Γ_{pp}, and the phase acceleration method, Γ_{pa}, performed approximately as well as the power persistence method creating a greater missed error and the phase acceleration method producing a greater false error.

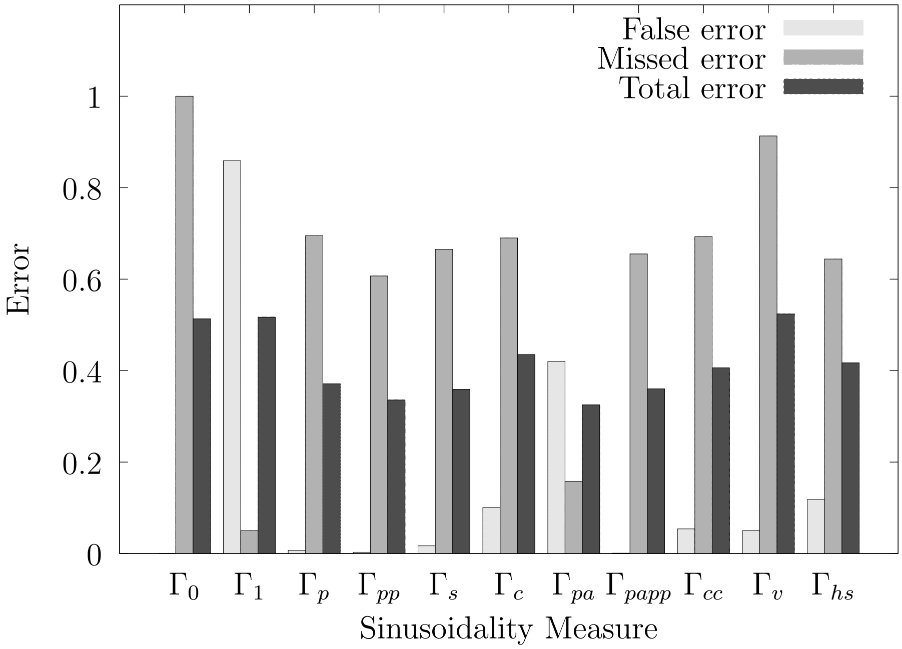

Figure 3.5: Sinusoidality error analysis for 32 equal amplitude sine waves with a random distribution of frequencies between 40 Hz and 10 000 Hz and white noise. The total energy of the sine tones is equal to the total energy of the white noise.

Graph 3.6 shows the sinusoidality error analysis for a harmonic sound with a fundamental frequency of 440 Hz and 31 harmonics each generated with 1 dB less energy than the prior. The energy of the harmonic tone is equal to the total energy of the white noise. The harmonic sum sinusoidality measure, Γ_{hs}, produced the lowest total error, 0.292, for this harmonic sound. The phase acceleration measure, Γ_{pa}, also performed well relative to the other measures.

Graph 3.7 shows the sinusoidality error analysis for a single sine tone with time-varying frequency logarithmically changing from 220 Hz to 880 Hz over one half second. The energy of the time-varying tone is equal to the total energy of the white noise. The sine tone is generated with approximately 19 dB less intensity than the white noise. As can be seen in graph 3.7, the sigmund sinusoidality measure, Γ_s, has the smallest total error of 0.140. This result is notable because

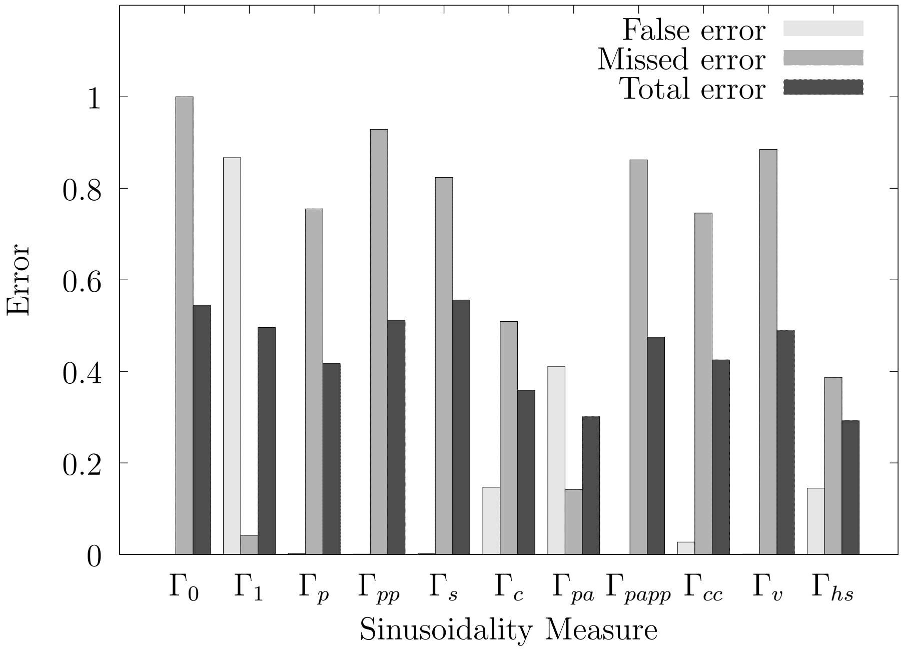

Figure 3.6: Sinusoidality error analysis for harmonic sound with a fundamental frequency of 440 Hz and 31 harmonics each generated with 1 dB less energy than the prior. The energy of the harmonic tone is equal to the total energy of the white noise.

the sigmund method has not produced a particularly low error for other sample sounds. Also, notable is the cross-correlation measure's low total error of 0.218.

Each of these four test signals suggest that different sinusoidality measures are most effective. The Charpentier and power measures performed well with a single sine tone in noise. The phase acceleration and power persistence measures performed well for inharmonic tones in noise. The harmonic sum measure did indeed perform best for a harmonic tone, but the phase acceleration and cross-correlation measures also performed well for this sound. The sigmund measure produced the lowest total error for a time-varying sine tone in noise, followed by the power measure and cross-correlation method.

The poor performance of the variance measure can be attributed to the simplifications to the algorithm in our implementation. Further work with this

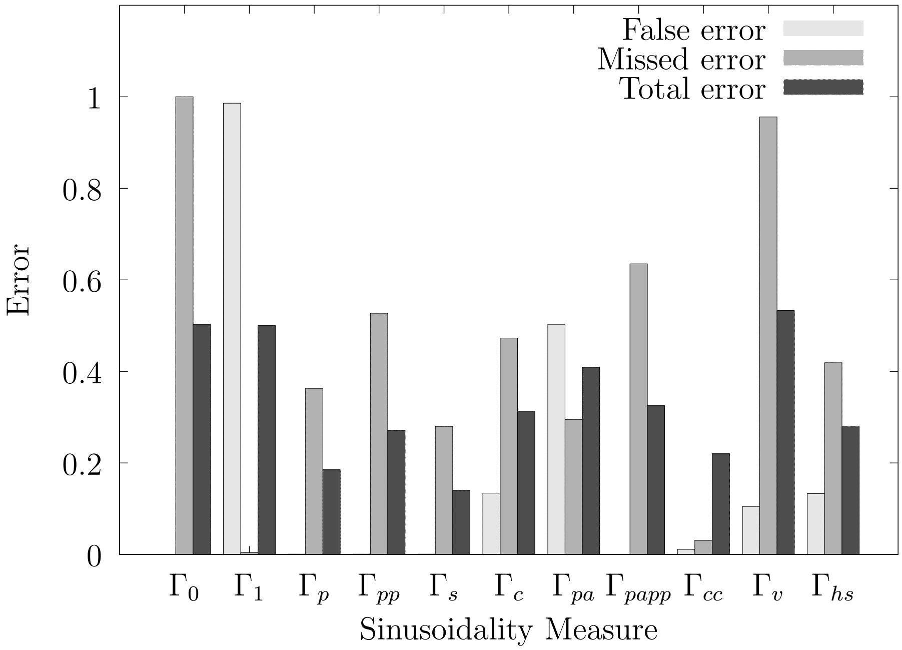

Figure 3.7: Sinusoidality error analysis for a single sine tone with time-varying frequency logarithmically changing from 220 Hz to 880 Hz over one half second. The energy of the time-varying tone is equal to the total energy of the white noise.

measure should not make these simplifications. The combined measure of phase acceleration and power persistence, Γ_{papp} did not create a better measure than one or the other of the measure alone, and it did not produce the least error for any of the test signals.

We can conclude that none of the sinusoidality measures tested is ideal for diverse types of sounds. Even when the sounds are restricted to monophonic tones as discussed above, no one sinusoidality measure presents itself as ideal. In the following section, methods of adding noise to a PV representation will be presented that rely on the sinusoidality coefficients for each PV frame in order to determine the amount of noise to be added to each channel. We employ a number of different measure depending on the type of sound being analyzed. The power, power persistence, phase acceleration, cross-correlation, and harmonic sum methods are all employed to analyze different sounds.

3.3.1 Spectral Flatness Scaling

Before proceeding to a discussion of our noise synthesis technique, a method of biasing the sinusoidality coefficients to counteract the effects of the amplitude scaling is introduced. Theoretically the long-term shape of white noise is flat across all spectral bands. Using the idea that noisy sounds have a flatter spectral shape than the peaky shape of pitched sounds, the spectral flatness measure (SFM) was devised to quantify the overall noisiness of a sound [63, 20]. Unlike the sinusoidality coefficients calculated for each spectral bin in the above methods, the SFM gives only a single value that characterizes the noisiness of a signal. Here we will devise a method of employing the SFM as an overall weighting in order to bias other sinusoidality coefficient measures. This is especially important when working with sinusoidality coefficients that are scaled by the maximum value of the spectrum as discussed in the power sinusoidality section 3.2.1. The SFM is defined as the ratio of the geometric mean of the amplitude spectra divided by the arithmetic mean of the amplitude spectra:

$$SFM = \frac{\left(\prod_{k=0}^{N-1} |X(k)|\right)^{1/N}}{\left(\frac{\sum_{k=0}^{N-1} |X(k)|}{N}\right)}. \tag{3.29}$$

The SFM varies from 1 for white noise close to 0 for a very peaky spectrum. This range is the inverse of the range we are using for our sinusoidality measures. Our method of biasing other sinusoidality measures uses the SFM as a power coefficient to each sinusoidality coefficient. The SFM is multiplied by a positive constant α in order to scale the SFM:

$$\Gamma_{biased}(k) = \left(\Gamma(k)\right)^{(SFM)(\alpha)}, \tag{3.30}$$

Where $\Gamma_{biased}(k)$ are the biased sinusoidality coefficients. We can see that a high SFM (near one) created by noise based signals, decreases the sinusoidality coefficients and a low SFM increases the sinusoidality measures. Before use in our noise synthesis system, the sinusoidality coefficients are biased in this way.

3.4 Sinusoid and Noise Phase Vocoder Synthesis

A method of synthesizing time-stretched sounds from phase vocoder analyses and sinusoidality coefficients for each spectral frame is presented in this section. In this method, the percentage of energy of each spectral bin that is noise-based is multiplied by a spectral domain noise signal. Before describing our sinusoid and noise synthesis algorithm, two existing alternative methods will be considered.

A method of segregating spectral energy separates the spectral channels into two groups, one the "sinusoidal" channels, and the other the "noisal" channels. For each frame, the individual bins are determined as belonging to one of two groups. This is the scheme proposed by Lippe and Settle for segregating bins [86, 51].

A threshold is set for each bin over which the bin is labeled as sinusoidal. Conversely, if the sinusoidal measure is below the threshold, the channel is labeled as noise based energy. Thus,

$$X_s(k) = (\Gamma_n(k) \geq T), \tag{3.31}$$

where $\Gamma_n(k)$ is a frame of sinusoidality measures between 0 and 1 for each channel, T is a stability threshold between 0 and 1, and $X_s(k)$ is a phase vocoder spectral frame with only the stable channels above the threshold not set to zero. Noise is then added to the spectra by a method shown in section 3.4.1. This method is less suited than the method proposed below because of the sharp cutoffs between adjacent bins created from the thresholding of energy.

A second method of adding noise to a PV spectrum involves noise modulation of each spectral bin to a degree determined by the sinusoidality of each bin. This method provides a single or unified spectral representation, as opposed to the dual representations of the other methods in which the sinusoidal spectra and noise spectra are separated for processing and recombined during synthesis. This distinction is analogous to the destinction between how noise and sinusoids are modelled in a dual manner in SMS modelling and how they are represented in a unified representation in the Fitz "bandwidth enhanced" method [27].

In Fitz's system the individual sinusoids of a sinusoidal model of sound are modulated with noise to increase their bandwidth. This process increases the noise level of each sinusoidal component. While the technique is well suited for creating a unified representation of noise in sinusoidal modeling synthesis, it is ill sited for PV based systems in which sinusoidal energy is necessarily spread across several bins of the spectrum. For example, a single sinusoid in a PV spectra will occupy at least three consecutive bins. Increasing the noise "bandwidth" of each of these bins would put energy in adjoining channels.

3.4.1 Dual Model Synthesis

Our method of segregating sinusoidal energy from noise based energy in a spectral representation divides the energy of each spectral bin into two parts corresponding to the sinusoidal energy and the noise-based energy. These separated spectra are processed separately and recombined during synthesis. Figure 3.8 shows our complete analysis synthesis system for time-stretching noise and pitched sounds. This process is discussed below.

The process starts with the STFT analysis data as a series of amplitude and phase spectra frames along with corresponding sinusoidality coefficients for each spectral bin. For each frame n, a new amplitude spectrum is created by using a sinusoidality coefficient spectra $\Gamma_n(k)$ to scale the amount of sinusoidal energy present in each bin. For each bin k of each spectral frame n,

$$|S_n(k)| = (\Gamma_n(k))\,(|X_n(k)|)\,, \tag{3.32}$$

where $|S_n(k)|$ is a new magnitude spectrum, here called "sinusoidal magnitude

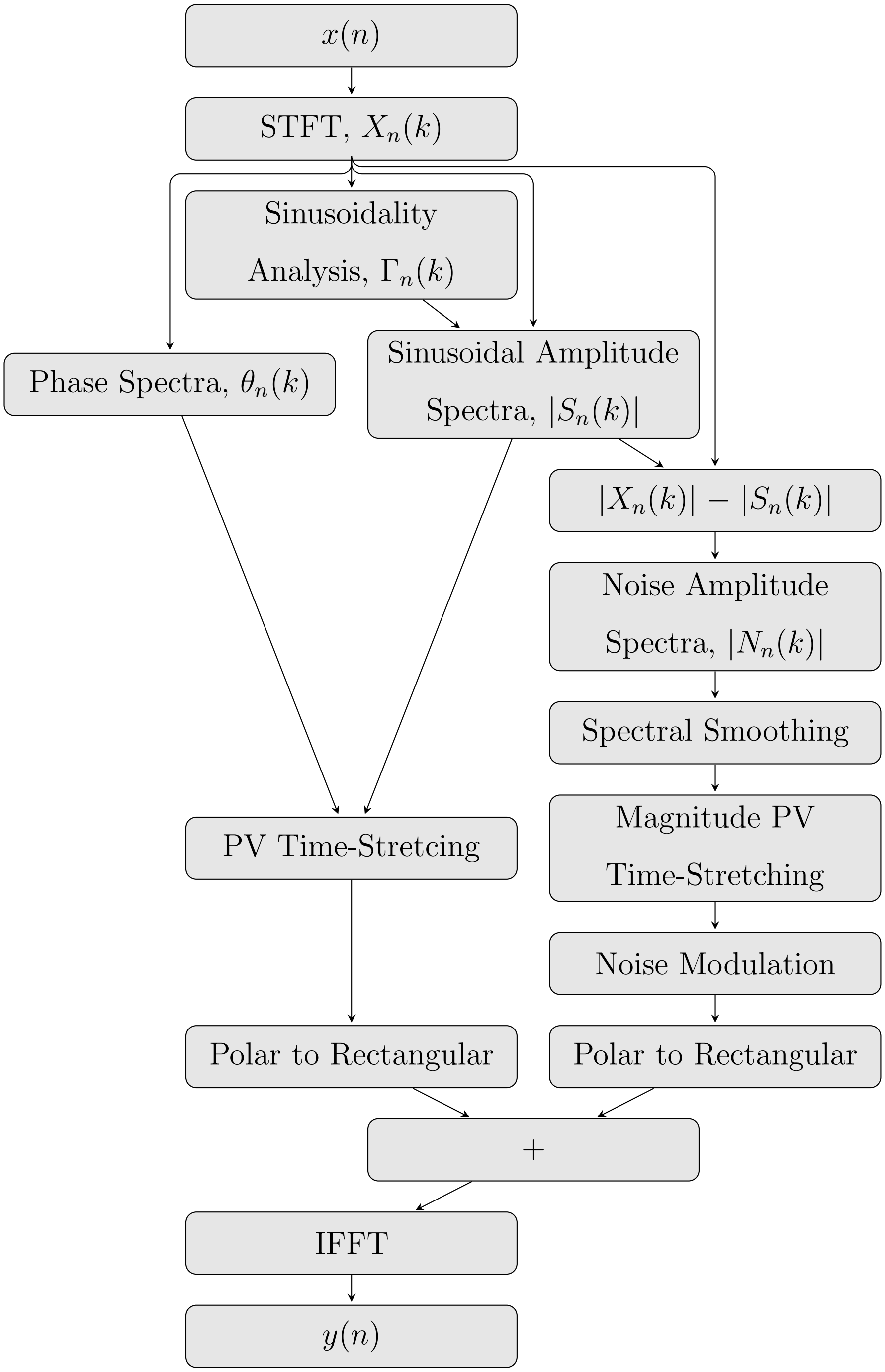
$x(n)$
STFT, $X_n(k)$
Sinusoidality Analysis, $\Gamma_n(k)$
Phase Spectra, $\theta_n(k)$
Sinusoidal Amplitude Spectra, $|S_n(k)|$
$|X_n(k)| - |S_n(k)|$
Noise Amplitude Spectra, $|N_n(k)|$
Spectral Smoothing
PV Time-Stretcing
Magnitude PV Time-Stretching
Noise Modulation
Polar to Rectangular
Polar to Rectangular
$+$
IFFT
$y(n)$

Figure 3.8: Sinusoidality analysis and noise synthesis system.

spectrum", in which each bin's amplitude is scaled by the corresponding sinusoidality of that bin.

Next, the corresponding "noise magnitude spectrum," $|N_n(k)|$, is calculated by subtracting the sinusoidal magnitude spectrum from the unaltered magnitude spectrum,

$$|N_n(k)| = |X_n(k)| - |S_n(k)|. \tag{3.33}$$

As can be seen here, the original magnitude spectrum can be recreated by adding the noise magnitude spectrum and the sinusoidal magnitude spectrum.

The noise magnitude spectrum is subject to a large variance in bin amplitude values between spectral frames. While this behavior accurately reflects the character of noise energy, it produces unwanted sonic artifacts when these random fluctuations are subject to PV time-stretching. For this reason the noise magnitude spectra are filtered in order to smooth each channel in time and frequency. The time frame smoothing is achieved by recursively averaging past frames of $|N_n(k)|$. Martin suggests a spectral noise smoothing filter for use on ambient stationary noise signals [57],

$$|NS_n(k)| = \alpha|NS_{(n-1)}(k)| + (1-\alpha)|N_n(k)|, \tag{3.34}$$

where $|NS_n(k)|$ is the smoothed noise magnitude spectrum, and the smoothing constant α is typically set, according to Martin, between 0.9 and 0.95. As the dynamic character of musical sounds is typically present in the noisy aspects of sound, we use a smoothing constant of 0.4, which is much smaller than the Martin suggestion.

In addition to this temporal smoothing, we smooth the noise magnitude spectrum across bins. This is simply achieved by using a running average of β bins, where $8 < \beta > 12$,

$$|NS_n(k)| = \frac{\sum_{l=-\beta/2}^{\beta/2} |N_n(k+l)|}{\beta}. \tag{3.35}$$

Both of these techniques are used to smooth our noise magnitude spectra.

From here, our smoothed noise magnitude spectra $|NS_n(k)|$ are time-stretched separately from the sinusoidal magnitude spectra above. Since they are only amplitude spectra without corresponding phase or instantaneous frequency spectra, the amplitude spectra are simply interpolated between spectral frames. These new noise magnitude spectra are then each multiplied by a different STFT analysis of white noise,

$$N_n(k) = |NS_n(k)|P_n(k), \tag{3.36}$$

where $P_n(k)$ is a new STFT of white noise for each n of the time stretched sound. Each new noise spectral frame is added to its corresponding sinusoidal spectral frame,

$$Y_n(k) = S_n(k) + N_n(k), \tag{3.37}$$

where $S_n(k)$ is the complex sinusoidal spectral frame produced by a traditional phase vocoder time stretch of the sinusoidal magnitude spectrum $|S_n(k)|$ with the original phase and derived instantaneous frequency values. $Y_n(k)$ is the resultant spectral frame that is inverted to the time domain by the IFFT and appropriate windowing. The final new sound $y(n)$ is shown at the bottom of figure 3.8.

Several differing sounds both musical and environmental were time-stretched with the new sinusoidality analysis and noise synthesis method. In each case, the original unaltered sound is followed by a traditional phase vocoder time-stretching with 8 times the normal length. Then, the new time stretched sound with the noise characteristics preserved is listed. Table 3.2 lists the sample sounds that can be found on the website (crca.ucsd.edu/~tapel/fppv/).

3.4.2 Conclusion

In each case, it can be heard that the generation of noise as part of the phase vocoder time-stretching creates a sound that more closely resembles the noise characteristics of the unaltered sound as compared to the traditional phase

Table 3.2: Example noise retention phase vocoder time-stretch sounds.

Track	Sound	Description
1	Sine then Noise	Original version of sound
2		Time-scale slowed by 32 with traditional PV
3		Time-scale slowed by 32 with noise maintained
4	2930 Hz and Clicks	Original version of sound
5		Time-scale slowed by 8 with traditional PV
6		Time-scale slowed by 8 with noise maintained
7	Brush Roll	Original version of sound
8		Time-scale slowed by 8 with traditional PV
9		Time-scale slowed by 8 with noise maintained
10	Breathing	Original version of sound
11		Time-scale slowed by 8 with traditional PV
12		Time-scale slowed by 8 with noise maintained
13	Noise and High Tone	Original version of sound
14		Time-scale slowed by 8 with traditional PV
15		Time-scale slowed by 8 with noise maintained
16	Sine fade into Noise	Original version of sound
17		Time-scale slowed by 8 with traditional PV
18		Time-scale slowed by 8 with noise maintained
19	Pencil Sharpening	Original version of sound
20		Time-scale slowed by 8 with traditional PV
21		Time-scale slowed by 8 with noise maintained
22	Wind	Original version of sound
23		Time-scale slowed by 8 with traditional PV
24		Time-scale slowed by 8 with noise maintained
25	Rain and Thunder	Original version of sound
26		Time-scale slowed by 8 with traditional PV
27		Time-scale slowed by 8 with noise maintained
28	Apple Bite	Original version of sound
29		Time-scale slowed by 8 with traditional PV
30		Time-scale slowed by 8 with noise maintained
31	Motorcycle	original version of sound
32		time-scale slowed by 8 with traditional PV
33		time-scale slowed by 8 with noise maintained
34		time-scale slowed by 8 with sine part only
45		time-scale slowed by 8 with noise part only

vocoder time-stretching method. Occasionally, the two parts of the time-stretched sound, sinusoidal and noise based, do not fuse perceptually as they do in the original sound. This is no doubt attributable to the dual nature of our synthesis system. Future work could consist of combining the two modes of synthesis into a unified method. However, it is my intention to implement the current system in widely used computer music languages before any further research.

Chapter 4

Vibrato and Tremolo Preservation during Phase Vocoder Time-Stretching

> *Concerning the vibrato modification of a recorded voice, there are manifold difficulties.*
>
> Daniel Arfib and Nathalie Delpart [5]

4.1 Sub-Audio Modulations

We have seen that the noise characteristics of a sound are altered during phase vocoder (PV) time-stretching. But PV time-stretching also affects the rate of vibrato and tremolo in musical sounds by slowing them proportionally to the amount of time-stretching. This behavior is for many applications an acceptable or preferred outcome. However, in many musical instrument performances, the particular vibrato and tremolo rates are important acoustic features that are critical to the musical perception of the performances. These sub-audio rate modulations, both in amplitude and frequency, are an important part of musical instrument identification, a performer's style, and musical expression. By preserving the original rates of vibrato and tremolo after a sound has been time-scaled, these important musical perception cues may be retained.

In this chapter we will review the current methods of vibrato and tremolo analysis in a spectral representation of sound, introduce the relevant background in second order Fourier transforms and long duration Fourier transforms, and develop a novel method of maintaining the original vibrato and tremolo rates of a recorded sound using second order analysis of the phase vocoder representation of sound.

4.2 Existing Methods

Carl Seashore performed extensive study of musical vibrato in the 1930's, in which he defined and characterized vibrato and tremolo of western musical instruments and voice [83]. He characterizes vibrato as a 4 to 8 Hz pulsation of pitch, loudness, and timbre, in which pitch change is approximately a semitone, and in which pitch and amplitude modulation rate are approximately constant. Seashore classifies both tremolo and vibrato as aspects of a single phenomenon he calls vibrato. Here we will separate vibrato as only a frequency modulation phenomenon and tremolo as an amplitude modulation phenomenon. When we wish to discuss any periodic or quasi-periodic low frequency modulation, whether amplitude, frequency or a combination or them, we will call them sub-audio modulations. Further discussion of the vibrato and tremolo characteristics of individual musical instruments can be found in Timmers and Dresian [88].

Research into sub-audio modulation of musical sounds typically involves automated detection of the existence of vibrato and or tremolo in a musical signal and estimation of the rate and extent of the modulation. Rossignol et al. suggest five methods of detecting and estimating the vibrato of a musical signal [79]. Two of the methods operate on the DFT spectrum of the signal: the first, comparing its shape to a synthetic spectrum, and the second comparing the shape to the shape at a different time. Three other methods operate on the fundamental frequency track, f_0, extracted by various methods, typically from the analytic signal generated by the Hilbert transform method. This f_0 signal is analyzed for its vibrato by calculating the DFT of the f_0 track, by extracting its fundamental frequency, or by computing the distance between local maxima in the f_0 trajectory. This first of these three methods, was originally developed by Herrera and Bonada [40].

Prior efforts to remove or modify vibrato from a recorded sound are generally based on sinusoidal analysis synthesis systems, that is, methods based on the MQ or SMS analysis synthesis systems [74, 84]. However, one method based on cepstral analysis will be presented here before sinusoidal based methods are reviewed.

4.2.1 Arfib and Delprat Cepstral Method

Daniel Arfib and Nathalie Delprat suggest a method of vibrato alteration of a sampled sound based on DFT, cepstral, and Hilbert pitch tracking techniques [4, 5]. Their method decomposes the pitch curve calculated in the cepstral domain into a mean frequency and an oscillatory frequency. The oscillatory frequency is modified before recombination with the mean frequency and transformation out of the cepstral domain.

Specifically, their technique is as follows. Starting with a phase vocoder representation of sound, each amplitude spectrum is converted to the cepstral domian. Each cepstral domain spectrum is separated into two parts (cepstral "liftering") consisting of the low coefficients which typically contain the spectral envelope spectrum and the high coefficients which contain the so called excitation spectrum. The amplitude of the excitation spectrum is converted back to the spectral domain, and the low spectral peaks are used to track the fundamental frequency and the modulation rate or vibrato. This time varying fundamental frequency is subtracted from the modulation rate to produce a vibrato rate signal. The frequency and amplitude of this signal is in turn calculated with a Hilbert transform based analytic signal method. These rates are modified before recombination with the mean frequency and transformation obtained from the cepstral domain. Finally, the inverse cepstrum is combined with the original phase estimates of the phase vocoder to create a new sound. They report successful alteration of vibrato rates with errors produced from the pitch detection method and extraction of the vibrato.

4.2.2 Maher and Beauchamp Vibrato Method

Maher and Beauchamp created a method of removing vibrato from vocal sound recordings based on the MQ sound representation [52]. In their system, the frequency of sinusoidal analysis tracks are smoothed to remove the vibrato. The system uses frequency modulated wavetables to synthesize tones with vibrato, but does not allow for modification of vibrato rates or re-introduction of vibrato to time stretched sounds. However, it does appear to be the first time that vibrato had been modified independently of other musical parameters in a sampled sound.

4.2.3 Marchand and Raspaud Order-2 Modeling

A recent system proposed by Sylvian Marchand and Martin Raspaud creates a second order sinusoidal system [56] that is conceptually similar to Schumacher's 2DFT analysis introduced below. Their method analyzes the sinusoidal tracks of a sinusoidal modelling system using sinusoidal analysis, i.e. each amplitude and frequency track of a sinusoidal analysis is in turn analyzed as a series of sinusoidal tracks. This method is used in a time-stretching algorithm which preserves the vibrato and tremolo of the recorded sound. As the modulations of each amplitude and frequency spectral track follow the sub-audio variations in each track, a sinusoidal analysis of each of these partials will capture this sub-audio modulation.

Their method uses the standard method of time-stretching with a sinusoidal model, (i.e., interpolating between the amplitude and frequency of spectral track peaks) performed in this new second order domain. As these tracks do not represent the amplitude and frequency of spectral energy, but rather, the amplitude and frequency of individual amplitude and frequency tracks. As will be seen, the idea of a second order spectral representation will be used in our PV based system presented below.

4.3 Related DFT Techniques

Before our sub-audio retention method is presented, two signal analysis techniques that are incorporated into the method will be reviewed. In our system, the 2DFT second order analysis technique of Schumacher will be expanded to an analysis synthesis system by using the long-term discrete fourier transform (LTDFT) of Hammer and Sundt in the second order domain.

4.3.1 Schumacher 2DFT Analysis

Robert Schumacher defined the 2DFT method of analyzing low frequency and sub-audio periodicities [82]. His method performs a single DFT along each of the time channels of the spectrogram of a sound. The amplitude spectrum of this second order DFT shows the spectrum of the low frequency and sub-audio periodicity of each channel of the input sound. Averaging the second order DFTs across all of the first order spectral channels produces a spectrum in which low frequency and sub-audio amplitude modulation of the original sound can be identified as spectral peaks. Schumacher used this method to analyze the aperiodicities in periodic waveforms. Schumacher's 2DFT analysis method will serve as part of the analysis, transformation, synthesis system developed here to maintain sub-audio modulation rates. As will be seen, the 2DFT analysis allows us to model sub-audio amplitude modulations as spectral peaks that are maintained during phase vocoder time-stretching operations.

4.3.2 Long-Term Discrete Fourier Transform

The idea that a single Fourier Transform can be performed over long durations of time-variying sound in order to create novel transformations was formalized by Øyvind Hammer and Henrik Sundt [35]. Their research was inspired by the composition "Z" by Paul Pignon in which a single DFT was taken from a guitar and saxophone improvisation, and manipulations to the DFT coefficients were performed before inverting the DFT. Hammer and Sundt call this transformation the long-term discrete fourier transform or LTDFT and suggest several manipulations that are possible in this domain. Among these, they suggest (i) remapping amplitude spectra values to differing bins in order to create time dispersion of frequency bands or frequency sweeps and (ii) multiplying the phase spectra by a constant in order to move frequency bands in time. Most all of the manipulations in the LTDFT domain create sounds which disrupt the integrity of the original sonic events. That is, sounds become unrecognizable because their frequency components have been moved to radically different locations in the output sound. Unlike the phase vocoder representation of sound as amplitude and frequency spectra which localize sonic events in time, the LTDFT does not encode time in a perceptually relevant manner. This makes it it particularly difficult to predict the outcome of transformations performed in the LTDFT domain.[1]

We will see in our sub-audio modulation retention method that applying the idea of the LTDFT in a higher order context helps us to model the globally relivant vibrato and tremolo rates. In general, a technique that produces abstract sounds in the first order LTDFT domain produces a useful model of sub-audio

[1] My experiments performing transformations in the LTDFT domain confirm this difficulty. Nevertheless, I have found that setting the LTDFT phase spectra to a constant produces interesting abstract sonic textures and that randomizing the phase creates a constant droning sound that has all the frequency content of the original sound. I use both of these techniques in my own compositional work. This technique will be discussed further in chapater 5.

modulation in a higher-order domain.

4.4 Second-order PV based Sub-Audio Modulation System

We are now in a position to introduce our sub-audio modulation analysis/synthesis method based on the 2DFT analysis and LTDFT analysis/synthesis methods.

The phase vocoder typically employes a frame length of 512 or 1024 samples. Using a sampling rate of 44 100 Hz and a frame length of 1024 samples, the lowest frequency, f, that can successfully be analyzed is found by

$$f = \frac{SR}{N} = \frac{44\,100}{1024} = 43.066\,\text{Hz}. \tag{4.1}$$

Sub-audio frequencies are not captured explicitly as energy in low-frequency bins but as a pattern of changes across multiple spectral frames. The frame length required to capture modulations as low as 4 Hz, is

$$N = \frac{SR}{f} = \frac{44\,100}{4} = 11\,025\,\text{samples}. \tag{4.2}$$

A phase vocoder analysis with such a long frame length will smear note onsets and transients under sound transformation such as time-stretching. However, a phase vocoder with this frame length does model sub-audio modulations as spectral energy, and consequently PV time-stretching extends these frequencies at their original rates. If the sound to be analyzed is legato or does not have sharp transients this method may be appropriate.

In order to maintain the frame rate of a traditional phase vocoder and analyze sufficiently long periods of sound to characterize sub-audio modulation, we analyze *each spectral channel* of a PV analysis with a single long DFT (LTDFT). In each of these “2LTDFT” spectra the sub-audio modulation can typically be seen as a prominent peak in the 4 to 8 Hz range. As will be seen, this method allows us to modify or remove the sub-audio modulation by manipulating this 2LTDFT peak. Our method operates by (i) removing the energy in 2LTDFT bins, (ii) PV time-stretching the unmodulated result, (iii) performing another 2LTDFT analysis of the lengthened sound, (iv) and imposing the shape of the original sub-audio energy on the 2LTDFT spectra. The analysis/synthesis system is now presented in detail along with results from differing sound types.

4.4.1 Removal of Sub-Audio Modulation

This section will present our method of sub-audio modulation removal as the first step in a three-part process of removing the sub-audio modulation, time-stretching the sound with the modulation removed, and adding the modulation back into this time-stretched version of the sound. As noted above, the removal and re-introduction of sub-audio modulation is performed in a second order spectral domain. This domain allows us to model and manipulate the sub-audio modulation of an audio signal. Fig 4.1 shows all the components of the second order analysis/synthesis system. The first step in our removal algorithm is a phase vocoder analysis, as shown in chapter 2. Windowed and overlapped FFT frames are calculated, the complex rectangular spectra are converted to amplitude and phase spectra, and the instantaneous frequency is calculated for each spectral bin with the phase difference method. As discussed in chapter 2, we call the amplitude spectrum $|X_n(k)|$, and the instantaneous frequency spectrum $\Delta\theta_n(k)$ for time frame n and bin number k.

Typically, higher order spectral analysis considers the spectral frame as the indivisible object of further analysis. For example, cepstral analysis performs a second DFT on the Spectral frame after calculating the log magnitude of the spectrum [64]. Instead we will consider the *time-evolution* of each spectral channel as our indivisable objects for further analysis, not the spectral frame. In order to

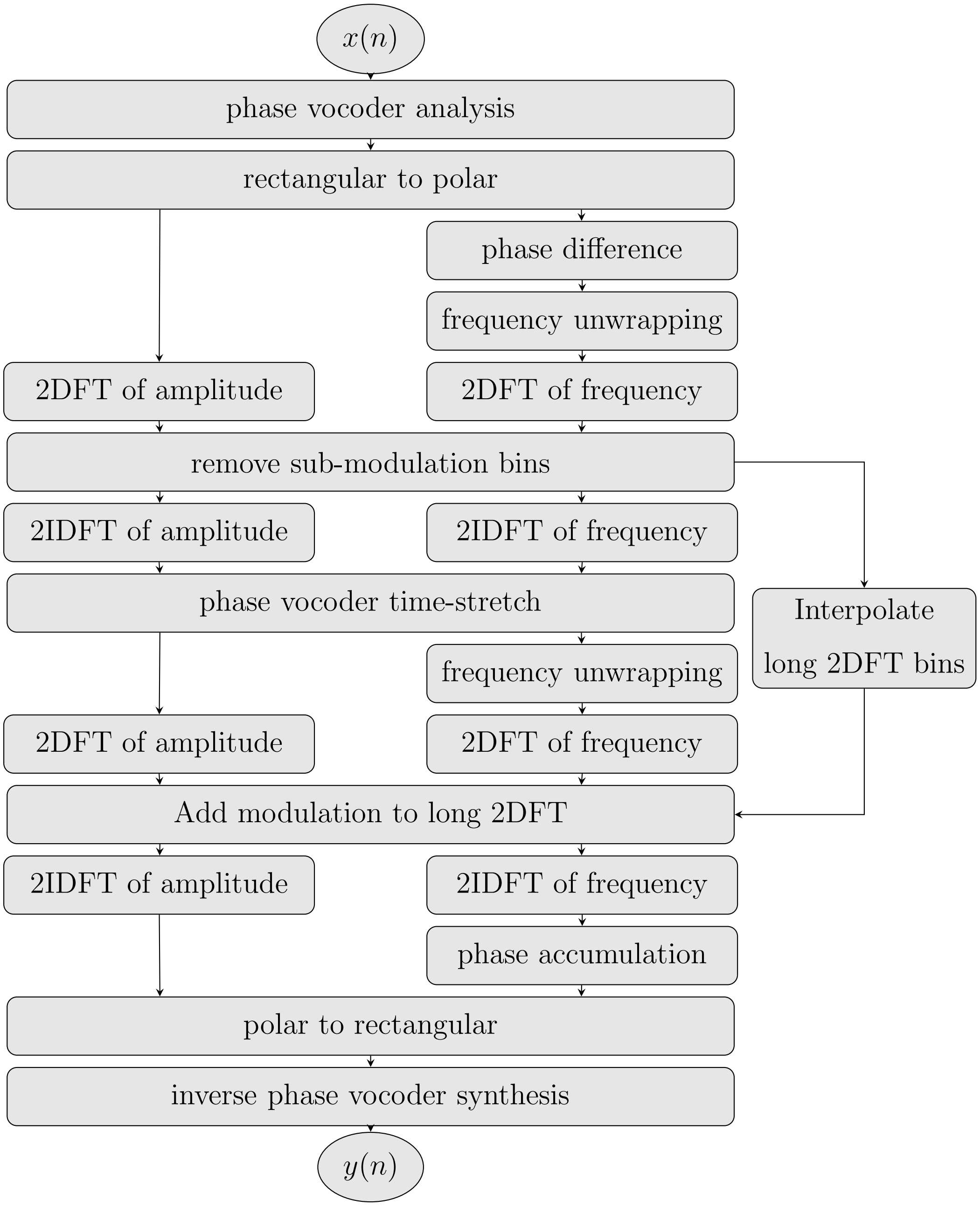

Figure 4.1: sub-audio analysis/synthesis method.

emphasize that the frame number n is the independent variable, the k amplitude channels are denominated $|X_k(n)|$, and the instantaneous frequency channels are notated $\Delta\theta_k(n)$. We now analyze each of these amplitude and frequency channels using a single FFT for each channel. Each second order spectrum of the amplitude channel is defined as,

$$\hat{X}_a(k,m) = \text{DFT}\big(|X_k(n)|\big), \tag{4.3}$$

where $\hat{X}_a(k,n)$ is the complex valued spectrum of each amplitude channel k and frequency bins m, and the hat symbol is borrowed from cepstral processing to denote a higher order spectra. Hereafter, we will call these spectra the "2LTDFT amplitude spectra" following Schumacher's nomenclature for defining the second order spectrum of amplitude channels. It is worth noting here that these 2LTDFT amplitude spectra are complex valued and can be converted to their corresponding amplitude, $|\hat{X}_a(k,m)|$, and phase, $\Delta\hat{\theta}_a(k,m)$, spectra if necessary. It is also worth noting that it is relatively easy to confuse the amplitude spectra of a frequency channel with the phase spectra of an amplitude channel or other combination of first and second order spectra.

As the arctangent function necessary to calculate phase from the complex spectrum produces principal values that are bounded by π and$-\pi$, the instantaneous frequency calculations are also bounded to this range. In order to calculate the corresponding second-order spectra of the frequency channels, the instantaneous frequencies must be transformed into a continuous function by unwrapping since discontinuities in the instantaneous frequency channels would be interpreted as false periodicities or noise in our 2LTDFT frequency analysis. After frequency unwrapping, each second order spectrum of the frequency channel is defined as,

$$\hat{X}_f(k,m) = \text{DFT}\big(\Delta\theta_f(k,n)\big), \tag{4.4}$$

where $\hat{X}_f(k,m)$ is similarly the complex valued spectrum of each frequency channel, defined here as the 2LTDFT frequency spectrum.

The center frequency of each of the 2LTDFT bins can be calculated as

$$f_k = \frac{SR}{(H)(N)} \tag{4.5}$$

where SR is the sampling rate of the sampled sound, H is the hop overlap of the first-order window and N is the number of of frames in the first-order spectra.

As an example, Figure 4.2 shows a time-domain signal of a saxophone note with prominent tremolo and vibrato. Visible in the 3 second sample is the slow amplitude variation at approximately 4 Hz. Figure 4.3 shows one amplitude channel (channel 10) after the phase vocoder analysis of this same saxophone sound, and Figure 4.4 shows the corresponding unwrapped instantaneous frequency channel. These two figures clearly show the sub-audio modulation that will be modeled by the 2LTDFT analysis.

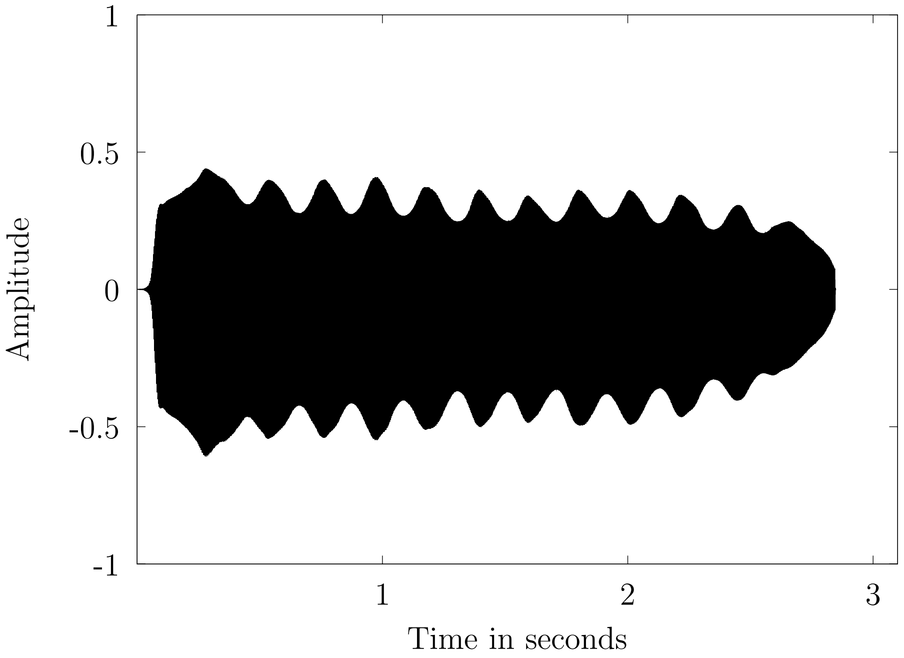

Figure 4.2: Saxophone sound showing amplitude modulation.

Figure 4.5 shows the 2LTDFT amplitude spectrum of the amplitude channel. As can be seen in these figures a prominent spectral peak at approximately bin number 16 is visible. The center frequency for this bin is approximately 4 Hz. This spectral shape consisting of a large DC component and a prominent spectral peak around the vibrato and tremolo rate is characteristic of many musical instruments which feature this low frequency modulation.

We can now remove this low frequency modulation by “spectrally filtering”

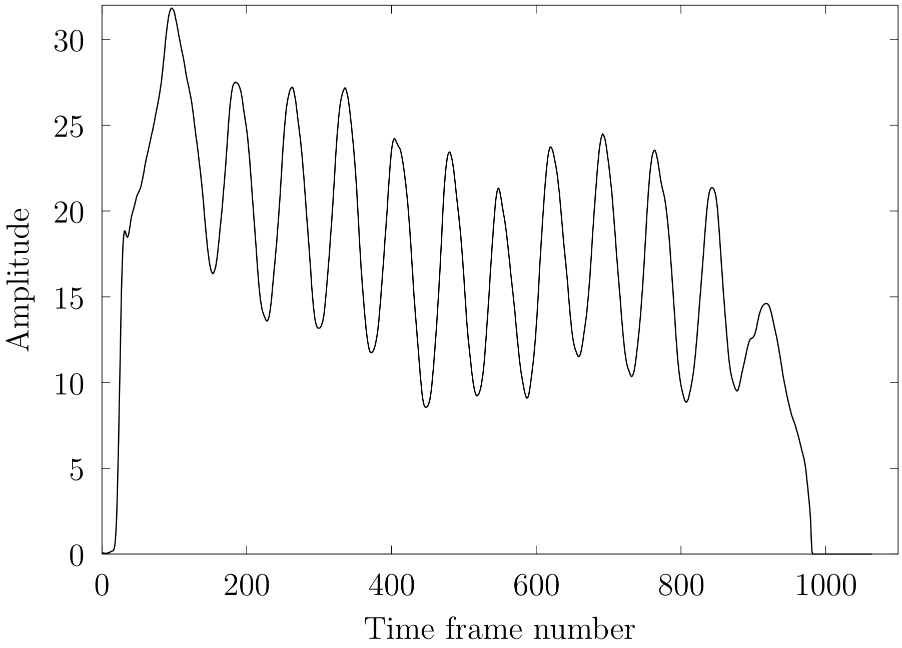

Figure 4.3: A single amplitude channel (channel 10) of the phase vocoder analysis of saxophone sound showing amplitude modulation.

each of the 2LTDFT frequency and amplitude spectra. Spectrally filtering simply means changing the amplitude of the complex spectral bins. Our method consists of multiplying each 2LTDFT spectrum by an inverted Hamming window centered at the 2LTDFT bin with the maximum sub-audio frequency component, i.e. the peak of the sub-audio modulation in the 2DFT spectra. Figure 4.6 shows the same amplitude of an 2LTDFT amplitude spectra after an inverted Hamming window centered on the modulation peak is multiplied by the spectrum. As can be seen, the energy in the peak bins is reduced to zero. This method smoothes the transitions to the adjacent bins of the 2LTDFT spectra. The original unmodified 2LTDFT bins are stored separately for use in re-introducing modulation after time-stretching.

The set of 2LTDFT amplitude spectra and 2LTDFT frequency spectra are then each converted back to the first-order spectral representation by the inverse FFT. Again, we note that this is a single global operation over the entire length

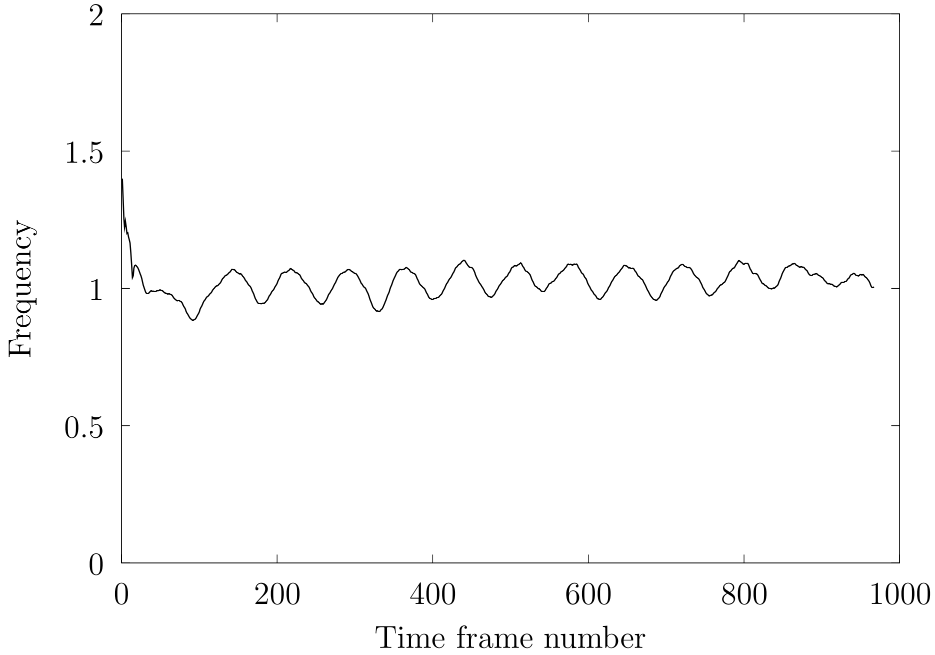

Figure 4.4: A single frequency channel (channel 10) of the phase vocoder analysis of saxophone sound showing frequency modulation.

of the spectral channels. No windowing or overlapping is performed to transform to or from the 2LTDFT domain. These new amplitude and frequency channels are now in the standard phase vocoder representation of amplitude and frequency channels which can be transformed back to a time-domain sound by means of phase accumulation, inverse FFT, windowing, and overlapping. When these steps are taken, the resultant sound has the sub-audio tremolo and vibrato removed. Here, we will not convert back to a time-domain sound, but rather continue working with the new sound in PV form.

4.4.2 De-modulated Lengthening

Next, a traditional phase vocoder time-stretch of this now "de-modulated" sound is performed. As the sound is already in a first-order phase vocoder representation, the initial steps of spectral analysis and instantaneous frequency calculation

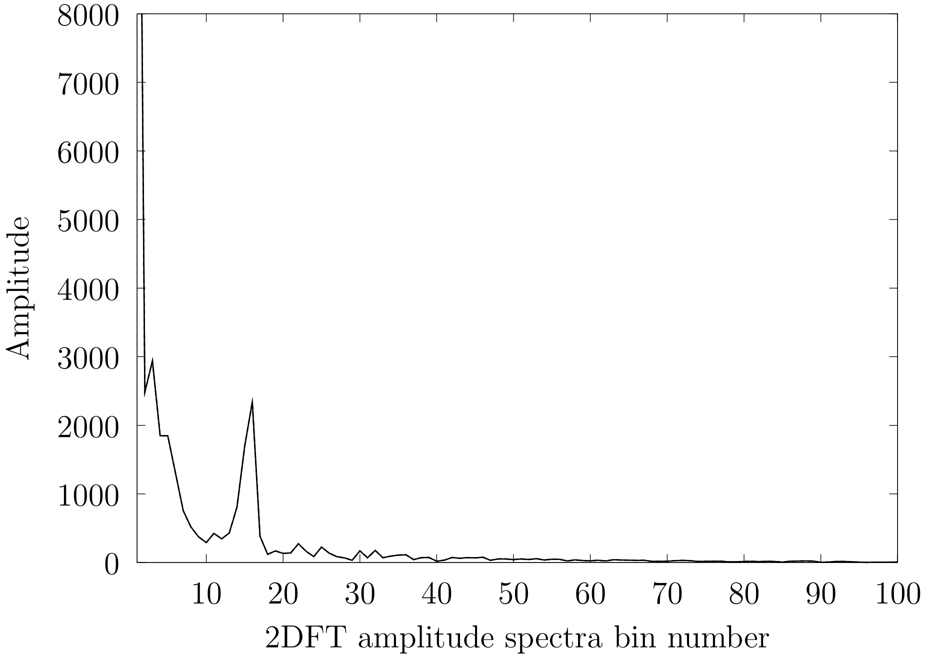

Figure 4.5: Amplitude of 2DFT amplitude spectrum of a single amplitude channel.

are already preformed. All that is necessary is to calculate the new phase frames by phase advancement and calculate the amplitude frames by interpolation as shown in chapter 2. Stretch factors or 2, 4, and 8 are used for computational efficiency in our examples.

4.4.3 Re-introducing Modulation

The final step of our method is to re-introduce sub-audio modulation at the original rate to our de-modulated and lengthened phase vocoder channels $Y_k(n)$. As above, each amplitude and frequency channel is transformed into the second-order 2LTDFT format by taking the FFT of each channel across time. Again, the instantaneous frequency values are unwrapped in time to make a continuous function before the FFT is performed. These new 2LTDFT spectra will be longer in proportion to the amount of time-stretching. Consequently, the original 2LTDFT sub-modulation bins cannot be directly combined into our new 2LTDFT spectra.

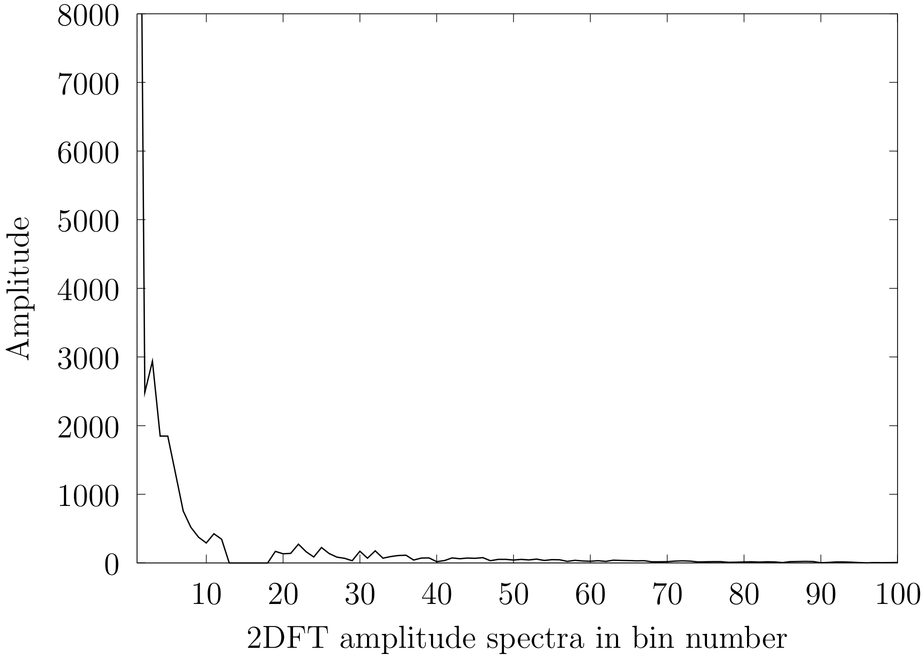

Figure 4.6: Amplitude of 2LTDFT amplitude spectrum of a single amplitude channel after removal of modulation.

In order to add the original modulation rate to the 2LTDFT spectra, the original 2LTDFT bins representing the sub-audio modulation are interpolated to fit the corresponding spectral length in the new longer 2LTDFT spectra. This interpolated peak is substituted for the existing data in the new range of bins. Figure 4.7 shows these bins interpolated over the new longer 2LTDFT length and substituted for the original 2LTDFT bins in the modulation range. After substituting the new interpolated 2LTDFT values, the inverse FFT can be performed on each spectral channel in order to return to the first-order phase vocoder representation of sound. Phase accumulation followed by the standard phase vocoder inversion is employed to generate the new sound. This sound exhibits the long term development, pitch and timbre of the orignal sound slowed by the new factor but with the original tremolo and vibrato rates maintained. Figure 4.8 a shows the final 3 second saxophone sound stretched to 6 seconds, where the original 4 Hz amplitude modulation rate is evident from a visual inspection.

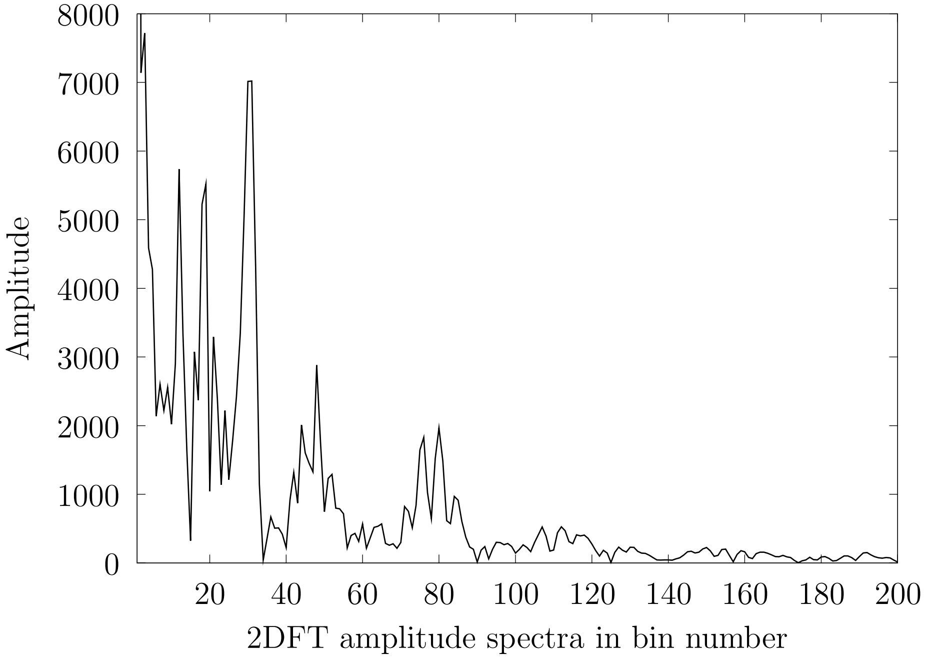

Figure 4.7: Amplitude of 2DFT amplitude spectrum of a single amplitude channel with addition of interpolated modulation bins.

4.5 Algorithm Examples

The algorithm was tested on several synthetic and musical signals with differing length, tremolo, and vibrato characteristics. Five example sounds are shown here which illustrate some of the features of the second-order analysis synthesis system. The demonstrations allow auditory evaluation of the algorithm to evaluate the efficacy and qualities of the method.

Unless otherwise mentioned all sound samples are monophonic and were sampled or generated at 44 100 Hz. A Hamming window is used for the first-order FFT with a window size of 1024 and a first-order FFT overlap factor between windows of 4. Table 4.1 lists the sample sounds that can be found on the website (crca.ucsd.edu/~tapel/fppv/).

Sound example 1 in table 4.1 is a 3 second synthetically generated amplitude modulated sine tone which simulates a typical musical tremolo rate. It is generated

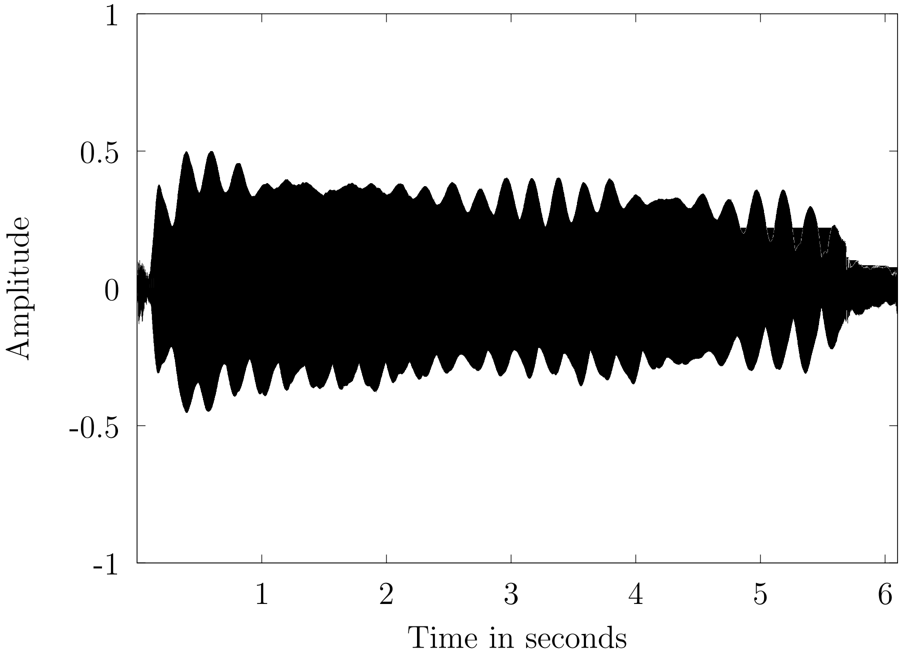

Figure 4.8: Time-stretched Saxophone sound showing original amplitude and frequency modulation.

with a 1000 Hz carrier frequency, a 5 Hz amplitude modulating frequency, and a 0.5 modulation depth. Sound example 2 in table 4.1 is the sound after 8 2LTDFT amplitude spectrum bins centered at bin 18 are removed. Sound example 3 in table 4.1 is sound example 2 stretched by a factor of 2 using a standard first order phase vocoder. Sound example 4 represents the sound after the original sub-audio amplitude modulation is imposed on example 3. In this example, the vibrato rate matches that of the original sound, but a time varying change in the vibrato can be heard. This unwanted variation has two parts, (i) glitching artifacts at the very beginning and end of the sound, and (ii), a gradual reduction in tremolo depth toward the middle and thereafter a gradual increase in tremolo depth to the end of the sound. Both of these problems are attributable to the phases of the 2LTDFT amplitude spectrum. When the phase spectrum of the 2LTDFT amplitude spectrum is altered by the addition of the interpolated bins from the first

Table 4.1: Example sub-audio modulation rate retention sounds.

Track	Sound	Description
1	Sine with AM	A synthetically generated Tremolo sound (AM sound)
2		Sound with sub-audio amplitude modulation removed
3		Sound with modulation removed and lengthened by 2
4		Sound lengthened by 2 with modulation re-imposed.
5	Sine with FM	A synthetically generated vibrato sound (FM sound)
6		Sound with sub-audio frequency modulation removed
7		Sound with modulation removed and lengthened by 2
8		Sound lengthened by 2 with modulation re-imposed
9	Saxophone	A saxophone tone
10		Sound with sub-audio modulation removed
11		Sound with modulation removed and lengthened by 2
12		Sound lengthened by 2 with modulation re-imposed
13		Sound lengthened by 2 with traditional phase vocoder
14	Violin	A violin tone
15		Sound with sub-audio frequency modulation removed
16		Sound with modulation removed and lengthened by 2
17		Sound lengthened by 2 with modulation re-imposed
18		Sound lengthened by 2 with traditional phase vocoder
19	Flute	A flute tone
20		Sound with sub-audio amplitude modulation removed
21		Sound with modulation removed and lengthened by 2
22		Sound lengthened by 2 with modulation re-imposed
23		Sound lengthened by 2 with traditional phase vocoder

2LTDFT, the evolution of the tremolo's depth is altered. The tremolo's frequency remains fixed because it is determined by the amplitude spectra of the 2LTDFT amplitude spectra. Attempts to alleviate this problem by randomizing the phases of the 2LTDFT amplitude spectra or leaving these phases unaltered resulted in sounds with less distinct tremolo characteristics, and it remains unclear how these artifacts can be reduced. We will see this problem in the other sound examples presented below.

Sound examples 5 through 8 in table 4.1 are the same sequence of sound manipulations for a frequency modulated sine tone which simulates a typical musical vibrato rate. Example 5 has a 1000 Hz carrier frequency, a 5 Hz frequency modulating frequency, and a 0.5 modulation depth. In this case, it can be seen that the problems with the phase of the 2LTDFT spectrum, discussed above, also produce perceptually salient artifacts in the resultant sound. The beginning and ending part of sound example 8 have no frequency modulation. However, the central portion of the sound has the correct rate of modulation as derived from the original sound.

Sound examples 9 through 12 in table 4.1 are the same sequence of manipulations for the saxophone sound pictured in the algorithm above. In this case, both the vibrato and tremolo were removed and re-imposed on the lengthened version of the sound. Although the artifacts discussed above can be heard in the version in which the modulation has been removed (example 10), the artifacts appear to be much less apparent than in the sine wave examples, when the vibrato and tremolo are re-imposed. This may be due to the relative complexity of the sound which masks the flaws in the process. Sound example 13 is a traditional phase vocoder time-stretching of the saxophone sound in which the slowing of the tremolo can be heard.

Sound examples 14 through 18 in table 4.1 represent a violin tone subjected to the same procedure. In this case only the 2LTDFT frequency spectrum was manipulated and not the 2LTDFT amplitude spectrum. As with the saxophone example, the artifacts are less perceptible than in the synthetic examples. Sound example 18 is a traditional phase vocoder time-stretch of the violin sound.

Sound examples 19 through 23 in table 4.1 represent a flute tone subjected to the process. As with the saxophone example, both the vibrato and tremolo characteristics are retained. In this case, the short glitching at the beginning and ending are present, but the variation in tremolo and vibrato depth appear absent.

4.6 Conclusions

We have seen that a method based on a higher order analysis of phase vocoder channel modulations can be used to alter the vibrato and tremolo characteristics of a sampled monophonic musical instrument sound. Our sample sounds show that the method is able to retain the tremolo and vibrato rates of musical sounds with fairly stable vibrato and tremolo rates. The method achieves better results on sounds which have prominent tremolo and less well with sounds with prominent vibrato. This may be attributable to our ability to more easily perceive inappropriate pitch changes than amplitude changes as discussed in the examples section above. The method was unsuccessful at manipulating vibrato and tremolo on human singing tones due to the added complication of formant frequencies present in the human voice.

The current version of the software precludes testing the algorithm on significantly longer sounds than the examples presented due to the computational complexity. Future versions of the software could be developed for more efficient computer languages and machines which would allow testing of the techniques efficacy on longer musical phrases.

A related line of research might be in devising other sonic transformations that are effectively carried out in the second order sound representation presented here.

Chapter 5

Negated Music

> *I unwrote that drawing because I was trying to write one with the other end of the pencil that had an eraser.*
>
> Robert Rauschenberg [41]

5.1 Introduction

This chapter concerns erasure and negation in the development of my own creative work involving what I call "negated music." Unlike the previous two chapters, this chapter will not introduce a new extension to the phase vocoder for musical uses. Instead, a particular artistic technique involving the phase vocoder spectral representation will be used for the author's own creative work. This chapter will first review works of art, both sonic and visual, involving negation and erasure with an emphasis on themes and ideas that have influenced my work in this area. Next, I discuss four of my sound works that trace my evolving engagement with the idea of negation with respect to my sound installations, sound sculptures, and conceptual audio works. This discussion will cover the technical procedures for their creation and the novel uses of the phase vocoder sound representation in the implementation of these works.

5.2 Erasure Based Visual Art

In 1953 Robert Rauschenberg persuaded Willem de Kooning to give him one of his drawings. Rauschenberg spent one month and forty erasers erasing the drawing, and claimed the finished erasure as his own original art work. This work consists of a paper which is covered with smudges created by an eraser. Art critic Calvin Tomkins discusses this work as a symbolic destruction of a leading artist father figure by a younger artist [89, 30]. However, Rauschenberg maintained that the work simply was intended to determine "whether a drawing could be made out of erasing" [9], yet Rauschenberg found his prior attempts at erasing his own works of art to be unsatisfactory compared to the completed *Erased de Kooning Drawing.* Rauschenberg's drawing points out that the marks left after an erasure tend to paradoxically highlight the original missing object. This heightened attention to the un-erased original makes the object of erasure significant to the success of works of art involving erasure. The choice of object of erasure is of paramount importance to the success of erasure based works of art.

The artistic use of erasure has been a subject of interest with the increases in digital photographic editing capability. Lev Manovich argues that the advent of digital manipulation of photographs has not caused a fundamental shift in the way we perceive photographic reality [54]. He instead suggests that digital photographic manipulation is part of a tradition of artistic and cultural manipulation of images that has existed long before digital techniques. Manovich cites Soviet era political photographs in which people have been removed as examples of this image making tradition. The relative ease of digital manipulation techniques has, nevertheless, allowed photographic artists to create works whose artistic content is based primarily on the removal of parts of a photographic image or objects within the image. The American artists Anthony Aziz and Sammy Cucher create photographic portraits in their *Dystopia Series* of 1995 in which the subject's facial features (mouth, eyes, ears, nasal passages) are digitally erased [13]. The erasure in this case consists of creating simulated flesh that takes the place of the missing features. The effect is striking, not because the figure has become more anonymous, but because the figure retains its individuality despite the horrific nature of the simulated mutilation. A similar technique is carried out by Venezuelian artist Alexander Apóstol in a series of photographs of urban buildings in which the doors and windows have been digitally removed [65]. The buildings become abstract monuments by the removal of the signs of their human use, and the effect is to heighten the inhuman nature of these buildings. Here erasure focuses our attention on latent characteristics of the buildings that would otherwise be unknown.

Charles Cohen's *Buff* series of photographs from 2001 to 2003 are different from the Aziz/Cucher and Apóstol photographs in that the photographic reality is decidedly broken in his removal of human figures from photographs that were originally pornographic [65]. Figures are removed as if by razor blade, leaving a pure white space occupying the location the figures had occupied. These negative spaces give a clear indication of the positioning and activity of the figures but give a new meaning to the images based on the negation of their prior use.

5.3 Erasure Based Sound Art

John Cage's well known silent piece *4' 33"* was preceded three years earlier by an unrealized electronic work he proposed in 1948. He proposed the concept of a piece called *Silent Prayer* in his article "A Composer's Confessions" as [69]:

> ...to compose a piece of uninterrupted silence and sell it to the Muzak Co. It will be 4 1/2 minutes long - these being the standard lengths of "canned" music, and its title will be "Silent Prayer." It will open with a single idea which I will attempt to make as seductive

> as the color and shape or fragrance of a flower. The ending will approach imperceptibility.

Although this piece is usually understood in the context of Cage's developing ideas about musical silence [69, 43], we can also think of *Silent Prayer* as an erasure of the Muzak that would otherwise have been played in retail and public spaces. Looking at this piece in relation to the idea that erasure and negation-based works of art are only significant in relation to the objects of erasure, the silence of *Silent Prayer* can be understood as a simple condemnation of the ubiquitous presence of particular types of recorded music in retail and public spaces. It appears that Cage's notion that silencing creates an aural space for listening to sounds that would otherwise be obscured is not yet present in *Silent Prayer*.

In 1996, the artist Jeremy Millar, while attempting to copy an audio tape of an interview he had conducted with the novelist J. G. Ballard, instead accidentally erased the master tape [13]. Millar now plays this erased tape as an artwork. With close listening one can perceive that something had been recorded on the tape, but is now inaccessible. The meaning of the erased silence of the Millar tape is different from that in Cage's *Silent Prayer*. Whereas Cage intended an erasure of the unwanted Muzak, Millar intends an attentive and regretful listening to the static byproduct of the unintended erasure. We are left concentrating on the Ballard interview despite its absence.

Matt Rogalsky's composition *S*, (2002) is a 24 CD set of "silence" created by editing out the words of 24 consecutive hours of a BBC 4 radio broadcast [45]. Each CD contains the electronically compiled "harvest" of the sounds of the gaps between the words from a single hour of radio broadcast. Unlike the Millar interview, which draws attention to the audio material that is missing, Rogalsky's erasure brings out the meaningful qualities of the sonic detritus that is left unerased. Referring to a later work of Rogalsky that uses the same technique, Jan Allen suggests that "the silences might signal hesitation; more often they stage emphasis. [the silences are] the zone in which what is suppressed or unsaid may be apprehended" [1]. Rogalsky's silences are understood here as linguistically rich sources of unspoken semantic content made prominent through their juxtaposition.

5.4 Towards a Negated Music

The previous sections have shown that erasure can be a significant artistic technique in contemporary visual and aural arts. The impulse toward negation in the arts does not always take the form of erasure. Sonic musical negation has evolved from a metaphorical or emotional negation, as in Cage's *Silent Prayer*, to the electronically mediated acoustic negations of Millar, Rogalsky, and the negation based sound works presented below

Musical theorists have speculated on what the nature of a negated sound might be. In his 1893 article "The Music of Negation," J. A. Fuller Maitland points out that one of music's expressive shortcoming's is its inability to "bring before us the absence of certain features; it can tell us what the heroine of an opera is feeling, but it is powerless to suggest what she is not going through" [53]. In 1922, Alexander Brent-Smith explained this shortcoming in terms of a lack of a sound that signals the negative in its presence [8]. He presents the example of setting to music a poem by Elroy Flecker: "Mute is battle's brazen horn, that rang for Priest and King." Brent-Smith states that "if [the composer] does not somehow suggest the horn the thought is incomplete, and if he as much as suggests a horn the thought is incorrect." He concludes that "not until music has discovered one single sound that shall negative [sic] its assertions" will this type of musical dilemma be solved. Presumably when it is found, this single negative sound would be played before, concurrently, or after the sounds that suggest horns in the composition. I believe Brent-Smith's suggestion of a sound which can negate its own evocative qualities or those of another sound is the first such speculation on the potential existence of an actual negated sound.

In 1974, Frederick Taylor posited a system of musical logic that could make more apparent what "can be meaningfully examined and critized" in a musical composition [87]. Peter Gibbins argues against Taylor's system of musical logic as a method of increasing our understanding of musical structure, pointing out that one of the shortcomings of Taylor's system is its lack of a concept of negation, and that mathematical negation is a key element in any system of logic [32]. Gibbins, like Brent-Smith above, acknowledges the difficulty of conceiving of a negated sound that would exist in a system of musical logic:

> It is difficult to see what the negation of a sound segment could be, for it would have to be some other sound segment. If we argue that consistency is the primitive notion and that a negation of a sound-segment is a second sound-segment inconsistent with the first, we are led to the result that y may be a negation of x while x may not be a negation of y. A 'negation' with this property is at best a peculiar sort of negation. For a sound-segment may have two 'negations' which are also 'negations' of one another [32].

Gibbins clearly gives serious consideration to the idea that a negated sound could exist as a sonic phenomenon and not just as a byproduct of musical notation manipulations. While it may be unclear what sonic "consistency" might be to Gibbins, his idea that a sound may have multiple negations is relevant to our understanding of the erasure based sound works of Cage, Millar and Rogalsky, as well as to the negated music concept presented below.

Stan Link discusses many possible meanings and functions of musical silence in his essay "Much Ado about Nothing" [50]. One of these meanings is that musical silence signifies absence or nothingness. Link states "quietness evokes nothingness as pointedly as human perception might allow." Later, Link wonders if silence is the "only token of negation in musical contexts." He also asks, "can negation involve sound as well as silence?" Link, like Brent-Smith and Gibbins, does not provide a recipe for producing a negated sound, but does suggest that the function of such a sound would be to "make us notice nothing," that is, a sound "which is blank regarding cognitive content – mute at its core." Unlike Brent-Smith's negated sound which would evoke the opposite of the meaning of the sound, or Gibbin's negated sound which would complete a system of musical logic, Link's negated sound would, like musical silence, evoke the idea of absence or nothingness.

My intention with the following sound works is to create a sonic realization of these ideas about how a negated sound might have meaning by creating my own technique for producing negated sounds. While the negated sounds in each work may not achieve these specific goals of the authors discussed above, they provide a first material instantiation of a negated sound.

5.5 Negation Based Sound Works

In this section the author's own soundworks involving negated sound will be presented to show the author's evolving technical and conceptual idea of negated sound.

5.5.1 *Irresonance*

Irresonance (2004) is a sound installation that uses feedback to resonate eight small brass tubes. One of these tubes is shown in Figure 5.1. Each of these 1 cm diameter tubes is mounted to the wall of an exhibition space and is equipped with a small piezoelectric loudspeaker at the bottom of the tube. The tubes are connected to each other via a single series connection, so that only a single strand of bare wire connects each tube to the next. This connection is shown in Figure 5.2. Each of these tubes is a different length, with the shortest being 20 cm and the longest 80 cm, and is topped with a flame dispersion cap designed for a bunsen burner. The differing lengths of the tubes, along with these bunsen burner tops, give each tube its unique resonant characteristics.

In addition to this primary visual component of the work, there are a number of hidden electronic components. A single microphone located near the center

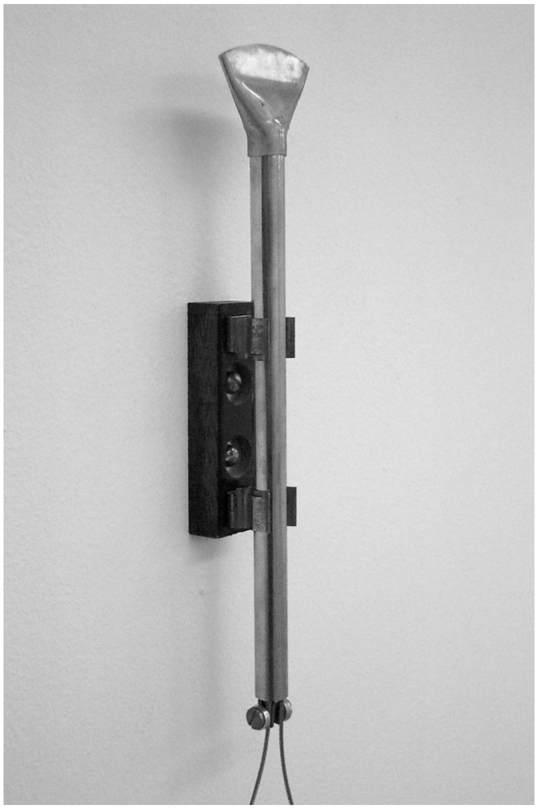

Figure 5.1: One tube of *Irresonance* sound installation.

of the installation space is placed near the ceiling. This microphone is meant to go unnoticed by the viewers and listeners. In addition, a computer, a preamplifier, and an audio amplifier are concealed in an adjacent space.

Irresonance functions as a live interactive sound installation, whose sound is constantly changing. The sound is influenced by the resonant characteristics of the tubes, the ambient sounds of the space, any sounds made by a viewer/listener in the space, and the state of the audio processing network in the computer. The main sonic idea of *Irresonance* is the creation of an acoustic feedback loop between the speakers in the tubes and the microphone in the space. The familiar high frequency howl of an electro-acoustic sound system that is feeding back on itself is carefully controlled by computer processing between the microphone and the speakers. At all times, the loudest frequency present in the mix of sounds captured

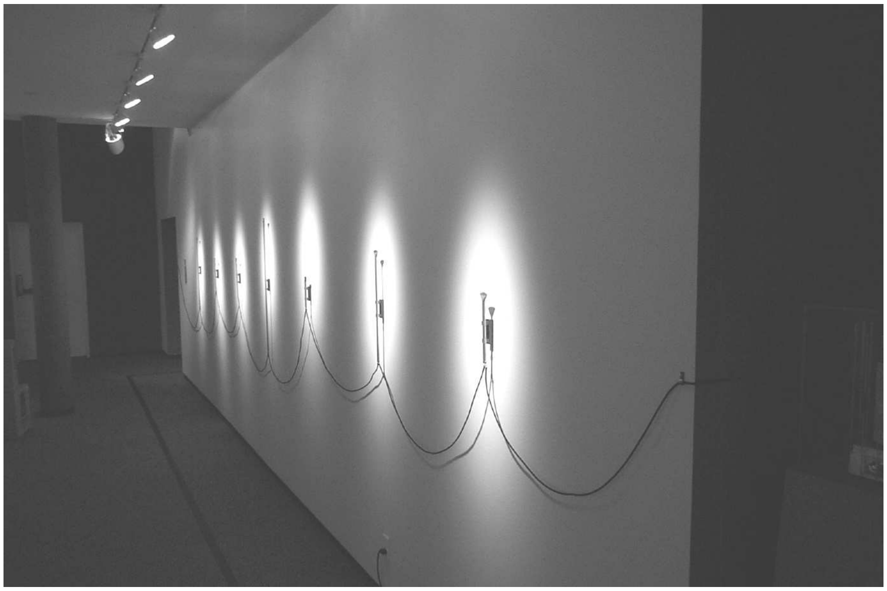

Figure 5.2: Installation view of *Irresonance* sound installation.

by the microphone is subtracted from the spectral domain representation of the sound. This results in a constantly shifting sound produced by the tubes because the computer combats the highly resonant tendencies of the different tubes. The title of the installation *irresonance* is a word that is no longer in regular use, and means, "lacking resonance". Here the tube with the loudest resonant frequency are removed from the mix of sounds produced, or "irresonated."

Even though each of the transducers in the tubes are activated by the same signal, the sounds produced by each tube are quite different. The resonance of each tube effectively filters frequencies that are not at a resonant frequency of that tube. In this way, a spatial component of the installation is created with only a single channel audio system. By giving the transducing element individual resonant characteristics, they create their own distinctive localized sonic developments in response to a single combined source.

The computer processing for this installation is performed in real time by Max/MSP [91]. The main irresonating function of the patch is carried out by reducing to zero the spectral bins around the fundamental frequency for each spectral frame. The fundamental frequency is determined by use of the `fiddle~` object of Puckette [59]. The frequency in hertz is converted and rounded to the spectral bin number corresponding to its spectral position in units of bin. For each spectral frame, the amplitude of this bin and its two adjacent bins on each side is set to zero. This new amplitude spectrum is combined with the original phase spectrum and transformed back to the time domain.

In addition to this removal of the most prominent spectral energy, the signal is compressed in dynamic range. The process not only prevents the sound from becoming too loud or too soft, but also allows softer tones to become more prominant in the spectrum.

This installation was my first work that deals with the idea of negating spectral components of a sound. The idea began as a method of controlling feedback in the installation, and grew into the primary way of creating a spatial and time evolving form in the installation. This negating of spectral components develops into the concept of negated music in the following installations and audio works discussed below.

5.5.2 *Inverse Music Box*

This section will describe the author's sound sculpture *Inverse Musicbox* (2005) and describe its relationship to negated music. The *Inverse Music Box* sculpture is based on a Sankyo paper strip manivelle music box mechanism. The Sankyo paper strip music box was invented by Komatsu Fumito and Tashiro Kazuo in 1968.[1] The sounding tines of this music box are activated by holes punched in paper cards that are fed through the mechanism by a hand crank.[2] In the *Inverse Music Box* the manivelle mechanism is attached to the top of a 1930s

[1] An interesting history of the Sankyo music box can be found in Murakami [62].

[2] This type of music box uses a "manivelle" mechanism, as differentiated from a spring or electric motor mechanism.

era "Sonora" tube radio external speaker cabinet, which is attached to a wall. The Sankyo mechanism is designed to produce twenty distinct chromatic pitches. In this sculpture the paper cards are permanently arranged in a loop through the mechanism. Figure 5.3 shows the *Inverse Music Box* installed. In addition to these visible mechanical aspects, the sculpture features a hidden computer, speaker, and microphone. Mounted inside the Sonora speaker box is a modern loudspeaker, and a piezoelectric microphone is mounted directly underneath the music box mechanism. The cabling for both of these transducers runs to a hidden computer.

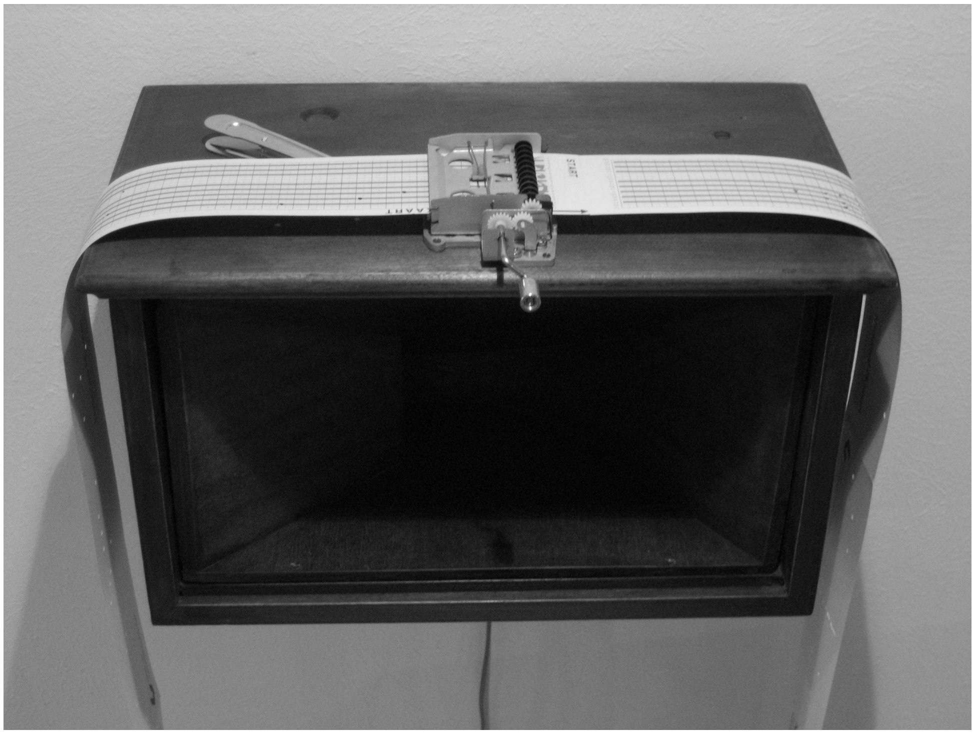

Figure 5.3: *Inverse Music Box* sound sculputre.

A viewer/listener is invited to turn the hand crank to advance the cards and trigger the notes, Notes can also be added with the hole punch provided. The inversion of this piece is performed by the hidden electronics. Instead of each note of the music box simply being amplified by the speaker, each struck note is *subtracted* spectrally from the sound the speaker is making and each of the twenty notes that is *not* plucked remains sounding. For example, if a chord of three notes is plucked, the sound produced by the speaker consists of a chromatic tone mass of the remaining seventeen tones. Only by triggering all 20 tones simultaneously does the speaker remain silent.

A real-time Max/MSP patch is used to calculate the “negated sounds” of this sculpture. The patch consists of two parts: a note onset detection unit, and a spectral subtraction unit. The note onset unit uses a simple amplitude thresholding function to create a trigger when any note or combination of notes is activated. The sound produced by the patch is created in the spectral domain. Without any triggered input sound, the patch creates a static droning sound from a single STFT of all the twenty notes playing simultaneously. In order to make sure that this spectrum contained all the frequencies of each note, this STFT was artificially constructed by summing the STFT of each individual note played separately. In order to create a continuous droning sound with this single STFT, the phases are advanced with the traditional phase vocoder method and combined with the static amplitude spectrum.

Each time a new note sets off the amplitude threshold trigger, the STFT taken at the time of the trigger is subtracted from this static combined amplitude spectrum. This new amplitude spectrum is now used to create the droning sound until a new note is triggered. Each time a note or combination of notes is triggered, it is subtracted from the original combined amplitude spectrum. The result is a continuous tone which suddenly changes to a new state that is missing any plucked note from its complex.

The *Inverse Music Box* makes a connection between the physical removal or subtraction of material in a musical score via a hole punch and the idea of removal of notes as a sonic phenomenon. Here the dual absence of the note, both in the paper card and the sonic complex, paradoxically draws attention to that note. The user removes the note from the paper and hears its removal from the sound.

5.5.3 *Whiteout*

Whiteout (2007) is a sound sculpture consisting of a handmade paper loudspeaker, an mp3 player, and a site-specific sound. The square speaker cone is approximately 50 cm on each side and is composed of a sheet of thick, pulpy, and rigid paper, attached to the wall by four push pins in the corners. Figure 5.4 shows the *Whiteout* installation. An approximately 3 cm diameter loudspeaker is affixed to the center back of the paper. The paper becomes the speaker "cone" since it is attached directly to the voice coil of the speaker. A small amplifier and mp3 player are connected to this speaker and are mounted nearby on the wall.

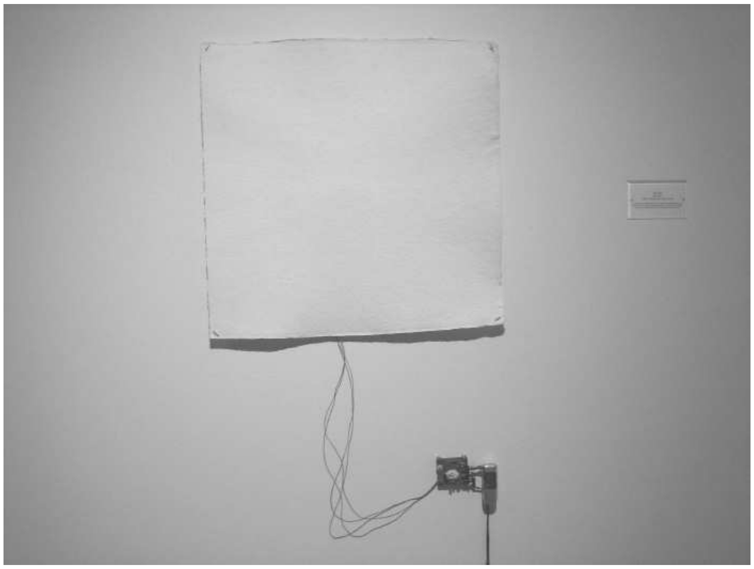

Figure 5.4: *Whiteout* sound sculpture.

Each time the sculpture is installed in a new location, a new sound is created to be played at that location using the following method. A recording of approximately five minutes is made of the ambient sounds of the installation space. This recording is transformed into the LTDFT domain as discussed in Chapter 4.

The LTDFT domain is simply a single FFT of the entire five minute sound. The amplitude and phase spectra are calculated from the LTDFT, and a transformation is made to the phase spectrum before recombining it with the unaltered amplitude spectrum. This new LTDFT sound is converted back to the time domain by an inverse FFT. As discussed by Hammer and Sundt, completely randomizing the phase spectrum results in a virtually un-changing drone sound with the amplitude of spectral components proportional to their overall presence during the five minute recording. In order to create a sound that is unrecognizable in relation to its original time evolution, like the drone sound, but has *some* spectral evolution, an ad hoc method of modifying the phase spectrum was devised via experimentation. This consists of squaring each value of the phase spectrum and wrapping the new phase values between 0 and 2π. Many other arbitrary transformations to the LTDFT phase spectrum are possible as outlined by Hammer and Sundt [35]. This squaring method has no particular significance beyond providing an interesting balance between change and stasis in the resultant sound.

The idea of this piece is to mirror the undifferentiated visual field of the slightly textured white paper to the continuous and slightly undulating sound created from all the sound present in the space. The white of the paper can be thought of as a combination of all the colors present in the space, and similarly the sound is a combination of all the sonic frequencies present in the space. This piece does not feature a negated component, but introduces the important sonic "ground" concept that is necessary in subsequent works.

5.5.4 *Negated Music Series*

The project *Negated Music Series* (2008) consists of a series of short video works. Each video is between 40 seconds and 1 minute 20 seconds in duration, and each consists of a single continuous shot in which a sonic event is accompanied by a soundtrack created by the use of the negated music process. Negated music is created from the average of all the sounds present during the entire recording, minus the sound that is happening at each moment in time. A technical description of the conceptual idea of negated music is presented below.

Negative music is a combination and refinement of two concepts presented above. The spectral subtraction idea of the *Inverse Music Box* and the long term average sound technique of *Whiteout* are combined. The signal processing method employed in the *Negated Music Series* is as follows: (i) the LTDFT of the original sound recorded concurrently with the video taping session is performed; (ii) the phase spectrum of this LTDFT is randomized; (iii) the inverse LTDFT is calculated using this new randomized phase spectrum; this creates a droning sound whose spectral content is proportional to the frequency content present over the entire duration of the sound; (iv) a phase vocoder analysis of both the original and the averaged sound is performed; (v) the amplitude spectrum of each STFT frame of the original is then subtracted from the STFT amplitude spectrum of the averaged sound to generate a new set of STFT amplitude spectra, (spectral bins whose difference is less than zero are set to zero in this new amplitude spectrum); and (vi) the new amplitude spectra are combined with the original phase spectra and transformed back to the time domain. Figure 5.5 shows this process as a flowchart. This process results in an abstract changing sound that can be thought of as one possible negation of a sound made audible by placing it in a context of the average of the sound.

The video works of the *Negated Music Series* examine the sounds of musical instruments and objects by showing video of an object in relation to the audio that has been subject to the negated music process. The project is conceived of as an open ended set of videos constructed in this manner. Currently eight videos are completed. They are: *Accordion*, *Clarinet*, *Clock Chimes*, *Drum*, *Player Piano*, *Spinning Mobile*, *Toy Piano*, and *Wind Chimes*. Further discussion of three of these are presented here.

The *Drum* video is a good introduction to the sonic nature of the negated

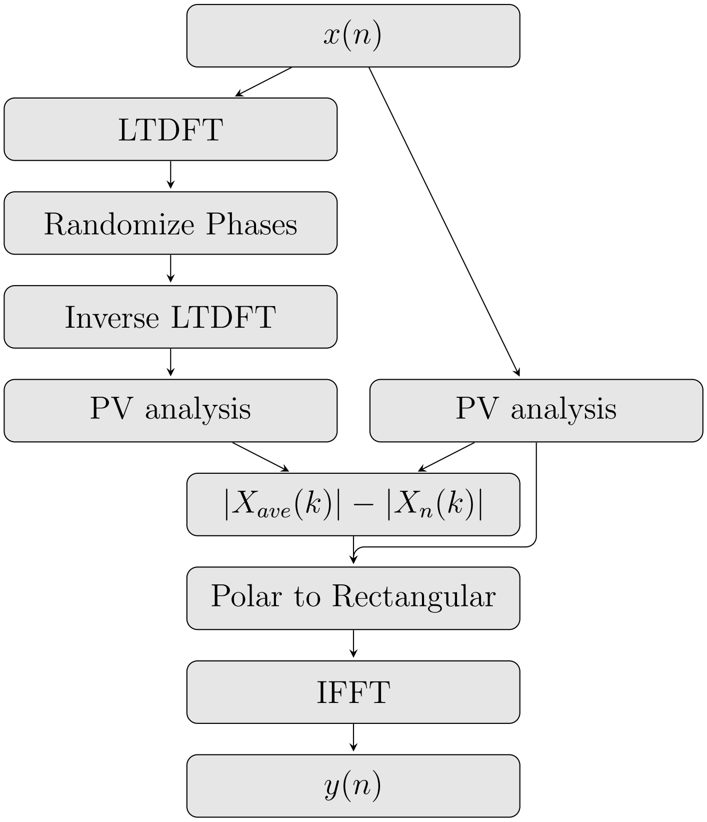

Figure 5.5: *Negated Music Series* analysis transformation synthesis method.

music process. The video depicts a snare drum for 22 seconds followed by a drum hit. This is the only perceptible sonic event that happens in the original video. The negated music process produces a sound that is continuous throughout the entire video, and contains all of the spectral components of the snare drum sound. When the drum is hit, this sound is subtracted from the averaged sound to quiet the sound during the hit. This video and soundtrack clearly shows the inverse nature of the negating process.

The *Clarinet* video exhibits other aspects of the process. In this video a single E flat tone is held for an entire breath, there is a brief pause, followed by another E flat held tone. During these tones, the player was purposely tuning the note sharp and flat at different times. After processing, the tone can be heard as a quiet high frequency sound during the original note's times, and as a tone much like the original clarinet E Flat during the breath rest in the middle. This again shows the inverse relationship to traditional sound production that exists with this technique. Figure 5.6 shows one frame of the *Clarinet* video. The *Toy Piano* video shows the inside mechanism of a toy piano music box that is playing the song "Cielito Lindo." Here we can hear the individual negated notes as we see them struck in the piano. This process results in a ghostly impression of the original melody with the original rhythm intact. Figure 5.7 shows one frame of

Figure 5.6: *Clarinet* video.

Figure 5.7: *Toy Piano* video.

the *Toy Piano* video.

The *Negated Music Series* is a piece of musical conceptual art, as opposed to a musical composition. As such, it is also intended to reference the ideas of negation in twentieth century art and music more broadly than our narrow discussions of musical negation presented here. Many twentieth century art movements can be seen as negations: abstract art's negation of the figure, conceptual art's negation of the art object, minimalism's negation of self expression, and musical minimalism's negation of musical development. Musical works take place through the continuous unfolding of sound through time. The specific conception of sonic negation presented here highlights the absent presence of a sound. The new sounds negate the continuous presence of sound and through this process suggest a negation of music's primary condition of existence.

Appendix A

Sinusoidality Error Analysis

Six different synthetic sounds are used to test the sinusoidality measures; (i) a sine tone without noise, (ii) pure noise, (iii)a sine tone with noise, (iv) 32 unevenly spaced sine tones with noise, (v) a harmonic sound with noise, and (vi) a sinusoid with time-varying frequency (chirp signal) with noise.

Table A.1: Sinusoidality error analysis for 440 Hz sine wave with no noise.

	False Error		Missed Error		Total Error	
Method	μ	σ	μ	σ	μ	σ
Zeros	–	–	1.000	0.000	1.000	0.000
Ones	–	–	0.000	0.000	0.000	0.000
Power	–	–	0.366	0.000	0.366	0.000
Power Persistence	–	–	0.508	0.055	0.508	0.055
Sigmund	–	–	0.719	0.040	0.719	0.040
Charpentier	–	–	0.023	0.035	0.023	0.035
Phase Acceleration	–	–	0.007	0.068	0.007	0.068
Phase Acc. × Power Per.	–	–	0.433	0.063	0.433	0.063
Cross-Correlation	–	–	0.481	0.000	0.481	0.000
Variance	–	–	0.587	0.030	0.587	0.030
Harmonic Sum	–	–	0.733	0.018	0.733	0.018

Table A.2: Sinusoidality error analysis for white noise only.

	False Error		Missed Error		Total Error	
Method	μ	σ	μ	σ	μ	σ
Zeros	0.000	0.000	–	–	0.000	0.000
Ones	1.000	0.000	–	–	1.000	0.000
Power	0.211	0.030	–	–	0.211	0.030
Power Persistence	0.043	0.016	–	–	0.043	0.016
Sigmund	0.235	0.035	–	–	0.235	0.035
Charpentier	0.272	0.017	–	–	0.272	0.017
Phase Acceleration	0.890	0.043	–	–	0.890	0.043
Phase Acc. × Power Per.	0.120	0.030	–	–	0.120	0.030
Cross-Correlation	0.251	0.026	–	–	0.251	0.026
Variance	0.090	0.014	–	–	0.090	0.014
Harmonic Sum	0.268	0.037	–	–	0.268	0.037

Table A.3: Sinusoidality error analysis for a sine tone of 440 Hz with white noise. The energy of the sine tone is equal to the total energy of the white noise.

	False Error		Missed Error		Total Error	
Method	μ	σ	μ	σ	μ	σ
Zeros	0.000	0.000	1.000	0.000	0.516	0.004
Ones	0.984	0.003	0.004	0.003	0.491	0.005
Power	0.001	0.001	0.366	0.002	0.189	0.002
Power Persistence	0.000	0.001	0.561	0.165	0.288	0.084
Sigmund	0.000	0.000	0.721	0.037	0.557	0.058
Charpentier	0.181	0.036	0.122	0.064	0.153	0.045
Phase Acceleration	0.491	0.015	0.038	0.078	0.262	0.038
Phase Acc. × Power Per.	0.000	0.000	0.495	0.190	0.254	0.096
Cross-Correlation	0.006	0.006	0.481	0.006	0.251	0.007
Variance	0.103	0.018	0.854	0.073	0.491	0.041
Harmonic Sum	0.217	0.010	0.489	0.002	0.358	0.006

Table A.4: Sinusoidality error analysis for 32 equal amplitude sine waves with a random distribution of frequencies between 40 Hz and 10 000 Hz and white noise. The total energy of the sine tones is equal to the total energy of the white noise.

	False Error		Missed Error		Total Error	
Method	μ	σ	μ	σ	μ	σ
Zeros	0.000	0.000	1.000	0.000	0.513	0.013
Ones	0.859	0.012	0.050	0.010	0.517	0.009
Power	0.007	0.003	0.695	0.059	0.371	0.033
Power Persistence	0.003	0.002	0.607	0.099	0.336	0.055
Sigmund	0.017	0.003	0.665	0.042	0.359	0.024
Charpentier	0.101	0.009	0.690	0.039	0.435	0.024
Phase Acceleration	0.420	0.020	0.158	0.081	0.325	0.039
Phase Acc. × Power Per.	0.001	0.001	0.655	0.090	0.360	0.049
Cross-Correlation	0.054	0.012	0.693	0.049	0.406	0.024
Variance	0.050	0.011	0.913	0.020	0.524	0.011
Harmonic Sum	0.118	0.016	0.644	0.066	0.417	0.043

Table A.5: Sinusoidality error analysis for harmonic sound with a fundamental frequency of 440 Hz and 31 harmonics each generated with 1 dB less energy than the prior. The energy of the harmonic tone is equal to the total energy of the white noise.

	False Error		Missed Error		Total Error	
Method	μ	σ	μ	σ	μ	σ
Zeros	0.000	0.000	1.000	0.000	0.545	0.006
Ones	0.867	0.017	0.042	0.012	0.496	0.009
Power	0.002	0.001	0.755	0.012	0.417	0.010
Power Persistence	0.001	0.001	0.929	0.027	0.512	0.017
Sigmund	0.002	0.000	0.824	0.013	0.556	0.019
Charpentier	0.147	0.034	0.509	0.047	0.359	0.025
Phase Acceleration	0.411	0.020	0.142	0.124	0.301	0.062
Phase Acc. × Power Per.	0.000	0.001	0.862	0.052	0.475	0.031
Cross-Correlation	0.027	0.005	0.746	0.035	0.425	0.018
Variance	0.001	0.001	0.885	0.004	0.489	0.010
Harmonic Sum	0.145	0.008	0.387	0.022	0.292	0.016

Table A.6: Sinusoidality error analysis for a single sine tone with time-varying frequency logarithmically changing from 220 Hz to 880 Hz over one half second. The energy of the time-varying tone is equal to the total energy of the white noise.

	False Error		Missed Error		Total Error	
Method	μ	σ	μ	σ	μ	σ
Zeros	0.000	0.000	1.000	0.000	0.503	0.032
Ones	0.986	0.005	0.004	0.003	0.500	0.033
Power	0.001	0.001	0.363	0.067	0.185	0.043
Power Persistence	0.001	0.001	0.527	0.201	0.271	0.119
Sigmund	0.001	0.001	0.280	0.111	0.140	0.054
Charpentier	0.134	0.014	0.473	0.355	0.313	0.190
Phase Acceleration	0.503	0.018	0.295	0.292	0.409	0.146
Phase Acc. × Power Per.	0.000	0.001	0.635	0.248	0.325	0.142
Cross-Correlation	0.011	0.025	0.424	0.109	0.218	0.056
Variance	0.105	0.017	0.956	0.058	0.533	0.024
Harmonic Sum	0.133	0.017	0.419	0.080	0.279	0.038

References

[1] Allen, J., 2005: Speaking through silence: thoughts on the limits of articulation and the nature of curatorial authority. In *Unspoken Assumptions: Visual Art Curators in Context*, 1–6. Ontario Association of Art Galleries.

[2] Alonso, M., David, B., and Richard, G., 2003: A study of tempo tracking algorithms from polyphonic music signals. *4th COST 276 Workshop*.

[3] Apel, T., 1993: *Transformation of Audio Signals by use of the McAulay-Quatieri Sinusoidal Model of Sound*. Master's thesis, Dartmouth College.

[4] Arfib, D., and Delpart, N., 1998: Selectrive transformations of sounds using time-frequency representations: An application to the vibrato modification. In *104th Convention of the Audio Engineering Society*, 1–7.

[5] Arfib, D., and Delpart, N., 1999: Alteration of the vibrato of a recorded voice. In *International Computer Music Conference*, 186–189.

[6] Arfib, D., Keiler, F., and Zölzer, U., editors, 2002: *DAFX - Digital Audio Effects*. John Wiley and Sons, LTD.

[7] Boulanger, R. C., 2000: *The Csound book : perspectives in software synthesis, sound design, signal processing, and programming*. MIT Press, Cambridge, Mass.

[8] Brent-Smith, A., 1922: The negative in music. *The Musical Times*, **63**(958), 839–840.

[9] Breslin, J. E. B., 1993: *Mark Rothko: A biography*. University of Chicago Press.

[10] Charpentier, F., 1986: Pitch detection using the short-term phase spectrum. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP*, **11**, 113– 116.

[11] Cooley, P., and Tukey, J., 1965: An algorithm for the machine computation of complex fourier series. *Mathematics of Computation*, **19**, 297–301.

[12] De Goetzen, A., Bernatdini, A., and Arfib, D., 2000: Traditional (?) implementations of a phase-vocoder: The tricks of the trade. In *Proceedings of the International Conference on Digital Audio Effects (DAFX-00)*. Verona.

[13] Dillon, B., 2006: The revelation of erasure. *Tate Etc.*, **8**.

[14] Dobson, R., 2006: http://www.dsprelated.com/showmessage/69091/1.php.

[15] Dolson, M., 1982: *A Tracking Phase Vocoder and its use in the Analysis of Ensemble Sounds*. Ph.D. thesis, California Institute of Technology.

[16] Dolson, M., 1986: The phase vocoder - a tutorial. *Computer Music Journal*, **10**(4), 14–27.

[17] Dorran, D., Coyle, E., and Lawlor, R., 2004: An efficient phasiness recuction technique for moderate audio time-scale modification. In *Proceedings of the 7th International Conference on Digital Audio Effects (DAFX-04)*.

[18] Dressler, K., 2006: Sinusoidal extraction using an efficient implementation of a multi-resolution fft. *Proceedings of 9th International Conference on Digital Audio Effects (DAFx-06)*, 247–252.

[19] Dubnov, S., 1999: Hos method for phase characterization in sinusoidal models with applications for speech and audio. *IEEE Signal Processing Workshop on Higher-Order Statistics*, 1–5.

[20] Dubnov, S., 2004: Generalization of spectral flatness measure for non-gaussian linear processes. *IEEE Signal Processing Letters*, **11**(8), 698–701.

[21] Dubnov, S., 2006: *YASAS - Yet Another Sound Analysis-Synthesis Method*. International Computer Music Conference.

[22] Dudley, H., 1939: The vocoder. *Bell Labs Record*, **18**, 122–126.

[23] Duxbury, C., Davies, M., and Sandler, M., 2001: Separation of transient information in musical audio using multiresolution analysis techniques. *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01)*.

[24] Duxbury, C., Davies, M., and Sandler, M., 2002: Improved time-scaling of musical audio using phaes locking at transients. In *112th Audio Engineering Society Convention*.

[25] Erbe, T., 1994: *Soundhack Manual*. Frog Peak Music.

[26] Ferreira, A., 1999: An odd-dft based approach to time-scale expansion of audio signals. *IEEE Transactions on Speech and Audio Processing*, **7**(4).

[27] Fitz, K., and Haken, L., 1995: Bandwidth enhanced sinusoidal modeling in lemur. *Proceedings International Computer Music Conference.*

[28] Flanagan, J., and Christensen, S., 1980: Computer studies on parametric coding. *Journal of the Acoustical Society of America*, **68**(2).

[29] Flanagan, J., and Golden, R., 1966: Phase vocoder. *Bell Systems Technical Journal*, **45**, 1493–1509.

[30] Galpin, R., 1998: Erasure in art: Destruction, deconstruction, and palimpsest. http://www.users.zetnet.co.uk/richart/texts/erasure.htm.

[31] Geoffroy Peters, X. R., 1998: Signal characterization in terms of sinusoidal and non-sinusoidal components. *Proceedings DAFX98.*

[32] Gibbins, P., 1976: Logics as models of music. *The British Journal of Aesthetics*, **16**(2), 157–160.

[33] Gordon, J. W., and Strawn, J., 1985: *An Introduction to the Phase Vocoder.* Digital Audio Signal Processing: An Anthology. W. Kaufmann Inc., Los Altos, CA.

[34] Griffin, D., and Lim, J., 1988: Multiband excitation vocoder. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **36**(8), 1223–1235.

[35] Hammer, Ø., and Sundt, H., 1999: Musical applications of decomposition with global support. In *Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99).*

[36] Hanna, P., and Desainte-Catherine, M., 2002: Detection of sinusoidal components in sounds using statistical analysis of intensity fluctuations. *International Computer Music Conference.*

[37] Hanna, P., and Desainte-Catherine, M., 2003: Using statistical analysis of the intensity fluctuations to detect sinusoids in noisy signals. url="citeseer.ist.psu.edu/658807.html".

[38] Hartmann, W. M., 1998: *Signals, Sound, and Sensation.* Springer-Verlag.

[39] Helmholtz, H., 1954: *On the Sensations of Tone.* Dover Publications.

[40] Herrera, P., and Bonada, J., 1998: Vibrato extraction and parameterization in the spectral modeling synthesis framework. *Proceedings of the Digital Audio Effects Workshop.*

[41] Hess, B., 2004: *Willem de Kooning 1904-1997: Content as a Glimpse.* Taschen.

[42] J, K., S, M., and S, W., 1994: Caractérisation du timbre des sons complexes. ii: Analyses acoustiques et quantification psychophysique. *Journal de Physique*, **4**(C5)), 625–628.

[43] Kahn, D., 1999: *Noise, Water, Meat: A History of Sound in the Arts*. MIT Press.

[44] Keiler, F., and Marchand, S., 2002: Survey on extraction of sinusoids in stationary sounds. *Proceedings of the Digital Audio Effects Workshop.*

[45] LaBelle, B., 2006: *Background Noise: Prospectives on Sound Art*. Continuum International.

[46] Laroche, J., and Dolson, M., 1997: About this phasiness business. In *International Computer Music Conference*. Thessaloniki, Greece.

[47] Laroche, J., and Dolson, M., 1997: Phase-vocoder: About this phasiness business. In *Proceedings IEEE ASSP Workshop on the Application of Signal Processing to Audio and Acoustics*. New Paltz, NY.

[48] Laroche, J., and Dolson, M., 1999: Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing*, **7**(3), 323–332.

[49] Laroche, J., and Dolson, M., 1999: New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing, and other exotic audio modifications. *Journal of the Audio Engineering Society*, **47**(11), 928–936.

[50] Link, S., 1995: Much ado about nothing. *Perspectives of New Music*, **33**(1/2), 216–272.

[51] Magnus, C., 2001: Real-time separation of periodic and non-periodic signal components. Unpublished paper.

[52] Maher, R., and Beauchamp, J., 1990: An investigation of vocal vibrato for synthesis. *Applied Acoustics*, **30**, 219–245.

[53] Maitland, J. A. F., 1893: The music of negation. *The Musical Times and Singing Class Cirular*, **34**(602), 207–208.

[54] Manovich, L., 2003: The paradoxes of digital photography. In *The Photography Reader*, editor L. Wells. Routledge.

[55] Marchand, S., 1998: Improved spectral analysis precision with an enhanced phase vocoder using signal derivatives. In *Proceedings of the International Conference on Digital Audio Effects.*

[56] Marchand, S., and Raspaud, M., 2004: Enhanced time-stretching using order-2 sinusoidal modeling. In *Proceedings of the International Conference on Digital Audio Effects*.

[57] Martin, R., 1994: Spectral subtraction based on minimum statistics. *Proceedings of EUSIPCO-94 Seventh European Signal Processing Conference*.

[58] McAulay, R. J., and Quatieri, T. F., 1986: Speech analysis/synthesis based on a sinusoidal representation. *Ieee Transactions on Acoustics Speech and Signal Processing*, **34**, 744–754.

[59] Miller Puckette, D. Z., Theodore Apel, 1998: Real-time audio analysis tools for pd and msp. *International Computer Music Conference*.

[60] Moore, F. R., 1990: *Elements of Computer Music*. Prentice-Hall.

[61] Moorer, J. A., 1978: Use of phase vocoder in computer music applications. *Journal of the Audio Engineering Society*, **26**(1-2), 42–45.

[62] Murakami, K., 1999: Sankyo paper strip manivelle - history. http://www.mmdigest.com/Archives/Digests/199908/1999.08.19.html.

[63] N. S. Jayant, P. N., 1984: *Digital Coding of Waveforms*. Prentice-Hall.

[64] Oppenheim, A. V., and Schafer, R. W., 1998: *Discrete-Time Signal Processing*. Prentice-Hall, Inc.

[65] Paul, C., 2003: *Digital Art*. Thames and Hudson.

[66] Portnoff, M. R., 1976: Implementation of digital phase vocoder using fast fourier-transform. *IEEE Transactions on Acoustics Speech and Signal Processing*, **24**(3), 243–248.

[67] Portnoff, M. R., 1980: Time-frequency representation of digital signals and systems based on short-time fourier transform. *Ieee Transactions on Acoustics, Speech, and Signal Processing*, **28**(1), 55–69.

[68] Portnoff, M. R., 1981: Time-scale modification of speech based on short-time fourier analysis. *Ieee Transactions on Acoustics, Speech, and Signal Processing*, **29**(3).

[69] Pritchett, J., 1993: *The Music of John Cage*. Cambridge University Press.

[70] Puckette, M., 2007: *The Theory and Technique of Electronic Music*. World Scientific Press.

[71] Puckette, M. S., 1995: Phase-locked vocoder. In *Proceedings IEEE ASSP Workshop on the Application of Signal Processing to Audio and Acoustics.* New Platz, NY.

[72] Puckette, M. S., and Brown, J. C., 1998: Accuracy of frequency estimates using the phase vocoder. *IEEE Transactions on Speech and Audio Processing*, **6**(2), 166–176.

[73] Quatieri, T. F., 2002: *Discrete-Time Speech Signal Processing: Principles and Practice.* Prentice Hall.

[74] Quatieri, T. F., and McAulay, R. J., 1986: Speech transformations based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **34**, 1449–1464.

[75] Rabiner, L. R., and Schafer, R. W., 1968: *Digital Processing of Speech Signals.* Prentice-Hall.

[76] Reich, S., 1967: *Writings about Music.* The Press of the Nova Scotia College of Art and Design, Halifax.

[77] Röbel, A., 2003: Transient detection and preservation in the phase vocoder. *Proceedings of the International Computer Music Conference.*

[78] Rodet, X., 1998: Musical sound signals analysis/synthesis: Sinusoidal+ residual and elementary waveform models. *IEEE Time-Frequency and Time-Scale Workshop.*

[79] Rossignol, S., Depalle, P., Soumagne, J., Rodet, X., and Collette, J. L., 1999: Vibrato: Detection, estimation, extraction, modification. In *Proceedings of the Workshop on Digital Audio Effects (DAFx-99).*

[80] Sarlo, J., 2004: Real-time pitched/unpitched separation of monophonic timbre components. *Proceedings International Computer Music Conference.*

[81] Schroeder, M., 1968: Period histogram and product spectrum: New methods for fundamental-frequency measurement. *The Journal of the Acoustical Society of America*, **43**(4), 829–834.

[82] Schumacher, R. T., 1992: Analysis of aperiodicities in nearly periodic waveforms. *The Journal of the Acoustical Society of America*, **91**(1), 438–451.

[83] Seashore, C. E., 1967: *Psychology of Music.* Dover Publications, Inc.

[84] Serra, X., 1989: *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition.* Ph.D. thesis, Stanford University.

[85] Serra, X., and Bonada, J., 1998: Sound transformations based on the sms high level attributes. *Proceedings of the Digital Audio Effects Workshop (DAFX'98).*

[86] Settel, J., and Lippe, C., 1994: Real-time musical applications using the fft-based resynthesis. *Proceedings International Computer Music Conference.*

[87] Taylor, F. E., 1974: Music and its logic. *The British Journal of Aesthetics*, **14**, 214–230.

[88] Timmers, R., and Desain, P., 2000: Vibrato: Questions and answers from musicians and science. In *Procdeedings International Conference on Music Perception and Cognition.*

[89] Tomkins, C., 1980: *Off the wall: Robert Rauschenberg and the Art World of Our Time*. Doubleday.

[90] Wishart, T., 1994: *Audible Design*. Orpheus the Pantomine, Ltd.

[91] Zicarelli, D., 1998: An extensible real-time signal processing environment for max. In *Proceedings of the International Computer Music Conference*, 463–466.

[92] Zivanovic, M., Roebel, A., and Rodet, X., 2007: Adaptive threshold determination for spectral peak classification. *Proceedings of Ninth International Conference on Digital Audio Effects.*