Fall 12-17-2011

Statistical Spectral Parameter Estimation of Acoustic Signals with Applications to Byzantine Music

Kyriakos Michael Tsiappoutas
University of New Orleans, ktsiappo@uno.edu

Statistical Spectral Parameter Estimation of Acoustic Signals with Applications to Byzantine Music

A Dissertation

Submitted to the Graduate Faculty of the
University of New Orleans
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Engineering and Applied Science
Physics

by

Kyriakos Michael Tsiappoutas

B.S. University of New Orleans, 2002
M.S. University of New Orleans, 2004
M.S. Illinois State University, 2007

December, 2011

Dedication

To my Alexander.

Acknowledgments

My wife, Elisabeta Pana, my son, Alexander, my mother and father, Georgia and Michael Tsiappoutas, who all supported me in more ways than one. My grandfather's brother, Hatzisavvas Psaltis, who taught me the music and instilled the curiosity in me.

My teacher, Professor George E. Ioup, whose courses taught me how to think as a physicist and presence how to carry myself decently. Professor James May who believed in me early on and spent many hours talking to me about things non-physical, yet very relevant to our physical world. Professor Juliette Ioup, Professor Huimin Chen, and Professor Ioannis Georgiou for their commitment to educating me.

The LaTeX community, who put together this remarkable tool used to typeset this dissertation. The Louisiana taxpayers for supporting my scholarship.

Table of Contents

List of Figures

List of Tables

Abstract

Digitized acoustical signals of Byzantine music performed by Iakovos Nafpliotis are used to extract the fundamental frequency of each note of the diatonic scale. These empirical results are then contrasted to the theoretical suggestions and previous empirical findings. Several parametric and non–parametric spectral parameter estimation methods are implemented. These include: (1) Phase vocoder method, (2) McAulay–Quatieri method, (3) Levinson–Durbin algorithm, (4) YIN, (5) Quinn & Fernandes Estimator, (6) Pisarenko Frequency Estimator, (7) *MU*ltiple *SI*gnal *C*haracterization (MUSIC) algorithm, (8) Periodogram method, (9) Quinn & Fernandes Filtered Periodogram, (10) Rife & Vincent Estimator, and (11) the Fourier transform. Algorithm performance was very precise.

The psychophysical aspect of human pitch discrimination is explored. The results of eight (8) psychoacoustical experiments were used to determine the aural just noticeable difference (jnd) in pitch and deduce patterns utilized to customize acceptable performable pitch deviation to the application at hand. These customizations [*Acceptable Performance Difference* (a new measure of frequency differential acceptability), *Perceptual Confidence Intervals* (a new concept of confidence intervals based on psychophysical experiment rather than statistics of performance data), and one based purely on music–theoretical asymphony] are proposed, discussed, and used in interpretation of results.

The results suggest that Nafpliotis' intervals are closer to just intonation than Byzantine theory (with minor exceptions), something not generally found in Thrasivoulos Stanitsas' data. Nafpliotis' perfect fifth is identical to the just intonation, even though he overstretches his octave by fifteen (15) cents. His perfect fourth is also more just, as opposed to Stanitsas' fourth which is directionally opposite. Stanitsas' tendency to exaggerate the major third interval A_4–F_4 is still seen in Nafpliotis, but curbed. This is the only noteworthy departure from just intonation, with Nafpliotis being exactly Chrysanthian (the most exaggerated theoretical suggestion of all) and Stanitsas overstretching it even more than Nafpliotis and Chrysanth. Nafpliotis ascends in the second tetrachord more robustly diatonically than Stanitsas. The results are reported and interpreted within the framework of Acceptable Performance Differences.

Keywords: statistical spectral estimation, fundamental frequency estimation, statistical signal processing, Fourier transform, autocorrelation, autoregression, autocovariance sequence, autoregressive moving average (ARMA) estimation, psychoacoustics, pitch discrimination, just noticeable difference (jnd), Yule–Walker equations, windows, Byzantine, uncertainty principle

Chapter 1

Introduction

1.1 Background

This is not a primer on Byzantine music and therefore only the bare music theory essentials will be provided here as an aid to the reader in understanding the purpose. One of the differences between Byzantine and Western music that is relevant to this dissertation is the completely different nature of scale intervals. It is explained below. For a discussion on the history of Byzantine Music and some technical aspects in English see Wallesz (1961) [88]; in Greek see Chrysanthos (1832) [18], Patriarchal Byzantine Music Committee (1883) [52], Panagiotopoulos (1981) [53]. Tsiappoutas (2004) [85], a masters thesis leading to this dissertation, gives more background on some technical aspects relevant to this research, as well insight into methodology of data collection and analysis, psychoacoustics of pitch discrimination, and general discussion on the topic of comparing theoretical Byzantine music intervals to those extracted empirically.

1.1.1 Western and Byzantine Music Intervals

In the well-tempered scale of Western music theory, frequencies within a scale are allowed to change only by discrete frequency quanta called semitones. Any musical interval is an integer multiple of this semitone. Two semitones make a tone. Three semitones make a tone and a half. For discussion on major and minor Western scales, please see Surmani et al. (2004) [83].

Think of a piano keyboard. Given any white key, the smallest frequency amount by which one chooses to go up or down the musical scale is the semitone. For example, pressing the black key above the white reference key causes the melody to increase in frequency by a semitone.

In Byzantine music there are no such uniformly fixed, equidistant quanta. One can choose to ascend by a fraction of a semitone—or any other variable frequency change for that matter. Clearly, then, a piano cannot play a Byzantine tune. It is possible for a violin to play a Byzantine tune, although Byzantine music is never performed nor accompanied by any other musical instrument but the human voice.

1.1.2 Current Debates

This dissertation's attempt to experimentally quantify Byzantine music intervals and compare them to the suggested theoretical intervals, is at the heart of an ongoing debate between schools of thought: the *traditionalists* and the *progressivists*. The traditionalists believe that true Byzantine music scale intervals are transmitted exclusively and solely though oral tradition specifically from chanters that have been the recipients of formal musical training in the conservatories of the Orthodox Christian Patriarchate of Constantinople. This stringent requirement limits the sample space of properly trained chanters considerably. It is argued that there are but a few remaining chanters able to perform the theoretically proposed musical intervals accurately.

The progressivists, on the other hand, can be further subdivided into two categories: the *"westernizers"* and the followers of *Simon Karas*. The westernizers believe that Byzantine music can be performed with the known well-tempered musical intervals without loss of fidelity. An example of that would be the use of musical instruments like the organ in Greek Orthodox Churches in North America. Simon Karas, on the other hand, is a contemporary music theorist and practitioner who contributed to many genres of traditional Greek and ethnic music. He had no significant formal training in any traditionalist–approved conservatory. According to his theory microtonal intervals exist, but are not the same as the traditional ones. It is worth noting that some of his proposed intervals are not discernible by the human ear—let alone performable by human voice.

1.1.3 Extant Theoretical Literature

The oldest *printed* Byzantine Music book is that of Bishop Chrysanthos of Madytos dating back to 1818, although subsequent editions (Chrysanthos (1832) [18]) are available as reprints in the market today (Koultoura Editions is credited with reprinting many long out–of–print Byzantine Music books). In his book, detailed mathematical accounts of how the musical intervals should be quantified are presented. The different scales that can be theoretically formulated are over 200 and all of them are performable by experienced chanters. The main ones, however, are probably less than ten, each of them employing unique microtonal intervals outside the semitone structure of Western music.

In 1883, the Ecumenical Patriarchal Committee of Byzantine Music matters devised an instrument that could play microtonal scales and refined the theoretical scales based not only on mathematical methods dating back to Pythagoras, but also based on their practical perception of the newly designed instrument which in essence provided a method for them to physically measure string lengths and construct ratios which then were linearized by means of logarithmic transformations (Patriarchal Byzantine Music Committee (1883) [52]).

The Byzantine Music theory bibliography expanded dramatically during the latter part of the 19-th century and later on. There are hundreds of books on the subject, most of them based more or less on the older ones.

In this dissertation we are using the intervals produced by the Patriarchal Committee of 1883.

1.2 Research Objective

The purpose of this research is to empirically quantify the microtonal intervals of Byzantine Music and compare them to theoretical intervals, though review and implementation of a number of classical and modern fundamental frequency tracking methods.

Other sound physical and perceptual characteristics are explored, such as human ability to resolve the acoustical discrepancies between theory and practice, perceptual brightness, and "jaggedness"—spectral irregularity.

1.3 Formulation of Problem

Traditionally (Kinsler et al. (1999) [34]), a continuous-time one-dimensional mechanical wave is represented by

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 y}{\partial t^2}$$

where the constant c^2 is $\frac{T}{\rho_L}$ with T being the tension, ρ_L the linear density of a string vibrating, and x and y indicate displacements on a two–dimensional Cartesian place. By analogy, the solutions to the above partial inhomogeneous differential equation are the same ones for a mechanical acoustic wave traveling in space at some time t and are generally complex $\mathbf{x} = \mathbf{A}e^{i\omega_0 t}$, where $\omega_0 = 2\pi f$ is the initial angular velocity, and $\mathbf{A} = a + ib$ is the complex amplitude.

Since only real–valued functions are practical for acoustical applications, only the real part of the $\mathbf{x}$ will be considered here. In other words,

$$Re\{\mathbf{x}\} = a\cos\omega_0 t - b\sin\omega_0 t$$

reduces to

$$x = A\cos(\omega_0 t + \phi) \tag{1.1}$$

by means of visualizing a phasor $A = \sqrt{a^2 + b^2}$ in magnitude rotating counterclockwise in the complex plane and forming a phase angle $\phi = tan^{-1}(b/a)$ from the (positive) real axis.

Equation (1.1) is central to classical methods of fundamental frequency estimation. It can be modified to represent a more specific acoustical signal which can then be digitized and analyzed by means of some *Fourier Transform* or *Autocorrelation* technique (this aspect will be explored further in Chapter 2).

A complex acoustical tone, i.e., a tone comprising more than one superimposed sinusoid—not complex in the sense of having a real and an imaginary part—will in general behave like a sum of weighted or warped sinusoids. Our purpose in this case is to estimate (instantaneously or track over time) the lowest of those frequencies of the tone, namely, the fundamental frequency f_0.

The fundamental frequency is the physical characteristic of a musical tone that accounts or explains most of what our ears perceive as *pitch*. But frequency is not the only physical aspect that affects how we perceive a tone's pitch. Other variables include how

loud or intense the sound is or how smooth, periodic, transient, (wide-sense) stationary, and deterministic it can be made. Even if all the conditions are experimentally controlled and clinically optimized so that f_0 and pitch have a perfect positive correlation, the human ear will not perceive a tone"outside the existence region" of approximately 20 Hz – 20 kHz (Pressnitzer et al. (2001) [59]). Nevertheless, f_0 seems to be so overwhelmingly the most important predictor of pitch[1], that the two are sometimes used interchangeably—terms like *fundamental frequency estimation or tracking* mean the same as *pitch detection algorithms* (Hess (1983) [29]).

Equation (1.1) can be customized in a manner more suitable for formulating the problem at hand (Beauchamp (2007) [2]). Consider the following modification as a fundamental musical sound model. The signal $s(t)$ is a sum of properly weighted sinusoids with time–varying amplitudes, frequencies, and phases. Some additive noise is incorporated into the model so as to make it more realistic and it too is time-varying, white Gaussian, and zero mean—although it does not have to be.

$$s(t) = \sum_{k=1}^{K(t)} A_k(t) \cos[\theta_k(t)] + n(t) \tag{1.2}$$

where

$$\theta_k(t) = 2\pi \int_0^t f_k(\tau) d\tau + \theta_{k_0} \tag{1.3}$$

and
t = time.
$A_k(t)$ = amplitude of the kth frequency component (partial or sinusoid) at time t.
k = partial number
$K(t)$ = number of sinusoidal partials (integer), which is a time-dependent quantity
$\theta_k(t)$ = phase of partial k at time t.
$f_k(t)$ = frequency of partial k at time t.
$\theta_{k_0} = \theta_k(0)$ = initial phase of partial k (phase at time = 0).
$n(t)$ =additive noise signal, whose short–term spectrum varies with time.

Note that the the initial phase of any given partial k, i.e., $\theta_k(0) = 2\pi \int_0^0 f_k(\tau) d\tau + \theta_{k_0} = \theta_{k_0}$, i.e., since the integral evaluates to zero, the initial phase is independent of the partial frequency, as noted above. Also, the phase derivative *is* the frequency, or $d\theta_k(t)/d\tau = 2\pi\ f_k(t) = \omega_k(t)$, that is to say, the angular frequency of the partial k at time t—a more direct link to Equation (1.1). Consequently, the phase is known for all times if we have knowledge of the initial phase and the frequency at that specific instant in time. This result holds only when the time-frequency scale is not altered; if it is, the phases among the different harmonics will also change.

[1]Many renowned researchers make this claim (that f_0 and pitch goes hand in hand over a wide range of theory and application), but this effect has never been quantified in the sense that "if you consider a number of variables that are known to affect how we perceive pitch, f_0 accounts for n% of it". An interesting psychoacoustics experiment would be to apply predictive modeling to both perceptual and physical variables in a controlled environment so that finally a number of shared variance is associated with f_0 predicting pitch with all other variables held constant.

In some synthesis applications the noise term $n(t)$ is left out due to the need for time expansion or stretching. In this case we assume that the noise is embedded in the amplitude and frequency functions for that particular partial we time-stretched. Incorporating the above into one model we have

$$s(t) = \sum_{k=1}^{K(t)} A_k(t) \cos[2\pi \int_0^t f_k(\tau) d\tau + \theta_{k_0}]. \tag{1.4}$$

and the formulation problem is reduced to estimating the parameters in Equation (1.4), namely, $K(t)$, $A_k(t)$, $f_k(t)$, and θ_{k_0} for $1 \leq k \leq K$.

Chapter 2 presents a number of ways tailored to specifically estimate $f_k(t)$ either tracking it as a function of time, or pinpointing average and/or instantaneous f_0 estimates.

1.4 Contribution of Dissertation

The main contribution of this work is to empirically extract the music scale intervals of traditional Byzantine chant (acoustic signal) through implementation of a collection of established statistical spectral parameter estimation pitch detection algorithms. Secondary contributions include the comparison between empirical and theoretical music intervals and the use of pitch perception literature to determine if those differences are discernible by human ear.

1.5 Data Sample

The music to be analyzed is performed by Iakovos Nafpliotis. The choice of this particular person is not accidental. He is indisputably the most renowned chanter of Byzantine music caught on tape. In this respect, this is the one person whose performance—and hence music intervals—will not be brought into question neither by the traditionalist nor by the progressivists. Since the choice of the performer is crucial to the generalization of the results, a signal sampling method must be devised to capture these intervals without resorting to new recordings. Both of these are discussed below.

1.5.1 The Chanter

Iakovos Nafpliotis (1864–1942) was born on the island of Naxos, Greece and moved to Istanbul, Turkey at the age of seven. He quickly distinguished himself as a music prodigy on accounts of his musical memory and voice. By the age of 14 (after having served as a student in St. Nikolaos church) he was extended an offer and ordained as a Canonarches of the first order in the official cathedral of the Ecumenical Patriarchate of Constantinople, the center of Orthodox Christianity. He served in the Patriarchal Church for 60 years under 14 Patriarchs[2], assuming every position within the hierarchy up to the ultimate title of *First*

[2]A Patriarch in the Eastern Christian Church is the equivalent of a Pope in Western Roman Catholic Church. As a matter of fact, the two titles were one and the same up until *The Great Schism of 1054*.

Arch–chanter of the Great Church of Christ[3]. After his retirement, since he was not a citizen of Turkey, he returned to Athens, Greece where he died at the age of 78.

Probably what makes Iakovos indisputably the golden standard of Byzantine music chant is the fact that he was taught the music from teachers who had knowledge of the old paleographic system of Byzantine music notation which was largely committed to memory. Iakovos himself is said to have received partial training on old notation, which took up to 20 years of apprenticeship. This link of the old and new along with the fact that he has always served in the Patriarchal Church—where by doctrine no music other than the approved was chanted—would give Iakovos a clear advantage over other chanters. This must be why even progressivists reference Iakovos' performances.

1.5.2 The Recordings

In 2008, Professor Antonios E. Alygizakis released five (5) Compact Disks (Alygizakis (2008) [1]) accompanied by a monograph summarizing his research on Iakovos Nafpliotis' legendary recordings. This monumental audio remastering is a result of a nearly two decades of working with the original 78 RPM phonograph records. The sound fidelity and quality of these Compact Disks is far better than that of the tapes that have been circulating in Byzantine music circles.

The original vinyl records were recorded during the period of 1914–1926 by a German–based music production company named *Blumental Record and Talking Machine—Orfeon Record* later known as *Odeon Records*.

1.5.3 The Signal Sampling Method

This section deals with the method of sampling snippets of sound from an already digitized signal, as described in Subsection (1.5.2). It does not describe the analogue–to–digital conversion, in which the term "sampling" has a different meaning (in the usual digital signal processing sense). The sampling frequency for the already digitized signal (Alygizakis (2008) [1]) is 44,100 Hz.

The sampling method is simple. First a music piece in the scale of interest is chosen. Ideally, it is one that the master chanter is chanting alone, without other accompanying voices. Each time the voice passes through a note of the scale, that snippet of sound is kept. A piece can yield between twenty (20) and forty (40) such snippets which are then concatenated to produce a signal of the same note, anywhere from one to two seconds long. This concatenated acoustical signal is what is fed through the algorithms presented in Chapter 2.

Of course, the sampling is not random, but with the scarcity of data one can hardly expect that our sample space for each note is expansive enough to afford the luxury of randomly selecting tone snippets. The time between the snippets in the actual music piece is not a constant, but this is irrelevant for our purposes.

[3]Translation from Greek " Άρχον Πρωτοψάλτης της Μεγάλης του Χριστού Εκκλησίας."

The issue of human perception in choosing which snippets are which tone is called into question. It would be better, one may think, to have the machine choose them and categorize them into "tone categories." But for this to be done, we first need to know the frequency of each tone, and this is the topic of this dissertation. There is no real answer for the issue of human perception in the data sampling, other than the ability of the data collector to distinguish acoustically which tone is which and, of course, his/her scientific integrity.

1.6 Frequencies, Cents, and Atoms

Throughout this dissertation frequency estimates will be given as either pure *frequencies* (in Hertz or radians/second), *cents* (a well known measure of frequency differential), or *atoms* (a measure closely related to cents, but used in Byzantine Music literature). A discussion on how to practically go from one to another is warranted.

A doubling in frequency creates the perception of the same tone being one octave higher. In a well-tempered scaled, this frequency doubling is divided into twelve (12) intervals, called semitones. A chromatic scale[4] is one for which all semitones enclosed by the upper and lower frequencies of a scale are played in progression. Or one could combine some of the semitones to create tones and progress in such a fashion that a diatonic scale[5] is created. A diatonic scale, then, could be constructed as an upward and then downward progression of T—T—S—T—T—T—S, where a T denotes a tone and an S denotes a semitone. For example, let's denote the tone C_4 as our initial frequency f_0. Then we have the mapping $C_4 \longmapsto f_0, D_4 \longmapsto f_1, E_4 \longmapsto f_2, F_4 \longmapsto f_3, G_4 \longmapsto f_4, A_4 \longmapsto f_5, B_4 \longmapsto f_6, C_5 \longmapsto f_7$. Let f_0 be the *fundamental frequency*. In a well-tempered scale, then, a semitone would advance the fundamental by a factor of $\sqrt[12]{2}$, i.e., by about 1.059463094, or 6%.

We, humans, are not thinking in terms of frequencies, however; we think in terms of logarithms of frequencies. We seem to have the ability to distinguish between high and low frequencies and we tend to think of them as high and low as well[6]. Western music scores can be thought of as a plot with a logarithmic y-axis and time on the x-axis. While frequencies are multiplicative, log frequencies are additive, and here hinges their great perceptual advantage. It is more convenient to visualize the frequency space—f_0 to f_7—as being divided in 1200 equal parts, each semitone enclosing 100 of those parts rather than thinking of a $\sqrt[12]{2}$ factor which if raised to the 12^{th} power gives a $2:1$ frequency ratio. When the octave is divided into 1200 parts those parts are called *cents*; when divided into 72 (or, sometimes 68) parts we call them *atoms*[7].

[4]This is the definition of a chromatic scale in Western music. In Byzantine music the definition is dependent upon the different microtonal intervals within the two bounding frequencies, which are well-defined by theory. The Byzantine chromatic scale will not be used in this dissertation.

[5]As with the chromatic scale, the term diatonic has a different meaning in Byzantine than in Western music. Here we are using *diatonic* in its Western sense.

[6]This is not the case with hue, for example. We never call one hue shorter or longer than another one, even though wavelength is the corresponding physical aspect of the perceived hue.

[7]In Greek, it is called *morion* (<GR μόριον) literally meaning "molecule." But *atom* is adopted here since the Greek term implies a particle that cannot be further divided.

The following should come handy when transforming between frequencies, cents, and atoms:

$$\Delta parts = \gamma \ log_2 \left(\frac{f_1}{f_0} \right)$$
$$\underbrace{\frac{\Delta parts}{\gamma}}_{\alpha} = \underbrace{log_2 \left(\frac{f_1}{f_0} \right)}_{\beta} \tag{1.5}$$

where, if parts is cents, then $\gamma = 1200$ and if parts is atoms, then $\gamma = 72$. Note that the logarithm is to the base two to denote the doubling of frequency[8].

1.7 Organization of Dissertation

The rest of the dissertation is organized as follows. Chapter 2 presents the mathematical theory of a number of frequency estimation algorithms along with mathematical formulations of some interesting psychoacoustical phenomena. Chapter 3 provides an abridged account of psychoacoustics in regards to human ability to resolve frequencies. Chapter 4 tabulates the results of the algorithms implemented in Chapter 2, provides discussion, and concludes this dissertation.

[8]Most scientific calculators do not offer the logarithm to the base two function, but almost all have the natural logarithm function (to the base e). A useful trick is to think of the above exponent as $2^\alpha = \beta$, take natural logarithms of both sides so that $\Delta parts = \gamma \ ln\beta / ln2$.

Chapter 2

Algorithms and Methods

2.1 Introduction

This chapter provides the theory behind the algorithms used to both track the fundamental frequency and give physical insight into the psychophysical aspect of the analysis.

The theory herein is kept to a minimum; enough mathematics are given to emphasize the concepts. A conscious effort has been made to keep the mathematics simple and intuitive. Following Hamming's

"The Purpose Of Computing Is Insight, Not Numbers"

this chapter's motto would be

"The Purpose Of Algorithms Is Insight, Not Mathematics".

Even though the material is presented as coherently as possible, please keep in mind that it originated from diverse sources. Notation is oftentimes kept as in the original papers, but when bits and pieces are put together to make a point within a context, notation may be modified. Sometimes material from an article is omitted—if it does not enhance our knowledge, or if it is too cumbersome mathematically—and sometimes equations outside the scope of the article but relevant to our subject are inserted to solidify understanding. Since part of the contribution of this dissertation is an overview of concepts tailored to the singing voice and no original algorithm was architected, all of this customization is to a point necessary. The author is trying to take an authoritative look into the algorithms—to the best of his ability—and the material is sometimes presented within a lens of constructive criticism. This approach should not be taken as negative criticism towards any of the algorithms, but merely as an attempt to illustrate the algorithms' usability and practicality for the problem at hand and maybe demonstrate the author's effort towards understanding the algorithm more deeply. However, mostly positive criticism will be encountered, because if the algorithm was relevant and suitable enough to be included here, a bias has already been realized.

The chapter starts off with the Fourier transform and quickly moves to the two algorithms implemented by the same open source code: phase vocoder and McAulay–Quatieri methods. A section devoted to general historical and current findings on fundamental frequency (pitch) detection is followed by a number of interesting approaches to analyzing the singing voice beyond frequency tracking. The chapter concludes with a brief outline of another twelve (12) frequency estimators.

2.2 The Fourier Transform

The Fourier transform is at the heart of many algorithms on frequency tracking and estimation, primarily due to its core property of enabling us to go from the time domain to the frequency domain.

Fourier transformation happens in pairs, in such a way that an original function in the time—or space[1]—domain can be transformed to another function whose independent variable is frequency. Traditionally (Bracewell (2000) [12]), the Fourier transform pair is denoted by

$$f(x) \supset F(s)$$

where $f(x)$ is the original function in x and $F(s)$ is the transform of the original function in the frequency variable s. The above symbolism is made more concrete by the continuous Fourier integral transform definition

$$\begin{aligned} F(s) &= \int_{-\infty}^{+\infty} f(x) e^{-i2\pi xs}\, dx \\ f(x) &= \int_{-\infty}^{+\infty} F(s) e^{+i2\pi xs}\, ds \end{aligned} \tag{2.1}$$

which conveys the reversibility of the transformation, i.e., the $+i$ transform of the $-i$ transform is the original function[2]. There is much to be said about the theoretical properties of the Fourier transform, so much that whole books have been and continue to be devoted to its remarkable powers as a fundamental analysis tool, but there are outside the scope of this dissertation. Two classic and, in my humble opinion, unsurpassed sources on the theory and application of the Fourier transform are Bracewell (2000) [12] and Papoulis (1962) [54]; an excellent more recent publication that emphasizes more the practical applications of it is Lyons (2009) [39].

Throughout this dissertation, only digitals signals are analyzed, not their analogue counterpart. The *Discrete Fourier Transform*[3] or DFT

$$\begin{aligned} F(\nu) &= \sum_{\tau=-N/2}^{N/2-1} f(\tau\Delta t) e^{-i2\pi(\tau\Delta t)(\nu\Delta f)} \Delta t \\ f(\tau) &= \sum_{\nu=-N/2}^{N/2-1} F(\nu\Delta f) e^{+i2\pi(\tau\Delta t)(\nu\Delta f)} \Delta f \end{aligned} \tag{2.2}$$

[1] Traditionally, engineers use time as the independent variable in the domain that the signal originally was sampled from and, even though this convention fits the needs of this dissertation perfectly, here we will adopt the one–dimensional space notation, i.e., the x–axis. This is a more general case easily extended to accommodate not only time but planes like pictures, for example.

[2] Note that with this notation reversibility is achieved no matter if $f(x)$ is even $[f(x) = f(-x)]$ or odd $[f(x) = -f(-x)]$.

[3] Not to be confused with the *Discrete–Time Fourier Transform* or DTFT, which is not a replication of one period but a transformation of the entire time–series (Lyons (2009) [39]).

is the one used here in lieu of the continuous transform due to the digital nature of modern computers. By using this slightly modified version of the DFT given in Bracewell (2000) [12], the frequency in Hertz would be $\frac{\nu}{N\Delta t}$ for $-\frac{N}{2} \leq \nu < +\frac{N}{2}$, where Δt is the sampling interval, Δf is the frequency resolution, and $\Delta t \Delta f = \frac{1}{N}$. This last term could be used to simplify the exponent of Equation (2.2) by replacing $\Delta t \Delta f$ by $\frac{1}{N}$. Please notice the periodic behavior of $f(\tau)$ and $F(\nu)$ along their respective independent variables τ and ν, which is at the heart of the DFT concept just like other operations like the autocorrelation and cyclic convolution. The DFT is implemented in commercial software like MATLAB®[4] by means of a *Fast Fourier Transform* algorithm which greatly reduces the number of flops[5] by making use of symmetries and cyclic properties of the DFT (Bracewell (2000) [12]).

2.3 Phase Vocoder—Harmonic Filter Analysis

Following Beauchamp's work (Beauchamp (2007) [2], Beauchamp (1975) [5], Beauchamp (1993) [6], and Beauchamp, Maher, and Brown (1993) [7]), a phase vocoder can be visualized as a series of band–pass filters, each allowing a sinusoid of a certain frequency to go through, with each sinusoid a multiple integer of the lowest frequency. This lowest frequency is usually the fundamental, but not necessarily. It can be any frequency. Only, if it is too far from the fundamental, the phase vocoder idea will not hold. Let's call this lowest frequency the *analysis frequency* or f_a. Then, since the upper harmonics (or partials[6]) are multiple integers of the basis, analysis frequency, a well–behaved tone (like the ones used in this dissertation) should be able to pass through this model without loss of its general characteristics. For this to happen, however, not only the signal has to be well–behaved in the sense that it possesses nice, clear partials of the form $f_k = kf_a$ for $k = 1, ..., K$[7], it does not vary too much in time, has decent signal–to–noise ratio, and its noise is independent and identically distributed and normal with a small standard deviation clustered around the mean, but f_a must be chosen so that it is as close as possible to the empirical f_0[8], otherwise this asymphony between the two will render the vocoder of little practical use.

Each filter is basically a window $W_k(f - f_k)$ and its maximum is a unit vector at the center, i.e., $f = kf_a$, and it goes to zero away from the center for $f \leq (k-1)f_a$ and $f \geq (k+1)f_a$, like a Gaussian. If the input to the model is periodic and has $f_0 = f_a$ and fixed partial amplitudes A_k, the output would be a sinusoid

$$s_k(t) = A_k \cos[2\pi k f_a t + \theta_{k_0}] \tag{2.3}$$

with frequencies at $f_k = kf_a$ and amplitude A_k which is an idealized case of Equation (1.4), i.e., constant parameters.

[4] MATLAB® is a registered trademark of MathWorks, Inc.

[5] One flop (*fl*oating point *op*eration) is equivalent to one complex multiplication and one complex addition; flops (*fl*oating point *op*eration per *s*econd) is number of complex multiplications and complex additions per second.

[6] The terms *harmonic* and *partial* are sometimes used interchangeably, but strictly speaking the first harmonic is the second partial.

[7] $K - 1$ is the number of harmonics, or, equivalently, K is the number of partials.

[8] This is one of the reasons the Fourier Transform of the signal is taken first, so that the f_0 obtained there can be used in the phase vocoder.

2.3.1 Frequency Deviation

In practice, however, amplitudes and frequencies do vary with time. Equation (1.4) accommodates this given that the frequencies of the partials are not that far away from the model partials kf_a and that the amplitudes are confined also within a close range of its median. To have a measure of how closely the phase vocoder models reality we can define a frequency deviation as

$$\Delta f_k(t) = f_k(t) - kf_a \tag{2.4}$$

which can be made relative to the partial number k

$$\frac{\Delta f_k(t)}{k} = \frac{f_k(t)}{k} - f_a \tag{2.5}$$

or even normalize it by the frequency analysis like

$$\frac{\Delta f_k(t)}{f_a k} = \frac{f_k(t)}{f_a k} - 1. \tag{2.6}$$

These equations come in handy for quick numerical checks that can enhance conceptual understanding. For example, each partial frequency can be thought of as whatever the model analysis frequency output is at any given bin or time adjusted by the frequency deviation

$$f_k(t) = kf_a + \Delta f_k(t), \tag{2.7}$$

or use the normalized frequency deviation formula to see by how much the frequency of a given partial varies about its model–predicted "ideal" value as a fractional deviation and express that in cents or atoms. For example, since we have seen on page 8 that a semitone is about a 6% change in frequency, if the $\Delta f_k / kf_a$ is about ± 0.06 then we know that the frequency of the kth partial fluctuates about its central frequency kf_a by about a semitone. Using Equation (1.5) we can express this fractional deviation in cents or atoms as a function of time (or instantaneously) as

$$\Delta cents(t) = \gamma \cdot log_2 \left(\frac{\Delta f_k(t)}{kf_a} \right). \tag{2.8}$$

Filter bank analysis or phase vocoder is not like a Monte Carlo simulation where data are sometimes produced and then compared to empirical results or subjected to statistical requirements to yield a predetermined dataset. The vocoder is an idea, a construct that signals are fed through and then automatically compared to the model. This theory–to–practice comparison is central to this dissertation. The same idea is used with music theory and empirical practice. A set of metrics is then constructed to gauge how much practice matches the theory, or how well the ideal situation models the data. Another such metric is *inharmonicity*.

2.3.2 Inharmonicity

If all partials track one another perfectly in integer multiples such that

$$\Delta f_k(t) = k\Delta f_1(t) \tag{2.9}$$

a tone is harmonic at each instant of time. A sound is then said to be inharmonic if *inharmonicity*

$$I_k(t) = \frac{\Delta f_k(t)}{k\Delta f_1(t)} - 1, \tag{2.10}$$

deviates from zero, with larger numbers in general giving larger inharmonicity. In practice, though, the first partial is not always the one possessing most of the energy or having a very prominent amplitude, which could lead to a poor inharmonicity estimate. Define the "relative–amplitude–weighted sum of the harmonic–normalized first five harmonic frequency deviations",

$$\Delta f_{c_1}(t) = \frac{\sum_{k=1}^{5} A_k(t)\Delta f_k(t)/k}{\sum_{k=1}^{5} A_k(t)} \tag{2.11}$$

which is based on two experimentally validated facts (Moore, Glasberg, and Peters (1985) [47]): (1) In most musical sounds the first five partials are the stronger ones, i.e., the first five harmonics interfere in such a way that a weighted average (weighted based on experimental knowledge, not theory) of them is what usually determines our perception of pitch, and (2) the structure of this relative dominance of the lower five harmonics is known at least collectively from empirical data.

A closer look into Equation (2.11) will help us point out one fundamental limitation of the harmonic filter bank analysis method. First imagine that all A_k=constant $\forall\ t$ and $k = 1, ..., K$, that is to say, each partial has a constant amplitude and also amplitudes are equal to each other. Then Equation (2.11) reduces to a straight average since there are no amplitude weights. If we let the amplitudes take on different values in time so that are not equal to each other, then the stronger amplitudes will be weighted accordingly. Further assume that the Δf_1 is taken to be large with respect to the central f_a. This will cause $k\Delta f_1$ to deviate further from the respective harmonic analysis frequency, kf_a, or the kth bin. Now, a frequency component is input into the model, and we are faced with the decision of whether we should slot this experimental f_k into the kth or the $(k+1)$st bin. We can devise a rule, that is really statistical in nature, even though it is not spelled out or viewed that way within Beauchamp's framework, and decide that the frequency will belong to the $(k+1)$st bin, if that one "datum" frequency deviation of the first partial but in the kth harmonic is greater than half the analysis frequency, i.e., decide $(k+1)$st bin if

$$k\Delta f_1 \geq 0.5\ f_a \tag{2.12}$$

or the kth bin otherwise. This argument could be developed into a more formal statistical statement of binary hypothesis testing, but it is not necessary here. Please refer to Kay (1998) [31] for an excellent discussion on the issue of statistically setting up decision rules for hypothesis testing. What *is* important to note is that even though a "false" decision has been made due to the large frequency deviation, the output of the kth bin will also include

the effects of the $(k-1)$st partial. This means that even small deviations will affect upper harmonics (in our case, so high that it does not pose a problem). With large experimental deviation from the central ideal analysis frequency, however, even the lower harmonics could be adversely affected when it comes to estimation accuracy. This is a fundamental limitation of the phase vocoder method that is in fact of little or no consequence to our purposes, due to the well–behaved nature of the acoustical signal.

2.3.3 Precision and Accuracy

Precision refers to variability in the data sample; the lower the variance in the data the higher the precision. Accuracy[9] refers to the distance between the sample mean and a standard, be it the population mean or that note A_4 should be tuned to 440 Hz.

The sample collection methodology of concatenating signal snippets really takes care of both of these issues, which are especially central to the robustness of the phase vocoder (and most other algorithms). We could think of $k\Delta f_1$ as a measure of precision and how close is kf_a to f_k as a measure of accuracy. The fact that the snippets are all of the same tone is really an attempt to minimize the spectral variability of $k\Delta f_1$, that is to say, keep the frequencies in the spectrum close to each other and hopefully clustered closely around its mean value, which optimally would be the "true" frequency of the tone. This is relevant to the precision, reproducibility, or internal validity of the data. Now, how much that precise central frequency is close to the true standard is a question of accuracy. The issue of accuracy is addressed in three ways: again through the sample collection methodology, through the use of the Fourier transform (and later on with other methods) to establish what the f_a should be for the phase vocoder, and through common knowledge of both the absolute frequencies of each tone itself or relative fractional deviations (or microtonal representation in cents and atoms) within the musical scale boundary values of doubling the frequency to achieve the perception of octave and also knowledge of the segmentation of this range, i.e., where the intervals fall.

Therefore, the method of concatenating snippets of sound is responsible for most of the good nature of the data sample. If instead a whole piece was fed into this particular model (or most models presented in this dissertation for that matter), we would have to rely on the machine to distinguish where one tone stops and where the next one begins, and the machine was not really trained to do this task in any meaningful or reliable way. This goes back to the discussion in section (1.5.3) and how human perception is involved in collecting the sample. This philosophical discussion will be omitted[10].

2.3.4 Heterodyne–Filter Analysis Method

An implementation of the harmonic filter bank is the heterodyne–filter analyzer (Beauchamp (1966) [3], Beauchamp (1969) [4]) which is rooted in classic Fourier analysis.

[9] Accuracy is really the *Effect Size*, i.e., $(\mu_{sample} - \mu_{standard})/\sigma_{sample}$, even though literature does not like to associate the two. The effect size could be normalized in units of standard deviation (because the units of variance are the sample units squared) or it could be just the straight mean difference.

[10] This issue could be a paper in its own right. Machine learning and artificial intelligence is a very interdisciplinary topic that touches at least physics, psychology, and engineering.

Let us think what operations need to be performed on the signal $s(t)$. Each harmonic component of the vocoder of $s(t)$ has to be shifted to zero along with all frequencies around this component, i.e., kf_a along with any other frequencies must be centered at $f = 0$. This is an attempt to "align" the real frequencies in the signal to the frequencies of the phase vocoder. This is done using the first partial of $s(t)$. By how much each signal upper partial will be "out of alignment" depends on how inharmonic that signal upper partial is relative to its respective vocoder partial. Then any frequencies above $f_a/2$ must be filtered out. The shift operation is achieved via a customized Fourier transform and the filtering via convolution with a low–pass filter which can be any of the known window functions.

Symbolically, the heterodyne (multiplication) operation

$$\check{s}_k(t) = e^{-i2\pi k f_a t} s(t)$$

of the signal with a complex exponential shifts the frequency components on and around zero and convolution of the heterodyned signal with a window $w(t)$

$$\tilde{c}_k(t) = w(t) * \check{s}_k(t),$$

with this symbol "$*$" denoting convolution, achieves the low-pass filtering. We also know that, by the convolution theorem (Bracewell (2000) [12]), convolution in the time domain is equivalent to multiplication in the Fourier domain and the above is equivalent to

$$\tilde{C}_k(f) = W(f)\check{S}_k(f)$$

where functions in capital letters denote Fourier transforms of the corresponding lower letter functions; in other words, $s(t)$ and $w(t)$ are the signal and the impulse response of the low–pass filter and $S(f)$ and $W(f)$ are the signal spectrum and the frequency response, respectively.

By design only frequencies below $f_a/2$ pass through the filter which means that the frequency range is $\pm\frac{1}{2}f_a$ about the analysis frequency—not only for the fundamental frequency but for each partial. Since $S(f)$ takes on values only within the $\pm\frac{1}{2}f_a$ band about kf_a, $\tilde{C}_k(f)$ is the equivalent of a symmetric band–pass filter around kf_a. We can rewrite the above equation to reflect this symmetry as

$$\tilde{C}_k(f) = W(f)\check{S}_k(f) = W(f)S(f + kf_a). \tag{2.13}$$

It should be clear that the heterodyne–filter analysis method is nothing more than taking a windowed Fourier transform of the signal $s(t)$, as given in Equation (1.4), customized for the kth harmonic and using the analysis frequency in the complex exponential, because this is what the idea of phase vocoder is modeling. More explicitly, the complex amplitude of the kth partial of the input signal $s(t)$ is

$$\tilde{c}_k(f) = \int_{-\infty}^{\infty} w(t-\tau)s(\tau)e^{i2\pi k f_a \tau} d\tau. \tag{2.14}$$

Since windowing is an important aspect of the heterodyne–filter method, a brief account of four window options is presented below.

2.3.4.1 Window Functions

Windows give a "window" into the signal, a glance on a portion of the signal since part of what windows do is cut off any information outside a range. If we are in time domain, for example, anything outside a time interval $2T$ can be brought to nothingness by a multiplicative window, essentially a rule that says all but when $|t| \geq T$ must be zero. Cutting up a piece of signal to work with is not all that windows do for us. Used in convolution, they can also smooth functions around the center of the window (and therefore signal) or multiplicatively bring the signal down to zero[11] nicely without creating any jump discontinuities which in turn create spectral leakage problems and scalloping loss (Lyons (2009) [39]).

Typically, the absolute value of the frequency responses of the windows of interest, or magnitude responses $|W(f)|$, is plotted across frequency to reveal information about the window's performance. With such a plot the main lobe and the sidelobes can be seen, but more visual detail is possible if the magnitude responses are plotted on a logarithmic (decibel) scale. Therefore, the *power* (or energy)

$$|W_{dB}(f)|^2 = 20 \cdot log_{10}\left[\frac{|W(f)|}{|W(0)|}\right] \tag{2.15}$$

is most commonly used to show the spectral energy, where it can be seen that each window's plot is normalized so that its main lobe peak is zero decibels. Windows are even functions of time, i.e., $w(t) = w(-t)$, which makes them symmetric both in the time and frequency domains (which is why only the positive axis is usually plotted in conventional magnitude response plots); their energy phase responses are also zero and their Fourier transforms are real.

The width of the main lobe and the height of the sidelobes provide information on the frequency resolution vs spectral leakage trade–off. Narrower main lobe width indicates better resolution in frequency than wider ones and shorter sidelobes indicate lower spectral leakage than higher ones. A host of windows have been proposed to fit the digital filter design needs of various applications. Excellent sources are Harris (1978) [27] and Nuttall (1981) [51], but here we follow Beauchamp's discussion (Beauchamp (2007) [2]) on the basic windows employed in the heterodyne–filter method of analysis. Section (2.4.2) takes a closer look into window design and time–frequency trade–offs.

The *rectangular window* is nothing more than multiplication of the entire signal by a normalized height train of impulses within a range and zeros outside that range. Its area is one, so to customize it to our case at hand we define its height to be f_a and its base $1/f_a$, i.e.,

$$w(t) = \begin{cases} f_a, & |t| \leq 0.5/f_a \\ 0, & |t| > 0.5/f_a \end{cases}, \tag{2.16}$$

[11] Actually, it doesn't have to be zero. If the amplitude value of the first and last DFT point is the same *low* value, the spectral sidelobes will be minimized.

which effectively retains one fundamental period $T_a = 1/f_a$. The rectangular window is probably the worst choice at least for our purposes. Its sidelobes roll off very slowly and therefore excessive leakage is left to deal with. Its response is inferior to other useful windows for $f > f_a$. However, it is worth noting that it provides the best frequency resolution.

A better choice is the *von Hann*[12] *window* with a smoother bell–shaped curve that gives practically no discontinuity at the ends of the sampling interval (it brings the signal down to zero and has a zero first derivative there also). Its resolution is no better than that of the rectangular window, but then again, none of the non–rectangular windows' frequency resolutions are—in fact, non–rectangular windows degrade the windowed DFT resolution by about a factor of two. The von Hann window is

$$\frac{w(t)}{f_a} = \begin{cases} cos^2(0.5\pi t f_a) = 0.5 + 0.5\ cos(\pi t f_a), & |t| \le 1/f_a \\ 0, & |t| > 1/f_a \end{cases} . \tag{2.17}$$

The *Hamming window* looks like a von Hann window only its peak is shorter and its tails do not bring the signal to a halt, but are raised above zero. Its two terms look like this

$$\frac{w(t)}{f_a} = \begin{cases} 0.5 + 0.426\ cos(\pi t f_a), & |t| \le 1/f_a \\ 0, & |t| > 1/f_a \end{cases} . \tag{2.18}$$

The *Blackman–Harris window* is a 4–term more sophisticated window option with window width $4/f_a$ and a peak amplitude at $0.6969 f_a$:

$$\frac{w(t)}{f_a} = \begin{cases} .25 + .3403cos(.5\pi t f_a) + .0985cos(\pi t f_a) + .0081cos(1.5\pi t f_a), & |t| \le 2/f_a \\ 0, & |t| > 2/f_a \end{cases} . \tag{2.19}$$

All the above windows can be put in one general form as

$$\frac{w(t)}{f_a} = \begin{cases} \sum_{p=0}^{P-1} \alpha_p\ cos(2\pi p f_a t/P), & |t| \le \frac{P}{2f_a} \\ 0, & |t| > 0 \end{cases} \tag{2.20}$$

where P is the number of terms in the window[13] and $\alpha_0 = 1/P$. We also keep in mind that the window functions are normalized by f_a and all areas under the curve are normalized to one. A general expression for the frequency responses of the above windows can be obtained by direct Fourier transformation of the general expression in Equation (2.20), i.e,

$$\begin{aligned} \frac{W(f)}{f_a} &= \int_{-\infty}^{+\infty} w(\tau) e^{-i2\pi f t} d\tau \\ &= \sum_{p=0}^{P-1} \alpha_p \int_{-\frac{P}{2f_a}}^{\frac{P}{2f_a}} cos(2\pi p f_a \tau/P) e^{i2\pi f \tau} d\tau \end{aligned} \tag{2.21}$$

[12]Named after Julius von Hann and erroneously oftentimes referred to as the "Hanning window."

[13]For the rectangular window $P = 1$, for the Hamming and von Hann windows $P = 2$, and for the Blackman–Harris window $P = 4$.

which is again normalized by f_a. The above Fourier transform can be written in terms of $sinc(\pi x)$ functions, where $sinc(\pi x) = \frac{sin(\pi x)}{\pi x}$, as

$$\frac{W(f)}{f_a} = \frac{P}{2} \sum_{p=0}^{P-1} \alpha_p \left\{ sinc\Big[\pi\Big(\frac{Pf}{f_a} + p\Big)\Big] + sinc\Big[\pi\Big(\frac{Pf}{f_a} - p\Big)\Big] \right\}. \tag{2.22}$$

Readers familiar with the classic papers of Harris (1978) [27] and Nuttall (1981) [51] will notice that the coefficients of the Hamming and Blackman–Harris windows do not add up to unity, as it is the usual practice. Instead they are expressed as a percent of the analysis frequency. In the case of the Hamming window in Equation (2.18), for example, the peak amplitude (which is the same as the central ordinate mentioned in Bracewell (2000) [12], since the windows are all centered about the origin) is $0.926 f_a$. The areas under the windows in both domains are still all unity nevertheless. This discrepancy is due to Beauchamp's (2007) [2] choice to make the width and the peak of the windows a function of f_a. Here we adopt Beauchamp's convention, since this section is devoted to phase vocoder, but some basic notes on the equivalence of the two approaches (normalized versus traditional windows) are provided below.

To show that the area under the windows and their response curves are equal to unity, we borrow the idea of *equivalent width* from Bracewell (2000) [12], where in page 167 he notes "The equivalent width of a function is equal to the reciprocal of the equivalent width of its transform," i.e.,

$$\frac{\int w(t)dt}{w(0)} = \frac{W(0)}{\int W(f)df}, \tag{2.23}$$

where the limits of integration in our case will be the actual window limits. If we manage to show that the central ordinates of the two domains are equal, then their integrals must be equal. Evaluating the area under the window curve in one domain, say time, will then tell us what the area under the curve is in the frequency domain.

First we show that the central ordinates in the time and frequency domain are equal. Directly from Equation (2.20) and Equation (2.22) for a 2-term window we obtain

$$\frac{w(0)}{f_a} = \alpha_0 cos(0) + \alpha_1 cos(0) = \alpha_0 + \alpha_1 \tag{2.24}$$

and

$$\frac{W(0)}{f_a} = \alpha_0 sinc(0) + \alpha_1 sinc(0) = \alpha_0 + \alpha_1 \tag{2.25}$$

since $cos(0) = sinc(0) = 1$, and in general the result for two terms

$$\frac{w(0)}{f_a} = \frac{W(0)}{f_a} = \alpha_0 + \alpha_1$$

will hold for many terms.

The area under the time window curve, again for a 2–term window, is

$$\begin{aligned}\frac{1}{f_a}\int_{-\frac{1}{f_a}}^{+\frac{1}{f_a}} w(t)\ dt &= \int_{-\frac{1}{f_a}}^{+\frac{1}{f_a}} \alpha_0\ dt + \int_{-\frac{1}{f_a}}^{+\frac{1}{f_a}} \alpha_1 cos(\pi f_a t)\ dt \\ &= \alpha_0\ t\Big|_{-\frac{1}{f_a}}^{+\frac{1}{f_a}} + \frac{\alpha_1}{\pi f_a}\ sin(\pi f_a t)\Big|_{-\frac{1}{f_a}}^{+\frac{1}{f_a}} \\ &= \left[\frac{\alpha_0}{f_a} - \Big(-\frac{\alpha_0}{f_a}\Big)\right] + \frac{\alpha_1}{\pi f_a}\left[sin\Big(\frac{\pi f_a}{f_a}\Big) - sin\Big(-\frac{\pi f_a}{f_a}\Big)\right]\end{aligned}$$

and since $sin\pi = 0$, the second term will always be zero. The same will be the case for all the terms past the first one for the more general cases of windows with many terms. Finally,

$$\frac{\alpha_0}{f_a} + \frac{\alpha_0}{f_a} = \frac{2\alpha_0}{f_a} = \frac{w(t)}{f_a}$$

and since for this particular 2–term $\alpha_0 = \frac{1}{P}$,

$$\frac{w(t)}{f_a} = \frac{2}{2f_a} \rightarrow w(t) = \frac{f_a}{f_a} = 1.0.$$

This is the case for any n–term window due to our choice of $\alpha_0 = \frac{1}{P}$ and because all sinusoids in the time domain vanish. By Equation (2.23) the area under the response function is also unity.

2.3.4.2 Harmonic Corruption & Window Limits

The brief discussion on windows above leads to the concepts of harmonic corruption and harmonic limits of windows. In what follows notation has been changed considerably from the source.

As we said before, when $f_0 = f_a$ the empirical and theoretical are completely aligned and the phase vocoder will yield perfectly accurate (and hopefully precise) frequency estimates. All the unwanted frequency components will be rejected by the model and none will make it into the output spectrum. When $f_0 \neq f_a$, each harmonic will be shifted in frequency by the frequency difference Δf and this shift will cause the vocoder to be less than perfect in canceling out unwanted frequency components. We need a measure of how the model rejects unwanted components.

Let $\Delta f = f_0 - f_a$. Then the analytical and fundamental frequencies of the partial k are $f_k^i = k\Delta f$ out of tune, where the index $i = -1, 0, 1$ will denote the position of the frequency component in question (0), and its immediate two neighboring partials ("-1" indicates the one below it and "$+1$" the one above it) for that particular k partial. So, if f_a is detuned from f_0 by Δf we can say that the amount that the kth harmonic deviates from

f_0 for the wanted component and the two immediately adjacent undesired components is

$$\begin{aligned} f_k^0 &= k\Delta f \\ f_k^{-1} &= (k-1)(f_a + \Delta f) \\ f_k^{+1} &= (k+1)(f_a + \Delta f) \end{aligned} \tag{2.26}$$

It should be intuitively clear that if the two immediately adjacent undesired components f_k^{-1} and f_k^{+1} are not removed (again, due to large Δf) they will corrupt the accuracy of the estimation of the desired component f_k^0. That's one thing. By how much the accuracy will be adversely affected, though, is another thing and it has to do with the energy of the neighboring components relative to the central one we want to pick up clearly. One could argue that a decent measure of this contamination is the energy differential between adjacent components, which is true. But windows, on the other hand, are responsible for what makes it into the final signal to be transformed and how much that energy is curbed. It turns out that taking amplitude differentials is equivalent to comparing the Fourier transforms of the windows, only the latter is more convenient computationally. Therefore, we will use the frequency response differentials $W(f_k^0) - W(f_k^{-1})$ and $W(f_k^0) - W(f_k^{+1})$ as a metric of harmonic contamination. The greater $\Delta W(f_k^i)$ is, the greater the likelihood for rejecting undesirable components. Table (2.1) in the next subsection contains the windows specs (harmonic window limits) for each of the three harmonics f_k^i and their frequency response differentials for comparison. Since this model is central to my dissertation, a practical example is given using the four windows presented in the previous section (2.3.4.1).

2.3.4.3 Example of Window Limits

One of the characteristics of the harmonic–filter analysis at least as implemented by heterodyning a signal by kf_a is that lower harmonics are resolved much better than higher harmonics. From personal experience using this method and due to its flexibility in isolating individual harmonics, the model resolution for lower partials (including the fundamental[14]) is superb. Fortunately, the first half a dozen or so harmonics are the ones responsible for most of what we perceive as a tone, hence our decision to weigh only the first five harmonics in Equation (2.11).

Suppose that no prior empirical knowledge of the fundamental frequency exists. Based on general music theory we know that the reference tone is A_4 at $f_1 = 440$ Hz. We also know from Byzantine music theory that the tone D_4, the f_0 to be estimated, is 42 atoms below the reference A_4, i.e., $\Delta atoms = 42$. A quick calculation from Equation (1.5) shows

[14]The topics discussed here are beyond strict fundamental frequency estimation. It's noteworthy that the whole spectral content of the signal is considered and analyzed in conjunction to psychoacoustic empirical results. However, the term is established enough to be adopted here.

Table 2.1 – Harmonic Analysis Limits for a hypothetical example with $k = 3$ and $\Delta f^{D_4} = 0.03 f_a$ for four commonly used window functions.

Window Type	W_{dB} for $f_3^0 = 0.09 f_a$	W_{dB} for $f_3^{-1} = -0.94 f_a$	W_{dB} for $f_3^{+1} = 1.12 f_a$	$W(f_3^0) - W(f_3^{-1})$	$W(f_3^0) - W(f_3^{+1})$
Rectangular	−0.1	−24.0	−19.6	23.9	19.5
von Hann	−0.2	−32.2	−32.3	32.0	36.1
Hamming	−0.2	−38.6	−44.1	38.4	43.9
Blackman–Harris	−0.4	−70.5	−92.1	70.1	91.7

that the theoretical fundamental frequency for D_4 should be

$$\begin{aligned} \Delta atoms &= 72 \cdot \frac{ln\beta}{ln2} \\ \frac{ln2 \cdot 42}{72} &= ln\left(\frac{f_1}{f_0}\right) \\ 0.404336 &= log_e\left(\frac{f_1}{f_0}\right) \\ e^{0.404336} &= \frac{f_1}{f_0} \\ f_0 &= \frac{440}{e^{0.404336}} \\ \Rightarrow f_0^{D_4} &= 293.6647\ Hz \end{aligned}$$

which is what western theory frequency tables would show (Benson (2006) [11], pg. 379). Figure (2.1) shows why the Byzantine and Western D_4 agree. Since no prior knowledge exist on the fundamental frequency of D_4, $f_0^{D_4}$, we make the analysis frequency for the phase vocoder purposes equal to the fundamental, i.e., $f_0^{D_4} = f_a^{D_4} = 293.665$ Hz. The phase vocoder method is then used to identify the fundamental frequency of the signal. Since $\Delta f^{D_4} = 0$ and the signal is well–behaved, the accuracy of the f_0 estimate is as good as it can possibly be (within the framework of phase vocoder).

Let us assume, however, that empirical spectrum analysis of the data suggests that the complex tone D_4 has a "true" $f_0^{D_4} = 302$ Hz, about 3% higher than what was used for the phase vocoder. The accuracy of the estimate then becomes a function of not only this $\Delta f^{D_4} = 0.03 f_a$, but also the harmonic number and the frequency resolution and spectral leakage of the window of our choice. Following Beauchamp (2007) [2] the harmonic analysis limits are tabulated below for this hypothetical yet realistic example.

Let us now consider the case of a higher harmonic, for example $k = 10$, with the same Δf^{D_4} and window functions for comparison. The results are shown in Table (2.2).

The window response difference columns (last two columns of the table above) are a measure of relative rejection of undesired components; the higher the number, the better that window "resolves" that harmonic in the sense that it rejects the undesired adjacent harmonics adequately enough (results here are for the Δf of about half a semitone we have been using). Note, however, that if the Δf is large enough, the harmonics above it are

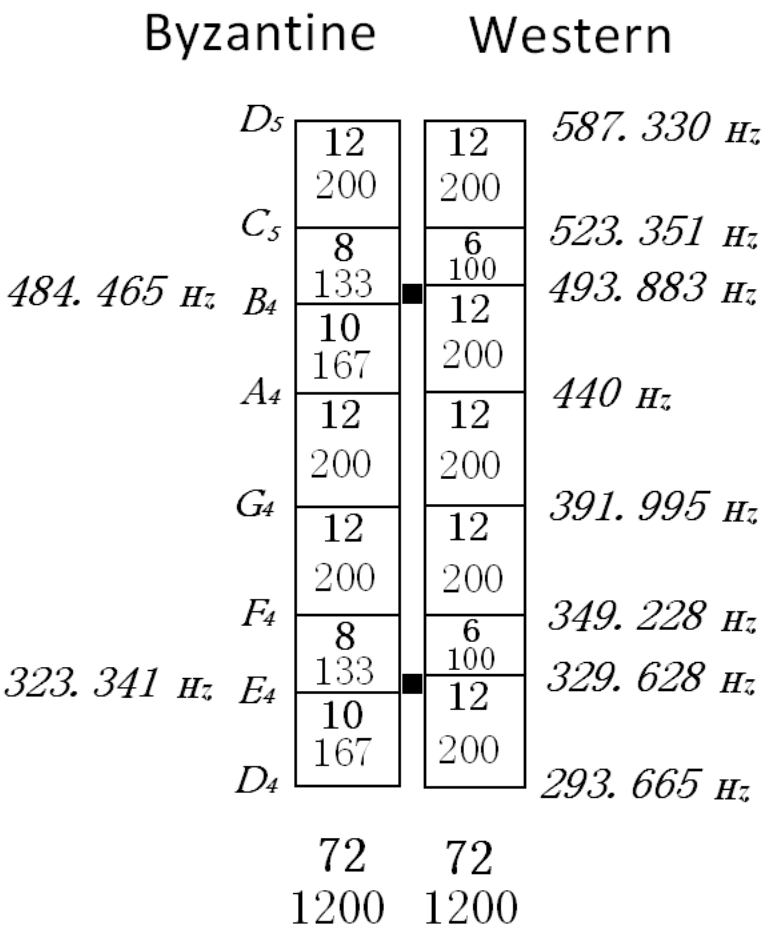


Figure 2.1 – *Byzantine vs Western scale.* Within interval boxes the upper **bold** number denotes atoms and the lower cents. Numbers outside the boxes are frequencies in units of Hz. Of all Byzantine music scales, the diatonic is the closest to the Western intervals. The subtle differences are shown in the two different frequencies (left) which create four different intervals. In fact, even though frequencies will differ as the scale progresses periodically in the upper or lower tones, the atoms, cents, or frequency ratios will be identical to the first *tetrachord*, the first four notes which enclose three intervals. A tetrachord, having a just intonation ratio of $4:3$, is a *perfect fourth*, i.e., $D_4 \mapsto G_4$. Along with the connecting interval $G_4 \rightarrowtail A_4$ it forms a *perfect fifth*, a $3:2$ ratio. Perfect intervals are consonant, that is why no matter if you are in a Byzantine or a Western scale D_4 will always be 293.665 Hz with respect to the reference tone A_4.

(hopefully) happening periodically in frequency, and at some point the Δf, if incremented just right, will coincide with a higher harmonic that is just an "echo" of the lower.

Direct comparison of the two tables gives at least three practical insights. First, notice how much smaller the frequency response difference numbers are for $k = 10$ as opposed to $k = 3$, i.e., higher harmonics are more difficult to isolate since the relative rejection of the neighboring corrupting components is difficult. Second, notice how W_{dB} for f_k^{-1} is higher than W_{dB} for f_k^{+1}, which means that the contribution of f_k^{-1} in corrupting f_k^0 is much higher than the corruption effect of f_k^{+1}. This is due to the relevant proximity of the frequency to be

Table 2.2 – Harmonic Analysis Limits for a hypothetical example with $k = 10$ and $\Delta f^{D_4} = 0.03 f_a$ for four commonly used window functions.

Window Type	W_{dB} for $f_{10}^{0} = 0.09 f_a$	W_{dB} for $f_{10}^{-1} = -0.94 f_a$	W_{dB} for $f_{10}^{+1} = 1.12 f_a$	$W(f_{10}^{0}) - W(f_{10}^{-1})$	$W(f_{10}^{0}) - W(f_{10}^{+1})$
Rectangular	−1.3	−9.7	−13.7	8.4	12.4
von Hann	−2.1	−14.4	−35.3	12.3	33.2
Hamming	−2.5	−17.7	−61.8	15.2	59.3
Blackman–Harris	−4.9	−33.4	−115.7	28.5	110.8

resolved and the adjacent components. If Δf is positive the above is true; if it's negative, the reverse is true (which means that f_k^{+1} is much closer to f_k^0). Third, the more sophisticated the window, the better the phase vocoder model isolates partials. Again, higher difference numbers means higher likelihood to get rid of unwanted neighboring frequencies, and clearly the relationship between window sophistication and better isolation is at least directionally in agreement with this statement.

Which window is best depends on (1) which partial needs to be resolved, (2) Δf (assuming the signal is well–behaved with clear partials at integer multiples of the fundamental), and, in general, (3) the spectral properties of the tone. From the above discussion, it would seem reasonable to use a more sophisticated window than not. But apart from computational expense, Hamming and von Hann windows lend themselves to better rejection due to their having a narrower frequency domain width ($2/f_a$ compared to $4/f_a$ for the Blackman–Harris window). This has to do with corruption of f_k^0 from components that are not strictly adjacent to it. A direct comparison between the Hamming and von Hann windows shows that von Hann is even better than Hamming when it comes to rejecting non–immediate components. On a related note, the Blackman–Harris window's main lobe width starts out narrower compared to the other windows which makes it more sensitive for appreciable Δf. Coupled with its wide window in the time domain (time-resolution issues), the Blackman–Harris window makes for a good option only for a handful of specialized situations. The Hamming window was the one implemented in SNDAN.

In our case we want $\Delta f \to 0$ to avoid having to analyze algorithm accuracy using limits. This is achieved by having a priori knowledge of the f_0 and setting that equal to f_a.

2.3.5 Analysis Step Implementation

This section describes some basic implementation aspects of the heterodyne–filter analysis, the implementation of phase vocoder. Since the input signal is already digitized at a sampling frequency of $f_s = 44,100$ Hz[15], we have a time–series of $s(n/f_s) = s(n\Delta t) = s(\tau)$

[15]Standard compact disc quality. Frequencies up to the upper human hearing limit 20,000 Hz are represented.

samples at $n = 0, 1, 2, 3, \ldots$. Using Equation (2.20), Equation (2.14) can be expressed as

$$\begin{aligned}\tilde{c}_k(t) &= f_a \int_{t-\frac{P}{2f_a}}^{t+\frac{P}{2f_a}} \frac{w(t-\tau)}{f_a} s(\tau) e^{i2\pi k f_a \tau} d\tau \\ &= f_a \sum_{p=0}^{P-1} \alpha_p \int_{t-\frac{P}{2f_a}}^{t+\frac{P}{2f_a}} cos(2\pi p f_a (t-\tau)/P) s(\tau) e^{i2\pi k f_a \tau} d\tau. \end{aligned} \tag{2.27}$$

We need to sample this equation with the ultimate goal of estimating its parameters. The input signal is divided up into i time frames with a frame rate of $f_a/2$. Using integer sample indices n and m we introduce discrete times at $t_n = n/f_s$ and $\tau_m = m/f_s$ which replace the corresponding continuous variables in Equation (2.27) above to give the sampled version

$$\tilde{c}_k(n/f_s) = f_a \sum_{m=n-N/2}^{n+N/2-1} w'[(n-m)/f_s] e^{-i2\pi k f_a m/f_a} s(m/f_s)/f_s \tag{2.28}$$

where $w'(\cdot) = w(\cdot)/f_a$ is the usual normalized window and $N \cong P f_s/f_a$ is the window length (in samples). If Equation (2.14) gives the complex amplitude of the kth partial, Equation (2.28) gives the sampled complex amplitude of the kth partial.

To simplify the above formula we make the following reasonable substitutions:

$$\begin{aligned} \tilde{c}_k(n/f_s) &\leftarrow \tilde{c}_k(n) \\ w'[(n-m)/f_s] &\leftarrow w'[(n-m)] \\ s(m/f_s) &\leftarrow s(m) \end{aligned}$$

to obtain

$$\tilde{c}_k(n) = f_a/f_s \sum_{m=n-N/2}^{n+N/2-1} w'[(n-m)] e^{-i2\pi k f_a m/f_a} s(m) \tag{2.29}$$

$$= \frac{P}{N} \sum_{m=n-N/2}^{n+N/2-1} w'[(n-m)] e^{-i2\pi P k m/N} s(m). \tag{2.30}$$

A couple of other considerations about Equation (2.30) seem necessary. The limits of the summation suggest an asymmetry about the median sample point. This is easily fixed by shifting the window function by half a point. Since the fast Fourier transform is used to implement the DFT of Equation (2.30), if N, the window length, is a power of two the computations will be more efficient[16]. For this reason, the signal must be resampled [Smith & Gossett (1984) [76] offer a flexible sampling–rate conversion method which is used in this implementation. To avoid aliasing due to undersampling, the new sampling rate f_s' must be higher than the regular f_s. We let

$$N = 2^M = 2^{ceil[log_2(P f_s/f_a)]},$$

[16]Note that this is not necessary, but it is more computationally efficient.

where $ceil(\cdot)$ is the ceiling function[17], so that

$$f_s' = \frac{N f_a}{P}. \tag{2.31}$$

As an example, suppose the tone D_4 is to be fed into the phase vocoder. The analysis frequency is $f_a^{D_4} = 293.665$ Hz, the sampling frequency is $f_s = 44,100$ Hz, and the 2–term Hamming window of width $2/f_a$ is used, i.e., $P = 2$. Then

$$N = \frac{P f_s}{f_a} = \frac{2 \cdot 44,100\ Hz}{293.665\ Hz} = 300$$

and the next power of two is $2^9 = 512$, which when used in Equation (2.31) gives a new sampling rate of $f_s' = 66,969.6$ Hz. If $f_a^{G_4} = 391.995$ Hz, then $N = 225$ and the next power of two is $2^8 = 256$, which makes $f_s' = 50,175.36$ Hz. All these are taken care of automatically based on the f_a to be analyzed.

2.4 McAulay–Quatieri—Frequency–Tracking Analysis

The phase vocoder method postulates an ideal model that experience is then referenced against. It does not lend itself to analyses where practical considerations prevent the signal from being less than perfect. For example, in voice signals particularly, a partial may or may not exist for the entire duration of the tone. It may "die" or a new one may be "born." This is particularly true of higher harmonics that do not carry much energy but are certainly audible and help form the quality of the sound. Another practical consideration is that harmonics are not likely to always be perfect integer multiples of the fundamental. It would be nice to have a way to reject all the unwanted components based on perfect harmonicity requirements (like the phase vocoder does), but it would also be nice to have a more realistic tool which will track the real components of a signal even with variable ratios among them.

The McAulay and Quatieri (1986) [43] method of frequency tracking provides this flexibility. It was designed for speech signals and it was later adopted to music applications by Smith and Serra (1987) [77]. Subsection (2.4.1) provides the algorithm. Subsection (2.4.2) discusses resolution issues for this method. Subsection (2.5) gives some information about the SNDAN user interface.

2.4.1 McAulay–Quatieri—Frequency–Tracking Algorithm

This section gives a brief outline of the actual frequency–tracking algorithm based on McAulay and Quatieri (1986) [43]. The idea is simple: pick the peak frequencies for each of the Discrete Fourier Transforms (DFTs) of each overlapping frame and then concatenate them to form frequency–vs–time plots. Along the way, the amplitude A_k and phase θ_k of

[17] The ceiling and floor functions map a real number to the smallest following or largest previous integer with respect to its argument.

each partial is also retained and used for other calculations or plots within the SNDAN user interface.

Since this is an established method for frequency estimation, extensive documentation exists in the literature. Some good references include McAulay and Quatieri (1986) [43], Smith and Serra (1987) [77], Serra (1989) [72], Beauchamp's book (2007) [2] and some of his papers, for example, Beauchamp (1993) [6], Beauchamp (1966) [3], Beauchamp (1993) [7], Beauchamp (1975) [5], Fitz, Walker, and Haken (1992) [18] [24], and Maher & Beauchamp (1994) [41]. Figure (2.2) summarizes the algorithm.

2.4.2 Time–Bandwidth Product—The Uncertainty Principle

Subsection (2.3.4.1) alluded to time–frequency trade–offs. A more detailed account is given here. The discussion starts out more generally with Bracewell (2000) [12], and Stoica & Moses (2005) [80], and concludes with some practical implications from Beauchamp (2007) [2] and Smith & Serra (1987) [77].

The window's *length*, M, is responsible for the energies in main lobe width and the sidelobes of the windows amplitude response. This limitation introduces the notion of frequency resolution and statistical variance trade–off. The window's *shape* introduces a trade–off between smearing and spectral leakage. It should be clear that there is always some kind of trade–off between the time and the frequency domains. We need a mathematical framework to help us discuss the balancing of these two domains. This well–known framework goes by several different names, comes in different notations, and gives insight into just how one can start thinking about optimizing a situation at hand given some concrete constants.

The uncertainly relation says that

$$\Delta t \; \Delta f \geq \frac{1}{4\pi}, \tag{2.34}$$

where $(\Delta t)^2$ is the second moment (or variance[19] of the square modulus of a function in t, $|f(t)|^2$, and $(\Delta f)^2$ is the variance of the square modulus of the Fourier transform of the function, $|F(f)|^2$. So the product of the variances of the energies of the time domain and the frequency domain cannot be smaller than a constant. This is in terms of energies (or powers, which is basically like the energies, only the influx of energy within a time window).

[18] Fitz et al. (1992) present an interesting idea in regards to how one puts together the consecutive DFT'ed frames. This tracking *hysteresis* claims that the final n–point of a frame (the end of a window frame as it is chosen based on various proposed algorithms) is not truly a final end N–point (the point that really is or should be the final point of a window frame to yield minimum leakage or distortion), unless it is persistent for a number of frames. So, it is not about the analysis of a signal, but how it is synthesized back to a coherent form after it has been processed. This made me think about cognitive theories for mildly autistic children. Maybe it is not all about parallel processing. Maybe it's as simple as a mismatch of putting together serial information *after* the input has been analyzed correctly. One can clearly see that the information is there somewhere, only the children seem confused about small portions of the information and where in time are situated.

[19] More precisely, the second moment is $<x^2>$, whereas the variance is the normalized $(x - <x>)^2 = <x^2> - <x>^2$. However, it is common practice to subtract the signal mean from the signal to make the second moment and the variance equal (see discussion in Bracewell (2000) [12], page 159).

McAulay–Quatieri Frequency–Tracking Algorithm

1. Calculate successive, overlapping Discrete Fourier Transforms (DFTs). Windowing and zero–padding for optimal partial isolation is used.

2. Identify each frame's spectral peaks. Each peak is determined by fitting a quadratic to the log of three DFT magnitudes $A_{\xi-1}$, A_{ξ}, $A_{\xi+1}$and the kth peak frequency is

$$f_k = (\xi + p)\Delta f_{DFT} \tag{2.32}$$

where

$$p = 0.5\frac{log(A_{\xi-1}A_{\xi+1}}{log(A_{\xi-1}A_{\xi+1}/A_{\xi}^2)}. \tag{2.33}$$

The A_k and θ_k for each peak are also computed using the real and imaginary parts of the transform. Parameters on each frame are stored for analysis and other FFT bin information is discarded. A logic based on an absolute or frequency–covarying amplitude threshold (Serra (1989) [72]) for picking global and not every local maximum is applied.

3. Peak frequencies of consecutive frames are concatenated. This is the algorithm's most crucial step as there is no perfect way of identifying "deaths" and "births" of frequency trajectories. A "link" index $\kappa_{k,i}$ determines where the connection should happen. All indexing and time information are retained for frequency plots and further calculations. A track is thus formed.

4. Information on peak data for each track is stored in a file.

Figure 2.2 – McAulay–Quatieri Frequency–Tracking Algorithm.

In terms of the Discrete Time Fourier Transforms (DTFTs), the uncertainly relation can be simplified outside the boundaries of periodicities (if we felt that $1/4\pi$ is too cumbersome for the eye) and say that the product of the equivalent duration and equivalent bandwidth is exactly one

$$\Delta t \ \Delta f = 1 \tag{2.35}$$

or put in another usually encountered notation

$$N_e \ \beta_e = 1 \tag{2.36}$$

where

$$N_e = \frac{\sum_{t=-(M-1)}^{M-1} w(t)}{w(0)} \tag{2.37}$$

and

$$\beta_e = \frac{\frac{1}{2\pi}\int_{-\pi}^{+\pi} W(s)ds}{W(0)} \tag{2.38}$$

where $W(s)$ is the Fourier transform of $w(t)$ and M is the window length (as opposed to N which is the sequence length). Equation (2.36) will always be unity for all functions no matter their mathematical properties, but Equation (2.35) only assumes its minimum value with Gaussian functions.

Once the window length has been determined there are not much we can vary but the window's shape. Since the window's length is so important, we can also think about Equation (2.35) as

$$\Delta t \ \Delta f = \frac{1}{M} \tag{2.39}$$

and the interdependence becomes more apparent. We can see, for example, that the equivalent bandwidth Δf is on the order of $1/M$ which is the equivalent *length* in the time domain, i.e. $\Delta f = O\left(\frac{1}{M}\right)$. The frequency resolution is $1/M$, but the second moment is proportional to M/N [for a derivation of this result and a discussion please refer to Stoica & Moses (2005) [80], Section (2.5.1) and Section (2.6.1)]. Therefore, the window length binds both the spectral resolution and the variance and the two are inversely proportional. This is why we said before that the window length should be chosen to optimize the choice of these two.

At the same time, with M fixed, $w(0)$ is also fixed and the area under the curve of the Fourier transform of the window equals the central ordinate of the original function, that is to say, $w(0) = \int_{-\infty}^{\infty} W(s)ds = 1$. This imposes another limitation: the main lobe width and sidelobes cannot be reduced simultaneously once M is set. But the main lobe width is associated with the windows frequency resolution which when allowed to become coarser leads to smoothing, or smearing; the sidelobe height is associated with the leakage of the window frame to be transformed. We then turn to the window's shape to address this trade–off. The more smoothly the data are weighted down to zero in one domain the more concentrated its energy will be in the other domain. That is why we said before that the shape of the window should be chosen with regard to the smearing vs leakage trade–off.

In practice, the signal is decimated into i bins. To accommodate this into our discussion we say that for two adjacent frequencies, say f_0 and f_1, to be resolved they must be apart by a frequency difference of Δf_w or more, where Δf_w is the bandwidth of the window and is defined as

$$\Delta f_w = B_w \frac{f_s}{M} = B_w \Delta f_b, \tag{2.40}$$

where B_w is the bandwidth of the window in number of bins, Δf_b is the bin separation frequency, M is the number of samples in the window, and f_s is the sampling frequency. Δf_w is also the lowest harmonic in a signal that can be adequately isolated, because the next harmonic will be twice this value (since f_a is set, then $f_k = k f_a$). We will assume that each spectral peak corresponds to a sinusoidal frequency component in the signal. Experience (see, for example, Beauchamp (2007) [2] and Smith & Serra (1989) [77]) shows that a 3–

bin bandwidth separation is enough to adequately resolve adjacent partials at a low enough frequency to fit our needs[20].

Since 3–bin frequency units resolve peaks in the magnitude spectrum accurately enough, we can take a look at a typical scenario where $f_s = 44,100$ Hz and $M = 2^{10} = 1024$. The time duration of the window is $N/f_s = 23$ ms. The bin separation frequency is $\Delta f_b = f_s/N = 43$ Hz and $\Delta f_w = B_w \cdot \Delta f_b = 129$ Hz, that is to say, if the lowest fundamental in the signal is about 130 Hz then the McAulay–Quatieri method should analyze it accurately. This is well below the lowest frequency analyzed for our purposes. The requirement $B_w = 3$ limits the maximum number of peaks K_i that can be detected within each bin i to $f_s/(6\Delta f_b)$. The window length must be chosen according to how much of a peak separation we would like the model to resolve. Assume that a desired minimum value of peak separation is 20 Hz. Since this minimum is spread over 3 bins, the real bin separation is 20/3=6.67 maximum. If the software automatically uses an $M = 2^{13} = 8192$ with the same $f_s = 44,100$ Hz, then a bin separation of 5.38 Hz is achieved. This in turn gives the new bin separation of 16.15 Hz (which is below the minimum of 20 Hz we required). Finally, the algorithm can resolve up to a $8192/6 \approx 1,365$ peaks.

2.5 SNDAN

Now that the phase vocoder and the McAlay–Quatieri frequency tracking methods have been presented, we devote a section on explaining their computational implementation code, SNDAN, and also provide some illustrative examples of its analysis using a sound snippet from Nafpliotis' recordings. Other analysis examples will follow in subsequent sections.

SNDAN: Musical Sound Analysis, Graphics, Modification, And Resynthesis Routines For Unix is an extremely powerful and versatile tool owing its development to Professor James Beauchamp of the *University of Illinois at Urbana-Champaign*. The author of this dissertation started using this tool in 2005. It is basically a library of routines written in C, and can be run from a UNIX or a Windows MSDOS prompt command. A couple of Graphical User Interface versions of SNDAN exist, but the steep learning curve of running the original code is well compensated by its tremendous flexibility. Here, only a very small portion of it is demonstrated and used. Other applications include resynthesizing the sound after it has been decomposed by either the phase vocoder or the McAlay–Quatieri frequency tracking method, applying smoothing filters specifically designed for vibrato, creating an impressive number of graphs and calculations, exporting data after operations for analysis in other software, etc. It is ideal for demonstrating sound engineering principles in the classroom. It has been used for other Ph.D. dissertations and Master's theses. The code, documentation, and everything one needs to get started can be found here http://ems.music.uiuc.edu/beaucham/. Another attractive feature of SNDAN is its extensive documentation on the mathematical theory behind the computations in books and

[20]In fact, $B_w = 2P$ with P being the terms in the window. But even more practically, taking a look at the zero crossings of the window response plot will confirm that for the rectangular, Hamming and Von Hann, and Blackman–Harris windows the B_w values are 2, 4, and 8 respectively. That we settled on $B_w = 3$ is more of a practical consequence (computationally cheap).

articles (some of them have been cited previously). A short history of its origins by James Beauchamp is quoted below.

> *The Origins and Development of SNDAN*
>
> SNDAN is an outgrowth of work that I did as early as 1966, when I gave my first paper at the AES on music sound analysis. Then I co–edited a book entitled "Music by Computers" in 1969 which contained a chapter by me on sound analysis/synthesis using main frames. I continued to use main frames for analysis/synthesis and small off-line computers for A/D and D/A until we first set up the Computer Music Project here in 1984–85. Back in the mainframe days, the analysis/synthesis package was called TONEAN and it was written in FORTRAN. Rob Maher arrived in 1985 to work on a doctorate in electrical & computer engineering, which dovetailed very nicely with the arrival of our first desktop Unix computers. I held a course for musically–inclined engineers and programmers in 1985, and out of that came several very useful products, all written in C, which we are still using, including Music 4C and g_raph, our graphics package. In the meantime, as part of his doctorate, Rob wrote the MQ and PV software, which, in a certain sense, is the most important part of SNDAN. (Rob left in 1989, and I still miss him, as he was probably the best assistant I ever had.) I wrote most of the stuff in add_syn, sig, and view_an (where monan resides). George Chaltas wrote g_raph in 1985 for the Tektronix protocol. Camille Goudeseune and I ported g_raph for EPS, and in 1996–97, Tim Madden extended the graphics considerably for the 3D ('pp') and 2D ('ftc') spectrum graphs, including the use of color to differentiate harmonics (3D) and to indicate intensity (2D). During 1991–93 Andrew Horner wrote several programs for sound analysis/synthesis based on PV output, most notably using the method of the Genetic Algorithm. Music 4C instruments for Spectral Dynamic Synthesis (Beauchamp and Horner) and, very recently, for Piano Wavetable Synthesis (Zheng Hua) have been developed based on analysis using SNDAN.
>
> As of this writing, 318 people have down–loaded SNDAN. Some people have used it for their academic theses. For example, Rebekah Brown at Indiana University used it for her doctoral thesis on intonation of violin performances in 1996 and John Hajda used it for his Ph.D. dissertation on musical instrument timbre in 1999. Jochim Krimphoff used it for his masters thesis on instrument timbre at IRCAM(Paris) in 1994, and it has been used by others in Stephen McAdams' Music Perception/Cognition group at IRCAM since then. It is also being used extensively by Andrew Horner at HKUST for sound analysis/synthesis projects which have been documented by many publications by him and other authors in JAES and CMJ since 1993.
>
> Two GUI versions of SNDAN have been written, both at UIUC. AnView was written for the black NeXT environment by Chris Gennaula and Camille Goudeseune in 1992–93. Armadillo was written for the MacOS/PPC environment in 1998–99. Instructions for obtaining these can be accessed via my web page at http://ems.music.uiuc.edu/people/beauchamp. These are not 1–to–1 implementations. They lack many things that SNDAN has, and they do some things that SNDAN doesn't do. Needless to say, for the unitiated user, they are a lot easier to use than SNDAN is.
>
> James Beauchamp
> University of Illinois at Urbana-Champaign
> 21st January 2000

The next subsections interject empirical insight into the so far theoretical treatment. A word about the demonstration data. A snippet of sound from one of Iakovos Nafpliotis' recordings is used as demonstration data. The tone is A_4 and preliminary analysis on this snippet shows $f_0 \approx 445$ Hz, the $f_s = 44,100$ Hz, the duration is 1.707 seconds and it was chosen specifically because it contains three distinct syllables: "me—nos—pros." The first two syllables have a musical duration of one beat and the last one of two beats, that is to say, had this been a 4/4 meter (which, in reality it is not[21]) the first two syllables would have been assigned a duration of 1/4 and the last one 2/4.

2.5.1 Visual Comparison of Phase Vocoder and McAlay–Quatieri Methods

The phase vocoder analysis was implemented using `pvan` and the McAlay–Quatieri analysis was implemented using `mqan`, both standard packages of SNDAN. Analysis data were dumped into an ASCII flat file using `andump` and were then analyzed in SAS®[22]. Plots were created using the `monan` and `mqplot` libraries and `antomq` was used to prepare data for one of the above analysis methods or to change formats from `an` to `mq`.

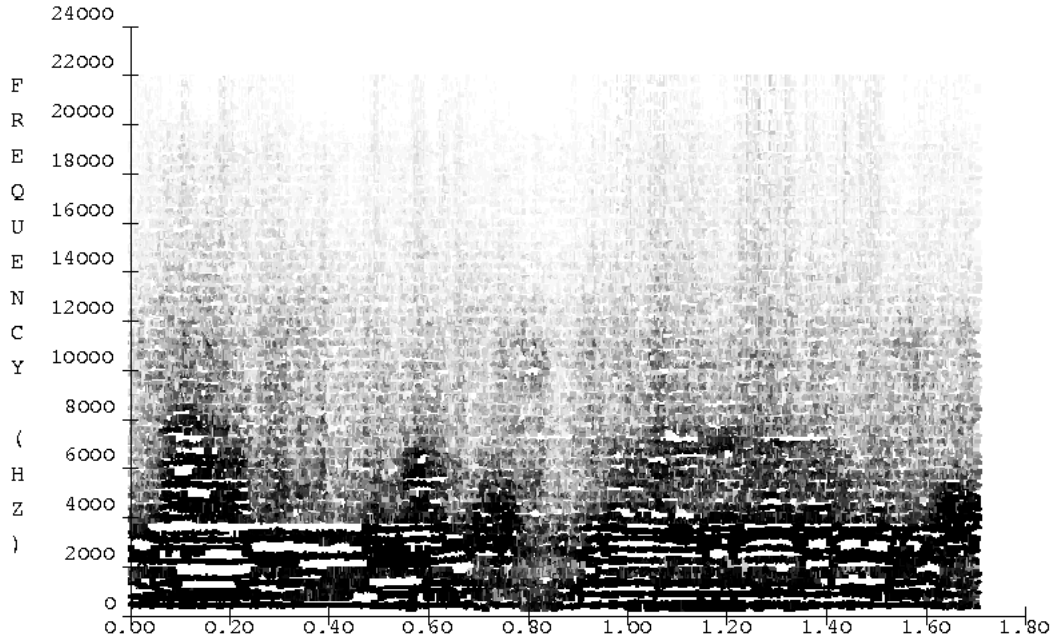
The analysis log informs us that the signal was segmented into 1591 frames each with a duration of 0.001124 seconds[23]. Forty-nine harmonics were resolved and $f_a = 445$ Hz (like we requested it to be). Figure (2.3) shows the two spectra for visual inspection. Forty–nine harmonics on top of $f_0 \approx 445$ Hz implies a maximum resolved frequency of 21,805 Hz, which is above the rough upper frequency capability of human resolution and about half of $f_s/2 = 44,100/2 = 22,050$ Hz. This indicates that the work on cleaning up the analogue vinyl records was superb—one does not get that high a quality in digital audio unless the analogue is nearly perfect. This fact in itself, should make us feel very comfortable analyzing this sample in terms of spectral content available. It may be beyond what we really need for estimating f_0, but higher harmonics are crucial for a host of perceptual analysis. The best samples of Nafpliotis' recordings analyzed before the release of Alygizagis' CDs (2008) [1] did not contain any frequencies above 3.5 kHz.

---

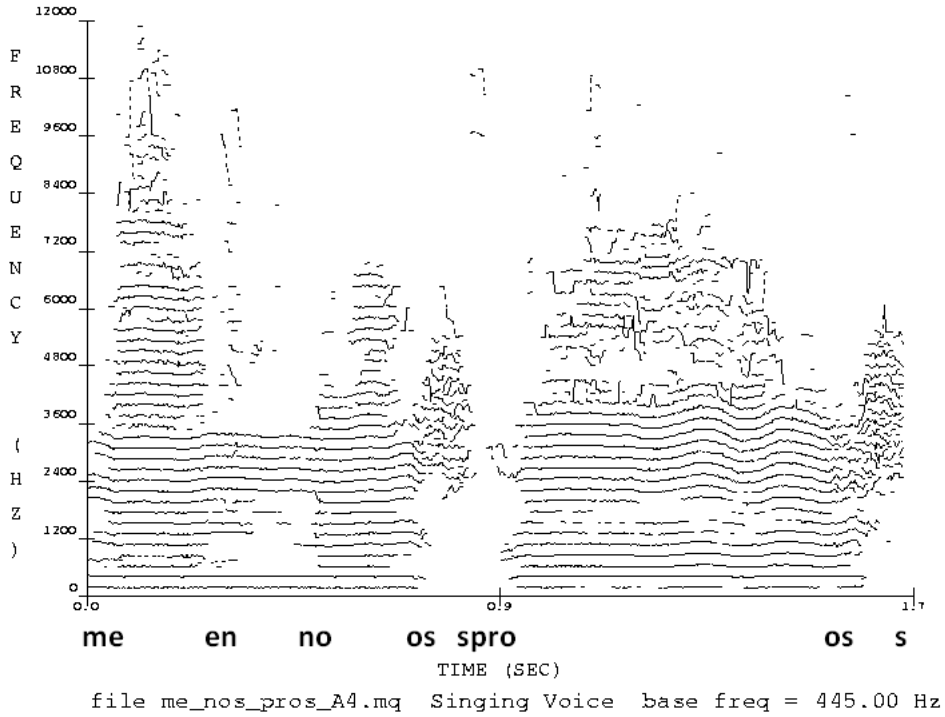
[21] This is another distinction between Western and Byzantine music metronomy. In Byzantine it is allowed to insert unequal meters among otherwise normal rhythm to emphasize prosodic meaning at the expense of broken rhythmic continuity. Along with scale intervals, this distinction is also approaching extinction under progressivistic trends. Future research on Byzantine music rhythm is also possible, along with frequency research. A good reference for research on metronomic tonality using spectra is Sethares (2007) [73].

[22] SAS® is a registered trademark of SAS Institute Inc.

[23] Remember that the phase vocoder uses either the phase spectrum or the provided f_a to fine tune where to look for f_0 and this compensates the coarse Δf of a DFT with the same sampling rate, f_s, and FFT window length, N. For example, in DFT we can use $\Delta f\ \Delta t = \frac{1}{N}$, with $\Delta t = \frac{1}{f_s} \approx 2.26 \times 10^{-5} sec$, $N = \frac{N_{TOT}}{frames} \approx 47 samples/frame$, to estimate $\Delta f \approx 938 Hz$, which is coarse. Here the vocoder is using f_a. Had f_a not been provided, the vocoder could have used $f_n = \frac{(\phi_2 - \phi_1) + 2\pi n}{2\pi(t_2 - t_2)}$, where ϕ_i is the phase of the sinusoid at t_i. The times could be selected to match those where the frequency peak is the maximum, or even closest to f_a (if f_a is known). If f_a is not provided, the vocoder chooses the f_n closest to the real frequency peak in the spectrum. This exploitation of phases dramatically improves f_0 estimation due to the powerful connection it reveals between time and frequency. The SNDAN implementation uses phases to suggest corrections on a user–provided f_a by least squares prediction. Note that the time resolution Δt is not compromised because the window length has not been increased at all. Sethares (2007) [73] provides a good discussion of this on page 118.



(a) Phase Vocoder Spectrum.

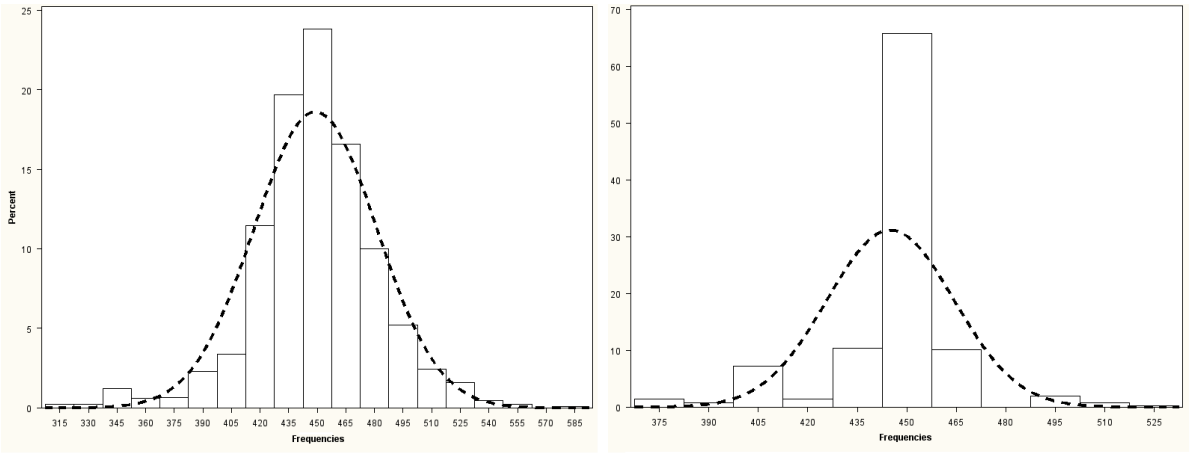


(b) McAlay–Quatieri Spectrum.

Figure 2.3 – Visual Comparison of Phase Vocoder and McAlay–Quatieri Spectra for the sample "me–nos–pros."

Note the "idealistic" nature of the phase vocoder spectrum as opposed to the "realistic" McAlay–Quatieri spectrum. The McAlay–Quatieri spectrum picks up every single

frequency. In doing so, notice how consonant sounds like "*n*" and "*s*" show up in the spectrum (see syllables below the spectrum). For example, "*n*," which is a voiced nasal (dental, alveolar) consonant, preserves 6 harmonics approximately in the 2.3 to 3.5 kHz range as if the nasal cavity didn't have much to do with altering the formant at this point in time, only a band–pass filter was applied by the tongue touching the palate. The fact that the "*n*" harmonics are holding up so well in the spectrum could be due to the fact that these syllables are sung, not spoken. Also, notice how stable the fundamental shows (second line in spectrum, the first one is an echoed subharmonic) even during the "*n*" sound. It's not up until "*s*" is uttered that the fundamental breaks—and for good reason. The voiceless (alveolar) fricative consonant "*s*" carries no vibrations from the vocal cords in it and consequently its frequency content is much more erratic, inharmonic and almost white–like than "*n*." Then, what is f_0 for "*s*"?



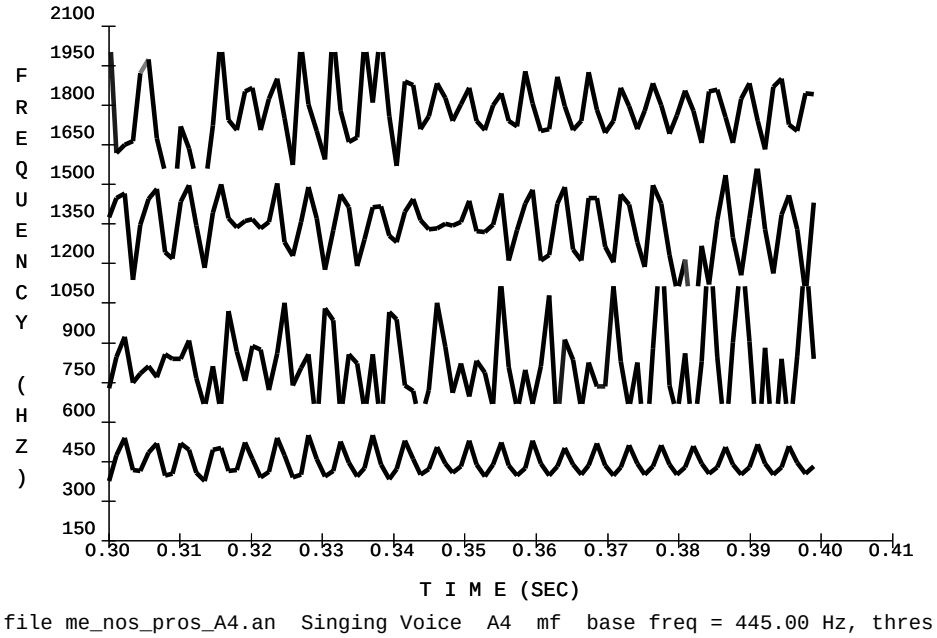
(a) Phase Vocoder Distribution. **(b)** McAlay–Quatieri Distribution.

Figure 2.4 – Visual Comparison of Phase Vocoder and McAlay–Quatieri Distribution. Frequencies are in units of Hertz.

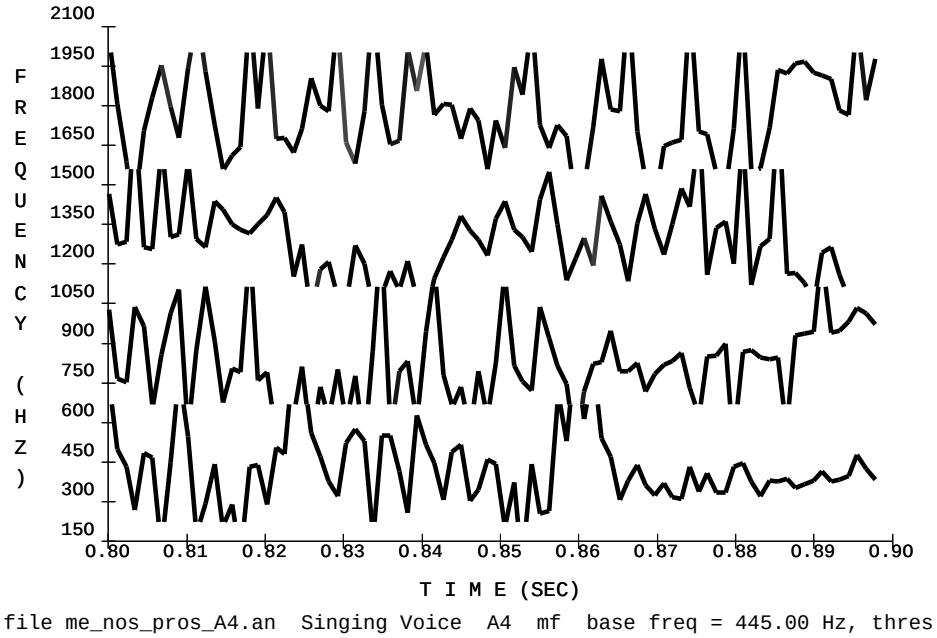
Figure (2.4) shows the distribution of the fundamental frequencies for the two methods. For "*n*" both algorithms should pick the fundamental frequency with no problem, but the McAlay–Quatieri method automatically sets the lowest instantaneous frequency within a frame equal to the fundamental. This means that for the fundamental of the "*s*" sound is in the thousands[24] and is also a non-integer multiple of the fundamental. Since this is undesirable, we first filter the data and then plot the McAlay–Quatieri frequency distribution. Also note that even with smaller sample size, the McAlay–Quatieri method exhibits reduced standard deviation compared to the phase vocoder, from about 30 to about 20 Hz. The dashed line shows a Gaussian curve fit to the data even though a Kolmogorov–Smirnov normality test indicates significant departure from Gaussianity (D=0.055; $p \leq 0.01$).

The realistic nature of McAlay–Quatieri method is preferable when looking for spectral patterns in the signal, but the idealistic nature of the phase vocoder is robust to hard–

[24]The peak picking method of McAlay & Quatieri could be thus improved, in my humble opinion, with an additional iterative step that first detected the overall f_0 and then filtered out any components that are clearly outliers via a 3σ method, some kind of confidence interval, or by a distributional tail.



(a) "n" Sound.



(b) "s" Sound.

Figure 2.5 – Visual Comparison of Phase Vocoder "n" and "s" Spectra.

to–handle parts of spectrum like consonants. Figure (2.5) shows a snapshot of how the consonants "n" and "s" are handled by the phase vocoder. Where the McAlay–Quatieri

method would have no components for "*s*" and only a couple for "*n*," the phase vocoder seems to be handling both of these consonants well despite their different spectral behavior. The f_0 for both can easily be detected in these examples, even though we see that for "*s*" the fundamental is quite irregular. As a visual comparison for how phase vocoder resolves vowels please refer to Figure (2.6).

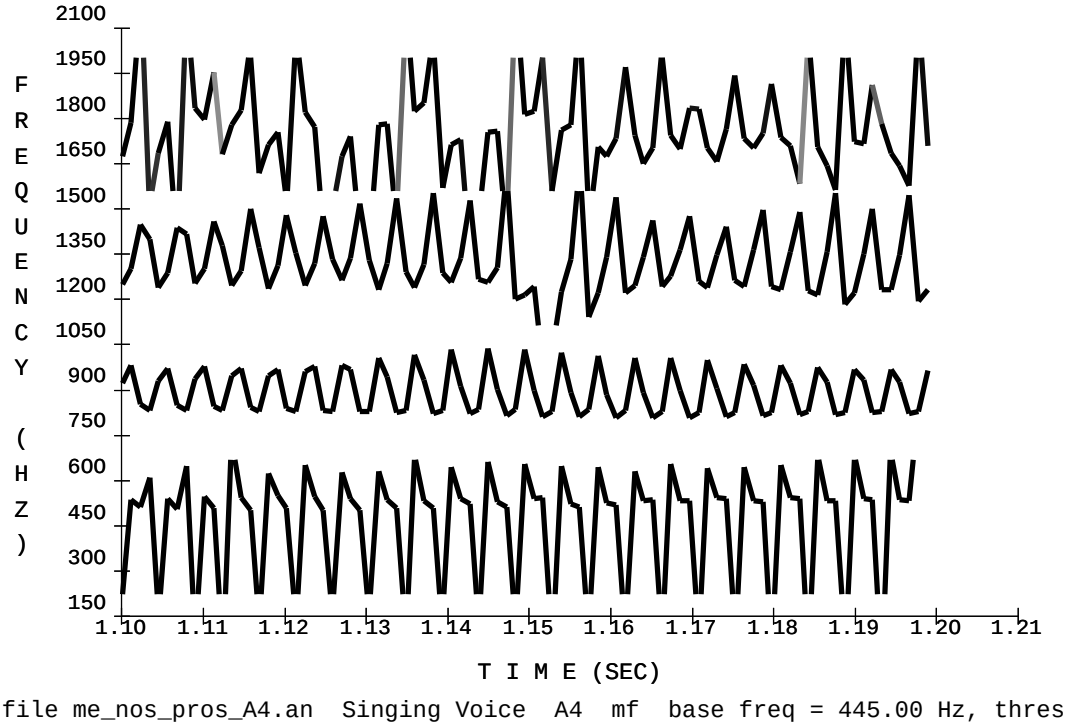
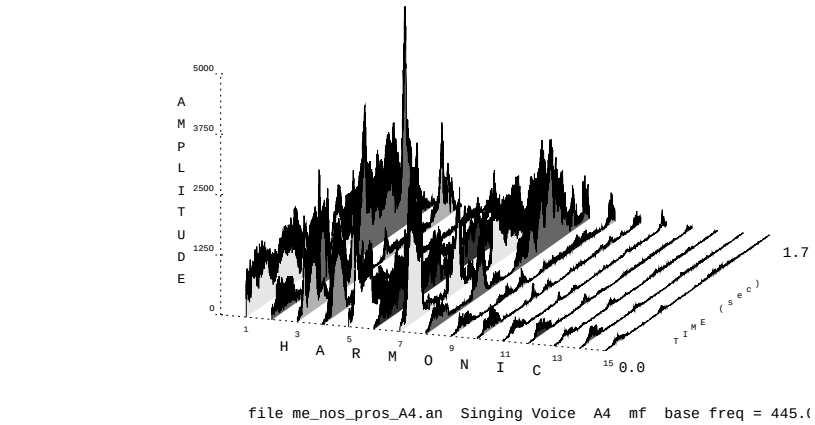


Figure 2.6 – Spectrum of the vowel "*o*" from the syllable "pros." Notice the semi–periodic vibration of f_0.

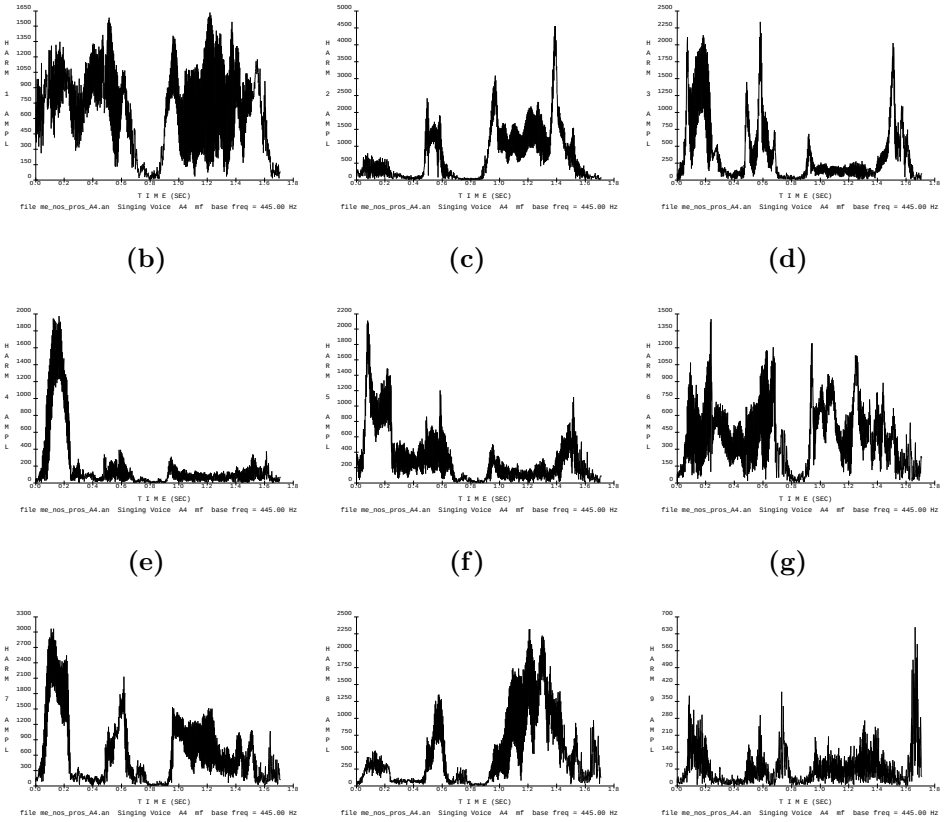
In these zoomed–in spectrograms (or sonograms rather) we can clearly see the effect of the convolution which made *each* partial bandlimited by half a bin around the analysis frequency. For example, since our $f_a = 445$ Hz the 3rd harmonic, or second partial, $f_2 = 1,335$ Hz has no content outside the $f_2 \pm 222.5$ Hz interval, i.e., outside [1112.5, 1557.5] Hz. Again, this is another layer of "idealism" on top of the concept of preserving only clear–cut integer multiples of the fundamental: the partials don't overlap at all. The effect of this truncation seems to have no effect in the resynthesized version of the signal (at least to the best of my ability to see differences in the spectra of the original vs the synthesized signal or or by listening to the two). Computationally, however, it is an advantage when we need to deal with messy and noisy signals.

Figure (2.7) shows a three–dimensional representation of the amplitude–harmonic–time where we can see that most of the energy of this signal is in the first eight harmonics. Subsequent pictures break down each harmonic amplitude across time.

Note how the first harmonic carries most of the vowel energy with big dips where the consonants occur and how the ninth harmonic in the 2 kHz range carries most of the energy



(a)



(b) **(c)** **(d)**

(e) **(f)** **(g)**

(h) **(i)** **(j)**

Figure 2.7 – *Amplitude (vs Harmonic Number) vs Time.* (a) Amplitude vs Harmonic Number vs Time; (b)–(j) Amplitude vs Time for the first nine harmonics of the snippet "me_nos_pros."

of the final "s" sound. As a matter of fact, a close examination of Figure (2.3b) will reveal that the ninth line is the first partial of the final "s" (tracking it almost all the way through

the end of time) and it also touches a little on the "s" sound of the syllable "pros." That's why the amplitudes in Figure (2.7j) are very large for the final "s" and reasonably large for the middle "s" sound. Higher partials have their amplitudes larger during these two instances of the fricative. Another example is the sound "o" in both consecutive syllables "nos_pros" in the second harmonic shown in Figure (2.7b) where only the "o" regions carrying energy and with the rest of the consonants and the vowel "e" repressed[25].

These and other interesting observations can be made by inspecting the amplitude of the harmonics over time. But now we move to another interesting concept, the Spectral Centroid.

2.6 Spectral Centroid

The root–mean–square amplitude (RMS) of a signal

$$A_{rms}(t) = \sqrt{\sum_{k=1}^{K} A_k^2(t)}, \tag{2.41}$$

is a well–known measure if its internal energy variability. Suppose now that each harmonic is weighted by a coefficient that is time–varying and normalized by the sum or total amplitude at every given point in time. Such a weight could be

$$\alpha_k(t) = \frac{A_k(t)}{\sum_{k=1}^{K} A_k(t)}, \tag{2.42}$$

which is basically a ratio of the harmonic to the total amplitude of the signal over time. If we weigh each harmonic by this value

$$BR(t) = \sum_{k=1}^{K} \alpha_k(t)k, \tag{2.43}$$

$$= \frac{\sum_{k=1}^{K} kA_k(t)}{\sum_{k=1}^{K} A_k(t)} \tag{2.44}$$

we have built a metric known as the *spectral centroid*, a measure traditionally associated with perceptual brightness (McAdams et al., (1999) [42]). Many instruments exhibit discernible

[25]Keeping in mind that the f_1 formant of vowel "o" is in the 400–600 Hz region, a case for Nafpliotis' *singer's formant* could easily be made, but this is a topic large enough for a paper on its own. Actually, another interesting phenomenon is how Nafpliotis projects his voice and how his formants have been sculpted from singing for 60 years in an architecturally peculiar acoustic environment. Just like an opera singer distributes the vowel frequencies in such a way to prevail over orchestral instruments, I suspect that a chanter needs to distribute his spectral power such that cancellations work out to deliver the vowels clearly and over greater distances. This is one example of why instantaneous frequencies are not a reliable measure of the overall tone and also why the tone length cannot be very short. This intentional formant–tone frequency adjustment makes the signal almost approach non–stationarity over longer time intervals, in the sense that statistics are intentionally altered slightly over time.

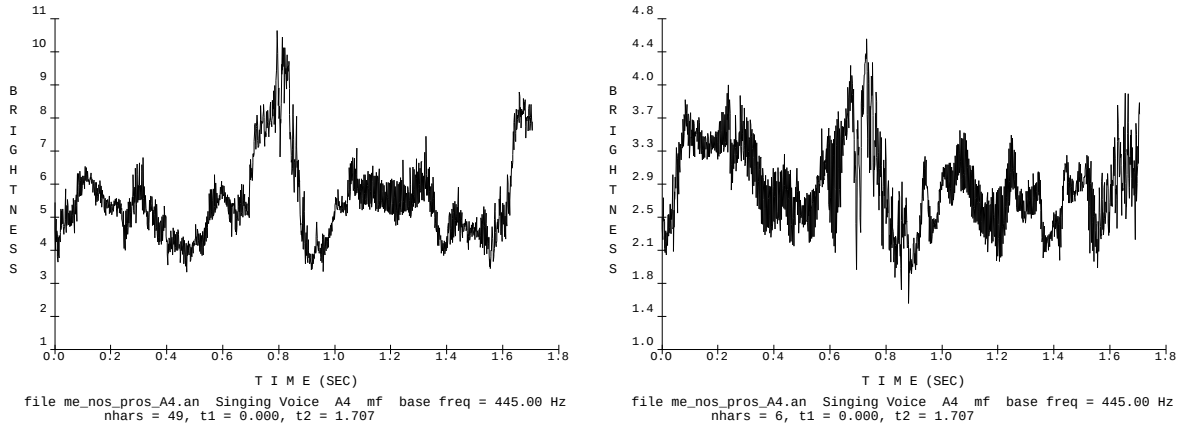
brightness differences across times, even though similar research on singing voice is less understood.

Instrumental brightness could be used as a benchmark to gauge Nafpliotis' brightness. Table 2.3 below tabulates some indicative instrumental centroid values. We see, for example, that Nafpliotis is not perceived as bright as a clarinet, but is brighter than a violin.

Table 2.3 – Normalized spectral centroids of some instrument sounds for comparison with Nafpliotis' voice.

Sound Source	Average Centroid	Maximum Centroid Value
Clarinet	6.4	11.1
Flute	3.4	11.2
Harp	1.6	15.2
Harpsichord	7.9	31.0
Nafpliotis	5.5	10.6
Saxophone (Alto)	4.1	9.8
Violin	4.6	7.5

If we plot the normalized centroid vs time [Figure (2.8a)] we observe that the *"s"* sound is clearly brighter than the rest of the signal where sounds like *"n"* and *"r"* are the least bright. Low–pass filtering the signal to keep only the first, say, six (6) harmonics, compromises the brightness of the *"s"* sound [Figure (2.8b)].



(a) **(b)**

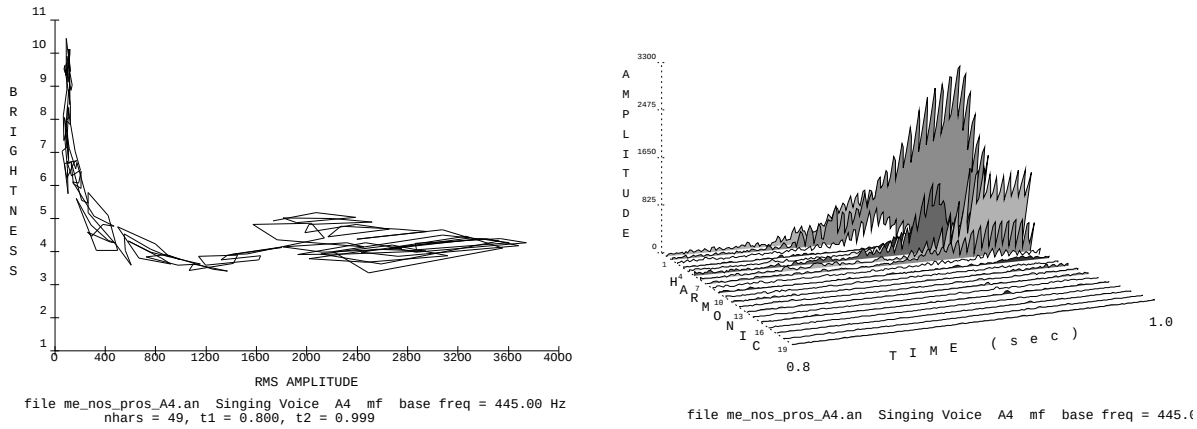
Figure 2.8 – *Spectral Centroid vs Time.* (a) All forty-nine harmonics used in calculation. (b) Only the first six harmonics used in calculation.

Note how the sound *"o"* of the last syllable "pros," which occupies the last half of the duration (last two out of four time beats) is brighter for the first beat than the second. This is exactly what a seasoned chanted would anticipate to see. There are two reasons for this: (1) The first of the last two beats is pronounced more emphatically than the second due to rhythm and (2) the way the lips close in preparation for pronouncing the fricative *"s"* change

the quality of the second part of *"o"* making it more like an *"ah"* sound. Previously, we alluded to the clarity of this signal (CDs compared to older renditions on vinyl records) and how its spectral richness provide for a variety of spectral observations that are associated with sound perception. This is a good example of that.

2.7 Normalized Centroid vs *RMS* Amplitude

The root–mean–square amplitude (RMS) of Equation (2.41), with its time component, can be plotted against the brightness of the sound *"s"*. The spectral centroid distribution over the RMS in Figure (2.9a) reveals a high brightness at low amplitude at the beginning of time, which then atones and slowly raises again towards the end of time, an indication that harmonics pick up more power as time progresses. This can be seen in Figure (2.9b).



(a) **(b)**

Figure 2.9 – *Brightness of fricative* "s". (a) Normalized Centroid vs *RMS* Amplitude. (b) Amplitude vs Harmonic Number vs Time. Note how low-harmonic amplitudes grows in energy as time progresses.

While the spectral centroid is perceptually associated with brightness, the timbre of the sounds seems to be associated with another metric, spectral irregularity.

2.8 Spectral Irregularity and Inharmonic Partials

There is no direct evidence to tie spectral irregularity to a sound's timbre empirically, even though Horner et al. (2004) [30] found that if the average random spectral error was kept to 24%, listeners could distinguish a sound that has been spectrally altered to increase irregularity 78%–90% accurately. McAdams et al. (1999) [42] used spectrally smoothed data (basically, low–pass filtering the spectrum) and had subjects listened to the smoothed and original versions. The timbre in the smoothed version was altered enough so that only 4% of

the sounds were judged as identical. Even though there is a connection between perceptual timbre and spectral irregularity, the latter is often called a measure of "jaggedness" (e.g., Beauchamp (2007) [2]).

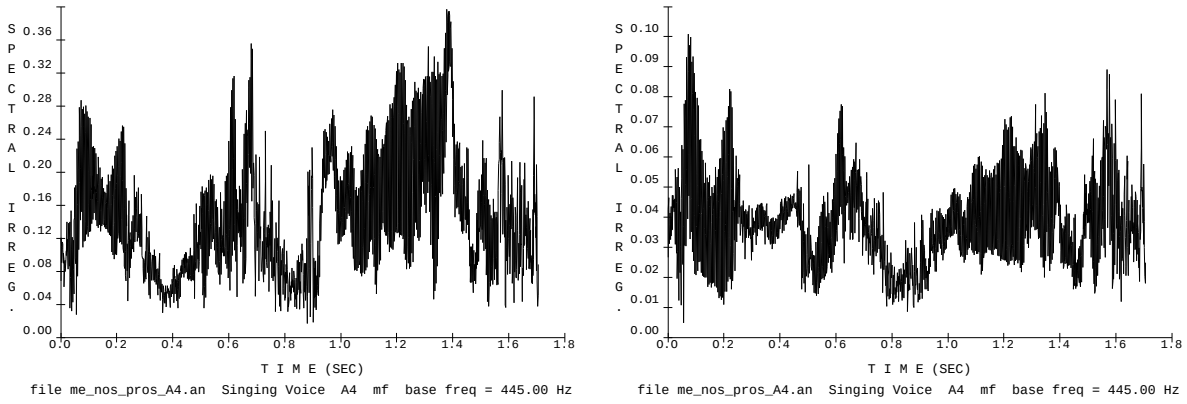
Let us define the spectrally smoothed harmonic amplitude as

$$\bar{A}_k(i) = [A_{k-1}(i) + A_k(i) + A_{k+1}(i)]/3$$

which is nothing more than the average of three consecutive amplitudes in a given ith analysis frame. To obtain an equation for spectral irregularity (SIR), let us further weigh the modulus of the original and smoothed differential with the amplitude sum over all harmonics but the fundamental and also normalize this metric by $A_{rms}(i)$ so that its value lies within the $[0, 1]$ interval (i.e., independent amplitude scaling), that is to say,

$$SIR(i) = \frac{\sum_{k=2}^{K-1} \bar{A}_k(i) \parallel A_k - \bar{A}_k(i) \parallel}{A_{rms}(i) \sum_{k=2}^{K-1} \bar{A}_k(i)} \tag{2.45}$$

Since spectral irregularity is the average absolute difference between the average of a harmonic amplitude and its two nearest neighbors and the harmonic amplitude itself, it is in essence a comparison of a spectrum to its smoothed version. We can artificially make $SIR(i)$ approach zero by averaging adjacent harmonics. Figure (2.10a) shows the spectral irregularity content over the entire time of our trisyllabic sound snippet and Figure (2.10b) shows an irregularity–reduced version. The author of this dissertation could clearly distinguish between the two versions when listened to.



(a) **(b)**

Figure 2.10 – *Spectral Irregularity vs Time.* (a) Spectral Irregularity vs Time. (b) Reduced Spectral Irregularity vs Time by averaging three adjacent harmonic amplitudes.

The idea of the phase vocoder is tightly tied with signal with harmonic partials. In practice, however, not all sounds (even musical ones) posses this property. Chimes, marimbas, xylophones, vibraphones, and cymbals are but a few indicative examples of partial inharmonicity. This dissertation will not discuss how models are generalized to include

situations like these. But even with harmonic musical signals, like piano, for instance, there exist some reasonable deviation or variance from the ideal partial. There is not much research on this specific topic for the singing voice, let alone Byzantine chant. In the case of piano, however, Lattard (1993) [35] and Fletcher (1964) [25] give modal frequency equations of a struck string (a plucked one would behave the same in this case) as

$$f_k = kf_0\sqrt{1 + B_k k^2} \approx kf_1[1 + (B_k/2)(k^2 - 1)] \tag{2.46}$$

where f_1 and f_0 are the fundamental and string frequency, respectively, and B_k is the so–called constant of inharmonicity. Let us define this deviation as usual by $\Delta f_k = f_k - kf_1$ and solve for the constant to obtain

$$B_k = \frac{2\Delta f_k}{(k^2 - 1)kf_1}, \quad k > 1. \tag{2.47}$$

As a loose benchmark, let us use B_k ranges for piano signals, which fall between 0.0001 and 0.001 if the fundamental is below 1 kHz and between 0.001 and 0.01 if the fundamental is above 1kHz. For our signal, $B_3 = 0.0015917$ and B_5 and B_6 have slightly lesser values [26]. Considering that the human voice is not by no means as controlled [27] as a plucked string, these values reveal greater insight on Nafpliotis' level of vocal singing mastery.

2.9 Steady Harmonics vs Vibrato sounds—The Singing Voice

Instruments like the piano are not able of producing vibrato sounds, but the singing voice and some other instruments like the violin are. Excessive vibrato "masks" the average frequency value of partials over time and makes the task of frequency trackers more laborious. The two frequency estimators we have considered so far, namely the phase vocoder and the McAulay–Quatieri methods, have been presented as "idealistic" and "realistic," respectively. The realistic one accounts for so much detail that sometimes tracking under certain conditions (consonant sounds, for example) is more challenging. However, due to its ability to threshold out spectral peaks below a given amplitude level (Figure (2.2) ITEM 2), and thus in essence de–noising the signal, is attractive when accounting for frequency and amplitude rapidly changing signals. Figure (2.11) below shows the amplitude for the first four harmonics for both methods over time. Figure (2.12) shows their normalized frequency deviation over time.

First notice that due to applying an amplitude threshold below which spectral peaks are ignored for the McAulay-Quatieri method, amplitudes are nullified at places [Figure (2.11)]. This compromise comes with a desirable return, however, i.e., frequency becomes much more legible and stable for the same method [Figure (2.12)] [28]. Also notice the stability of the normalized deviations among the various harmonic numbers in Figure (2.12)—they

[26] These values are not producible by SNDAN. Raw data were exported and analyzed in other software.

[27] The choice of the word "controlled" was intentional. There are factors other than the vocal folds at work affecting the final frequencies of a singer's envelope, and these alterations might be very intentional. In a sense, human voice frequencies have more "degrees of freedom" than mechanically induced frequencies.

[28] There is another reason for this fortunate result: the McAulay–Quatieri method uses a wider spectral window by definition.

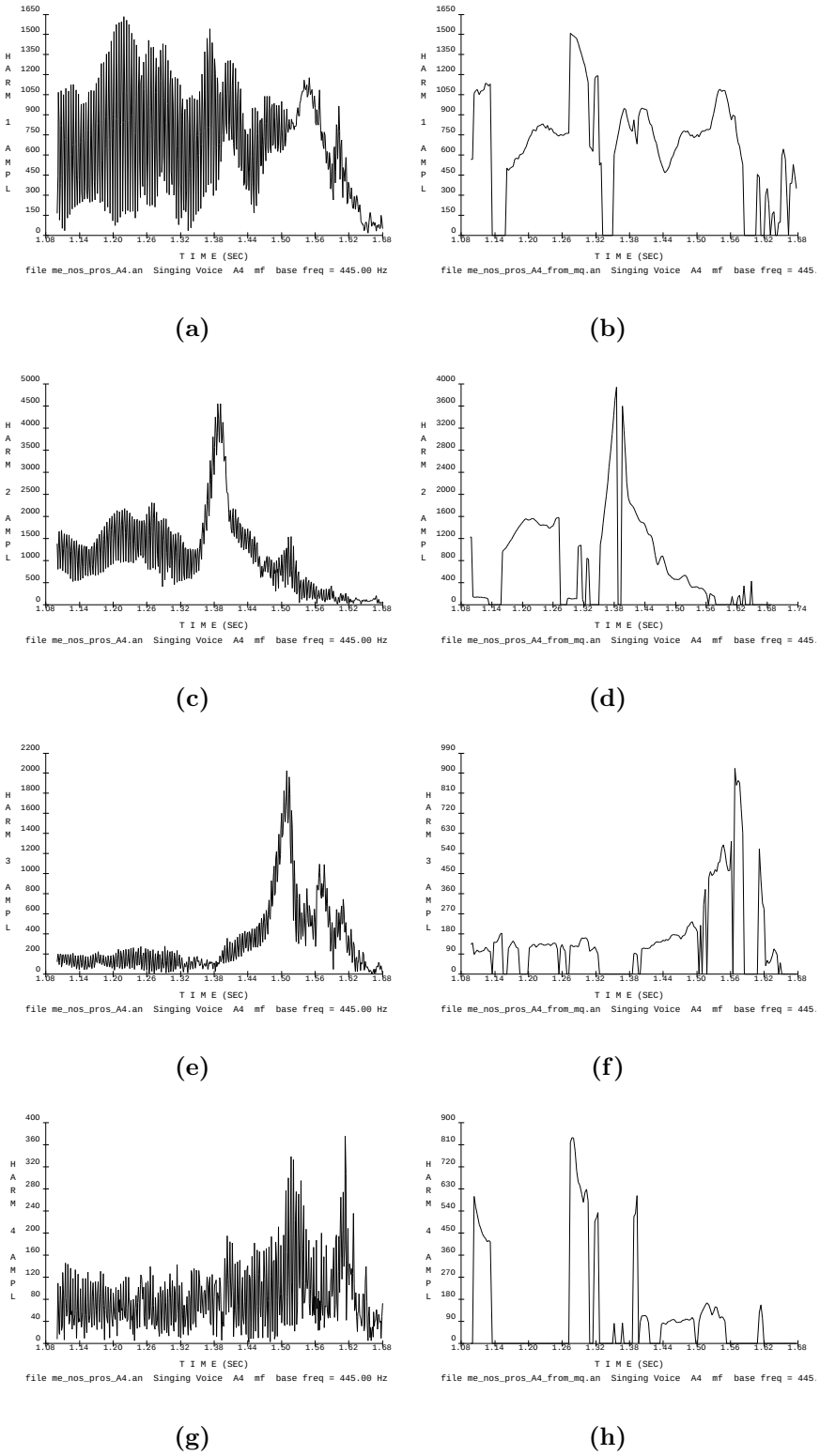


Figure 2.11 – *Harmonic Amplitude vs Time of vowel "oh."* (a), (c), (e), (g)—Phase vocoder method; (b), (d), (f), (h)—McAulay–Quatieri method.

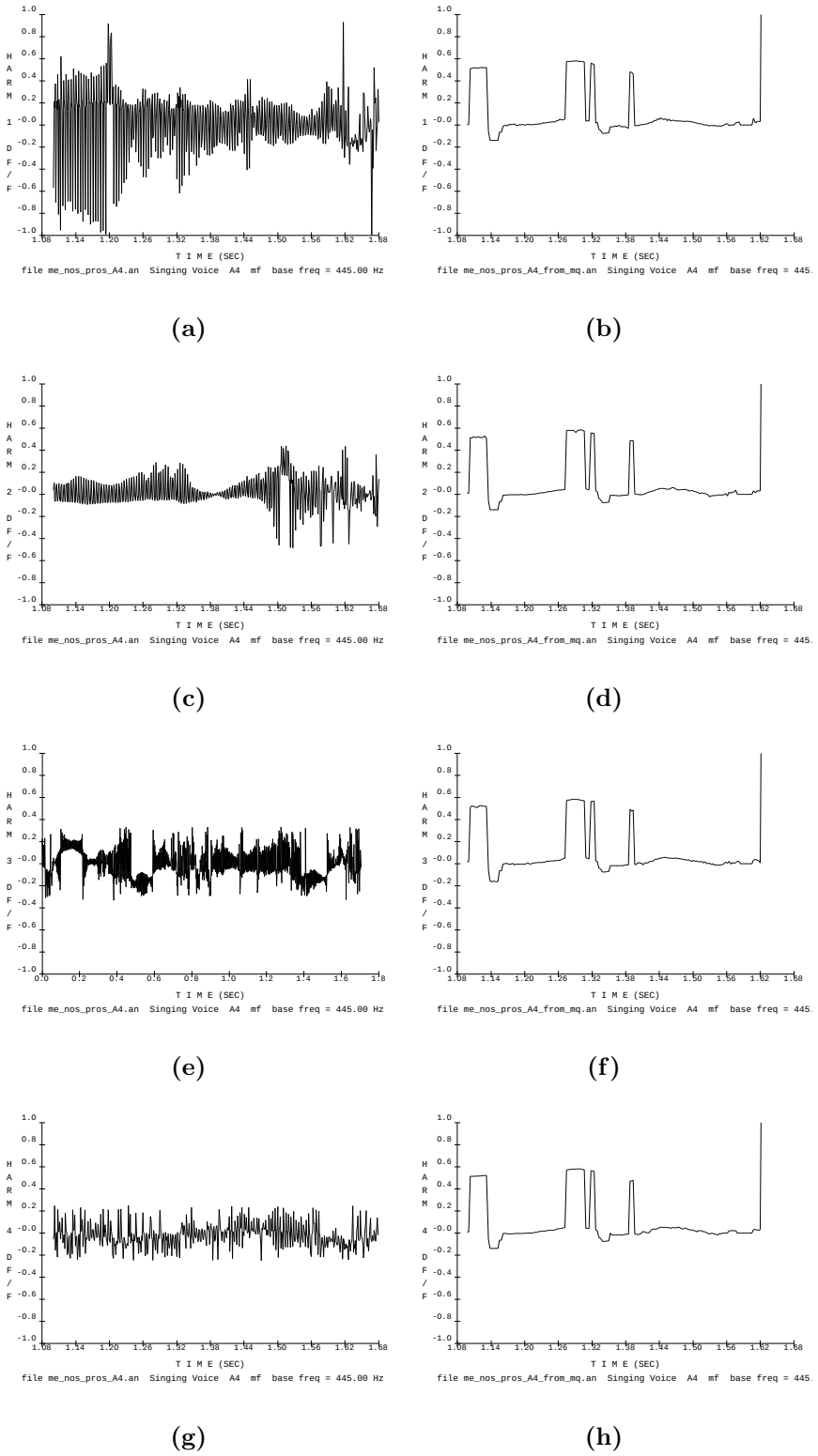


Figure 2.12 – *Harmonic Frequency Deviation ($\Delta f/f$) vs Time of vowel "oh."* (a), (c), (e), (g)—Phase vocoder method; (b), (d), (f), (h)—McAulay–Quatieri method.

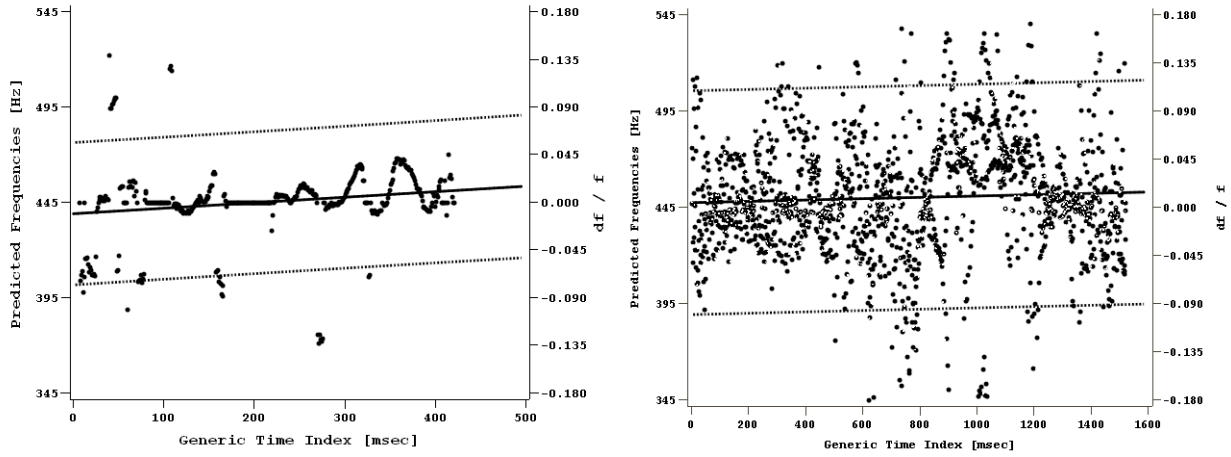


Figure 2.13 – *Generalized Linear Model.* Regression line (solid) and 95% confidence intervals (dashed) for frequencies (left vertical axis) and frequency residuals (right vertical axis). (a) Line of best fit for the McAulay-Quatieri method; (b) Line of best fit for the phase vocoder method.

are almost identical. Just by looking at these two graphs we should be convinced that the McAulay–Quatieri method is superior to the phase vocoder for singing voice applications such as this. As a matter of fact, the same is concluded by Beauchamp (2007) [2] where the data analyzed were a tenor's voice singing the vowel "ah" in G_3 (192 Hz) for about 3 seconds.

The fact that the McAulay–Quatieri method outperforms the phase vocoder for vowel snippets was hinted—even though not spelled out—back in Figure (2.4) and the discussion that went with it. It was seen there that the variance of this method is reduced by about a third with respect to that of the phase vocoder. We could take this idea and turn it around on itself to help us answer the interesting question "which analysis frequency f_a minimizes the frequency variance?" One could plot the residuals over a generic time index (generic because the peaks in Figure (2.12) have been filtered out), fit a least squares line to it which in turn would be used to "suggest" what value of f_a minimizes the least squares. This regression is shown on Figure (2.13) and it is also encouraged by SNDAN's online documentation. In fitting the line of best fit and averaging over the entire time the suggested values for f_a for the McAulay–Quatieri and the phase vocoder methods are 445.1795 and 460.9782, respectively [29]. This parameter is another metric that points to the superiority of the McAulay–Quatieri method, but is not bias-reducing strictly speaking, even if it may be tempting for some to call it that. I would disagree with the suggestion to re–do analysis using this new, "suggested" f_a. These two numbers help enhance our understanding by very little compared to the standard deviations we calculated earlier; both metrics are indicative of variability alone. Here it may be a good time to refer back to our discussion on precision and accuracy (see section (2.3.3). Strictly speaking, minimizing the residuals to fit another f_a that would lead to less variance

[29] $\hat{f}_{mq} = 438.8781 + 0.028889 \cdot (time\ index)$ and $\hat{f}_{pv} = 446.9851 + 0.002647 \cdot (time\ index)$

is certainly technically feasible (we just did, actually). However, this is similar to using a "self–predicting variable" to forecast future values. The variance explained by it is internal and inherent to the model to begin with, it does not add any new information nor it picks up any left–off variance in the system to account for. The information we should be looking for at this point should be bias–reducing, not merely variance–reducing. To reduce the bias (which is tantamount to increasing accuracy) one could look into external analysis (outside the model one has built), like for example, other spectral methods that would give us a better idea of where in the distribution f_a should fall.

Singing voice is capable of inducing vibrato effects, even though traditional Byzantine chant makes use of it much differently. Western music uses vibrato more frequently, more methodically (part of curriculum), and more prominently (larger frequency modulation) than Byzantine chant. In the context of singing voice, Byzantine chant is more steady than classical Western singing [30]. SNDAN offers the vibrato reducing package `fv`, which in effect replaces harmonic frequency deviations with kf_a, and a spectral irregularity reducing package `ri`, which applies smoothing filtering in a frame–by–frame fashion that could result in artificial vibrato reduction also. These packages will not be presented in this dissertation.

This concludes the presentation of mathematical theory of algorithms and methods that are related to SNDAN. The rest of this chapters presents algorithms that are more statistical in nature rather than time–varying.

2.10 Autoregression Models

Following the discussion of Stoica & Moses (2005) [80], an *autoregressive moving average* signal, abbreviated as ARMA(n, m), can be modeled as

$$y(t) = \frac{B(z)}{A(z)}\, e(t), \tag{2.48}$$

where $e(t)$ is normally distributed noise with zero mean and some variance σ^2 and B(z) and A(z) are polynomials in the unit delay operator z^{-k} operating on the signal $y(t)$ by means of lagging it by some constant k as in $z^{-k}y(t) \equiv y(t-k)$ expressible as

$$\begin{aligned} A(z) &= 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-n} \\ B(z) &= 1 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_n z^{-m}, \end{aligned}$$

where n and m give the order of the polynomial. Letting any one of the two polynomials be unity, by nullifying there order, results in either an *autoregressive* signal or a *moving average* signal, traditionally abbreviated as AR(n) and MA(m), respectively. If, therefore, $B(z) \equiv 1$, then the system is modeling an AR(n) signal, which is the topic of this section. The autoregressive model can be viewed as an all–pole infinite impulse response filter with

[30] Traditional chant is also more economical on amplitude variations as well. Some amplitude modulation exist naturally (due to lyric emphatics mostly), but *dynamics* usually ranging in Western music from *pianissimo* to *fortissimo* are not part of Byzantine music theory.

white noise as its input. Writing it in terms of a straight forward sum on the time–series signal we obtain

$$y(t) = \sum_{i=0}^{m} a_i \; y(t-i) + \epsilon_t. \tag{2.49}$$

Isolating the f_0 of a tone that is quasiperiodic and well–behaved in general, is often equivalent to resolving its spectral peak which, luckily, happens to be narrow–band—the peak one is trying to model is a narrow spike. This is why AR(n) models are useful in practice, because $B(z) \equiv 1$ is already restricted and zeroes of $A(z)$ are placed inside the unit circle [31]. To estimate the AR(n) coefficients, it is more convenient to introduce the *autocovariance sequence* of $y(t)$ as $r(k) = E\{y(t) \cdot y^*(t-k)\}$ and write a covariance structure equation for ARMA(n, m) as

$$r(k) + \sum_{i=1}^{n} a_i r(k-i) = 0, \quad \forall \; k > 0. \tag{2.50}$$

The next two subsections provide an AR(n) model parameter estimation method and an algorithm for its recursive solution.

2.10.1 Yule–Walker Equations

Short proofs that the *Yule–Walker* or *normal* equations is a solution to the AR(n) model can be readily found in Stoica & Moses (2005) [80], Kay (1993) [32], and Shumway & Stoffer (2006) [74], to mention a few, and therefore it will not be shown here. These are

$$\begin{bmatrix} r(0) & r(-1) & \dots & r(-n) \\ r(1) & r(0) & & \vdots \\ \vdots & & \ddots & r(-1) \\ r(n) & \dots & & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{2.51}$$

We can re–write Equation (2.51) as

$$R_{n+1} \begin{bmatrix} 1 \\ a_n \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \end{bmatrix} \tag{2.52}$$

to show explicitly that we are solving for the parameters a_n and σ^2.

2.10.1.1 Levinson–Durbin Algorithm

Since the order n of the model is not a known fact, but rather a trial–and-error process, one needs to test and assess model performance in a way that is both computationally efficient

[31] Given the impulse response function $H(z) = 1/A(z)$, any value of z that makes $H(z) \to \infty$ is a pole of $H(z)$ at that locus. Hence the name *all–pole* filter, since there is not way to obtain zeroes for $H(z)$.

Levinson–Durbin Algorithm

Initialize

$$a_1 = -\frac{\rho_1}{\rho_0} = k_1$$

$$\sigma^2 = \rho_0 - |\rho_1|^2/\rho_0$$

For $n = 1, \ldots, n_{max}$, do:

$$k_{n+1} = -\{\rho_{n+1} + \tilde{r}_n^* a_n\}/\sigma_n^2$$

$$\sigma_{n+1}^2 = \sigma_n^2(1 - |k_{k+1}|^2)$$

$$a_{n+1} = \begin{bmatrix} a_n \\ 0 \end{bmatrix} + k_{n+1} \begin{bmatrix} \tilde{a}_n \\ 1 \end{bmatrix}$$

Notational conventions:

- ρ_k means either $r(k)$, the unbiased autocovariance sequence, or $\hat{r}(k)$, the biased autocovariance sequence.
- $\tilde{x} = [x_n^*, \ldots, x_1^*]^T$
- $(\cdot)^*$ means the complex conjugate of a scalar, or the conjugate transpose of a vector.
- k_i are the *reflection coefficients*, and
- $-k_i$ are the *partial correlation coefficients*.

Figure 2.14 – Levinson–Durbin Algorithm.

and optimized to find the best model order. If one starts solving the Yule–Walker methods starting from an order of one and keep doing that iteratively up to a predefined maximum order, the computation will be cumbersome, in the order of n_{max}^4 flops [32]. The *Levinson–Durbin algorithm* is to the Yule–Walker Equations as the Fast Fourier Transform Algorithm is to the Fourier Transform: It reduces the machine calculation of $\{a_n, \sigma^2\}_{n=1}^{n=n_{max}}$ from about n_{max}^4 to n_{max}^2 flops, i.e., by two orders of magnitude. The algorithm is presented in Figure (2.14).

2.11 YIN

de Cheveigné & Kawahara (2002) [21] introduced the YIN algorithm based on the interplay of autocorrelation and cancellation and named after the oriental "yin and yang" philosophy of natural balance. The term "cancellation" refers to de Cheveigné's earlier auditory neural modeling work (de Cheveigné (1998) [20]) which presented a case of excitatory–inhibitory concurrence of signals (artificially replacing "coincidence" with "anti–coincidence") that when lagged successfully result in a "cancellation model of pitch perception" which

[32] Here n_{max} is some predefined value for the model order and by flops we mean number of complex multiplications and complex additions computed by the machine.

The YIN Algorithm

1. Take autocorrelation of signal x_t

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau}$$

   where τ is the lag, W is the window size, and t is the time index.

2. Minimize the difference function

$$d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2$$

   with respect to the lag τ by taking first derivative and setting to zero.

3. Calculate the cumulative mean normalized difference function

$$d'_t(\tau) = \begin{cases} 1 & \text{if } \tau = 0, \\ d_t(\tau) / [\tau^{-1} \sum_{j=1}^{\tau} d_t(j)] & \text{otherwise} \end{cases}$$

4. Impose an absolute threshold to avoid "octave errors." Find the smallest lag τ that also minimizes d'_t and also require that the partial is deeper than the given threshold. If this does not exist over a range, replace this local with the global minimum of the function.

5. Parabolic interpolation of each local minimum of d'_t and its immediate neighbors is fit by a parabola to optimize selection of dips.

6. Best local estimate replaces unstable d'_t values due to non–stationarity of data. This search for better f_0 estimates within short time intervals screens the spectrum for statistically unstable estimates and replaces them with an optimum one (not an average).

Figure 2.15 – The YIN Algorithm.

makes direct analogies to neural networking and autocorrelation algorithms. The algorithm gained well–deserved (in my humble opinion) popularity due to its flexibility of acoustic signals that can analyze (speech, music, underwater, etc.) and also due to the author's deep and remarkably unique inside into human perception and psychophysics. The YIN algorithm is presented in Figure (2.15).

Whereas the autocorrelation function maximizes the product of the signal with a delayed version of itself, the difference function minimizes the squared differences of the original and lagged signal. This reduced the error rates from about 10% to about 2% in the empirical demonstration of the paper using speech signals. Step 3 attempts to de–emphasize

erroneous higher harmonic picking instead of the fundamental ("octave errors," which the author rightfully deems as an unfortunate popular term, because, even though we are hopeful the partial above the fundamental is an integer multiple of it, in practice we often see it is not) and step 4 attempts to reduce the error of picking up subharmonics by means of a threshold which echoes the one used in the McAuley-Quatieri method (see Figure (2.2) ITEM 2).

2.12 Quinn & Fernandes Estimator

Before this ARMA–based frequency detection algorithm is presented, some fundamental definitions may prove useful; most of them come directly from Kay (1993) [32].

The expected value (mean) of an *unbiased estimator* $\hat{\theta}$ is the "true" θ for all possible values of θ, i.e.,

$$E(\hat{\theta}) = \theta \qquad \forall\ \theta. \tag{2.53}$$

An estimator $\hat{\theta}$ is *consistent* if the asymptotic probability that it is biased is zero, i.e., asymptotically unbiased. In other words, in the limit as the sample size increases, a consistent estimator is unbiased, in the sense that $\hat{\theta}$ approaches the "truth" (θ), loosely symbolically shown as $\hat{\theta} \to \theta$. More precisely,

$$\lim_{N\to\infty} Pr\{|\hat{\theta} - \theta| > \epsilon\} = 0 \qquad \forall\ \epsilon > 0, \tag{2.54}$$

which says that as sample size N grows without bound, the probability that the difference between the estimate and the "truth" (biasedness) is more than a positive number, is identically null; there is no chance that truth and estimation will not be exactly aligned as we consider more and more samples.

An *efficient estimator* meets the following two criteria asymptotically, i.e., as $N \to \infty$:

$$E\{\hat{\theta}\} \to \theta \tag{2.55}$$

$$var\{\hat{\theta}\} \to CRLB, \tag{2.56}$$

that is to say, the estimator is asymptotically unbiased (consistent) *and* approaches the Cramér–Rao Lower Bound (CRLB). An efficient unbiased estimator achieves the CRLB, i.e., it is a *Minimum Variance Unbiased* (MVU) estimator. In practice, the CRLB may not always be achievable, but if the MVU estimator exists, it is usually preferred, no matter if it is efficient or not. As a matter of fact, efficiency is measured by how close $\hat{\theta}$ comes to the CRLB. But to quantify efficiency further, we need to define the bound.

The *Cramér–Rao Lower Bound* is a theoretically derived variance for an unbiased estimator, below which empirically is impossible to go. It is the lowest possible variance the unbiased estimator can ever achieve. This mathematical device is useful in a number of practical ways: in the best case scenario we experimentally derive the variance of an estimator and it happens to be the same as the theoretical (CRLB) for all the values of the measurable parameter; then we know we found the *minimum variance unbiased* estimator. In the worst case scenario we compute the variance of the unbiased estimator from the data

and we find that it is really far away from the theoretical minimum (CRLB); in that case we know that this unbiased estimator is far from efficient, and the CRLB gives us a benchmark of how low it needs to go to start approaching reasonable efficiency (reasonableness here dictated by the application at hand). It can be defined as follows: The variance of any unbiased estimator $\hat{\theta}$ cannot be less than a lower bound $I(\theta)^{-1}$, i.e.,

$$var\{\hat{\theta}\} \geq \frac{1}{I(\theta)} \tag{2.57}$$

and $I(\theta)$ is the *Fisher information*, that is to say, knowledge extracted from the data about the estimated parameter. It is the negative of the expected value of the second derivative of the log–likelihood function with respect to the unknown parameter, i.e.,

$$I(\theta) = -E\left[\frac{\partial^2 ln\ p(\mathbf{x};\theta)}{\partial\theta^2}\right], \tag{2.58}$$

where $p(\mathbf{x};\theta)$ is the *likelihood function*, i.e., a probability density function that is comprised of a *fixed* data sample vector $\mathbf{x}$ and the parameter θ is *unknown*. The only condition that must be satisfied for the CRLB to be derivable (if it exists) is that this probability density function $p(\mathbf{x};\theta)$ is *regular*, i.e., the first derivative with respect to the unknown parameter of its logarithm exists and that its expectation value is zero over the whole range of the parameter.

Back to defining efficiency more succinctly, within the framework of the CRLB, efficiency is the ratio of the inverse of the Fisher information criterion and the variance of the estimator, or

$$efficiency(\hat{\theta}) = \frac{I(\theta)^{-1}}{var(\hat{\theta})} \tag{2.59}$$

and it obviously cannot be more than unity.

Off–the–shelf frequency estimation techniques based on ARMA models are more often than not asymptotically biased, i.e., inconsistent (for a proof see Quinn and Hannan (2002) [61]). In cases were the model is tweaked to be consistent, it is not likely at all it will be efficient. This is because the maximum likelihood estimators for frequency in general yield variances that are in general lower compared to ARMA–based frequency estimators by two orders of magnitude. Quinn and Fernandes (1991) [62] attempted to remedy some of the inconsistency and inefficiency drawbacks by starting out with an ARMA(2,2) model and trying to equate the two coefficients by iteratively making them closer and closer to each other.

The original model could be written as

$$y(t) - \beta y(t-1) + y(t-2) = x(t) - \alpha x(t-1) + x(t-2) \tag{2.60}$$

and we are trying to make $\alpha \to \beta$ or, if possible, equate them. Assuming that α is fixed and β is unknown, we estimate it by means of a Gaussian maximum likelihood function by

The Quinn & Fernandes Algorithm

1. Initialize $\alpha_1 = 2cos\hat{\lambda}_1$, with $\hat{\lambda}_1$ being initial estimate of true λ_0.
2. Calculate $\xi(t) = y(t) + \alpha_j \xi(t-1) - \xi(t-2)$, for $j \geq 1, t = 0, \ldots, T-1$ where initial $\xi(t) = 0$ for $t < 0$.
3. Calculate $\beta_j = \alpha_j + h_T(\alpha)$, where
$$h_T(\alpha) = 2 \frac{\sum_{t=0}^{T-1} y(t)\xi(t-1)}{\sum_{t=0}^{T-1} \xi^2(t-1)}$$
4. If $|\beta_j - \alpha_j| < \epsilon$, where ϵ is sufficiently small, then $\hat{\lambda} = cos^{-1}(\beta_j/2)$. Otherwise, let $\beta_j = \alpha_{j+1}$ and return to Step (2).

Figure 2.16 – The Quinn & Fernandes Algorithm.

minimizing

$$\sum_{t=0}^{T-1} x_{\alpha,\beta}^2 = \sum_{t=0}^{T-1} [\xi(t) - \beta\xi(t-1) + \xi(t-2)]^2 \tag{2.61}$$

with respect to β where we assumed that $x(t)$ is independent and identically distributed. Upon minimizing this quadratic in β (which may or may not minimize depending on the nature of the data), we set $\alpha = \alpha + h_T(\alpha)$ and the step is repeated until α and β are "sufficiently close." To put this into physical context, we use a trick that ties ARMA models and periodic signals. The ARMA(2,2) model

$$y(t) - 2cos\lambda\ y(t-1) + y(t-2) = x(t) - 2cos\lambda\ x(t-1) + x(t-2)$$

can filter out selected frequencies from a sinusoid

$$y(t) = Acos(\lambda t + \phi) + x(t) \tag{2.62}$$

and a representation of that would be the poles and zeroes placed on a complex plane unit circle diagram. In a sense, the autoregressive part acts like an infinite impulse response filter, and the moving average part like a finite impulse response filter; zeroes and poles are placed over the frequencies one wishes to filter out. If a time–series sequence $y(t)$ can be described by the sinusoid above, it also satisfies the ARMA model above. Then we can estimate λ in $\alpha = \alpha + h_T(\alpha)$ using $\alpha = 2cos\lambda$ [33]. Figure (2.16) summarizes the algorithm.

[33] Doubling the amplitude of this sinusoid (and of Step (2) in the algorithm) was an experimental, not theoretical, suggestion.

Pisarenko's Algorithm

1. Calculate the eigenvector x of the matrix C corresponding to its smallest eigenvalue.
2. Find the zeros of $x_1 + x_2 z + x_3 z^2$.
3. Assuming that these form a complex pair, estimate the fundamental frequency λ by the argument which is positive.

Figure 2.17 – Pisarenko's Algorithm.

2.13 Pisarenko Frequency Estimation

Pisarenko (1973) [55] is interesting in itself, as an idea, because it utilizes the autocovariance matrix of an autoregression of order two model [AR(2)] directly to derive the fundamental frequency. Consider the AR(2) model

$$y(t) + \beta_1 y(t-1) + \beta_2 y(t-2) = x(t)$$

with Yule-Walker estimates for $\beta = [\beta_1 \;\; \beta_2]^T$ being

$$\hat{\beta} = - \begin{bmatrix} C_0 & C_1 \\ C_1 & C_0 \end{bmatrix}^{-1} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}.$$

If our sequence is of the form of Equation (2.62), then it can be shown (Quinn & Hannan (2003) [61]) that

$$C_j \rightarrow \frac{A^2}{2} \, cos(j\lambda) + \gamma_j \tag{2.63}$$

where γ_j is the autocovariance sequence. In vector terms, if C_j is a matrix of the form

$$C = \begin{bmatrix} C_0 & C_1 & C_2 \\ C_1 & C_0 & C_1 \\ C_2 & C_1 & C_0 \end{bmatrix}$$

then it almost surely converges to

$$\frac{A^2}{2} \begin{bmatrix} 1 & cos\lambda & cos(2\lambda) \\ cos\lambda & 1 & cos\lambda \\ cos(2\lambda) & cos\lambda & 1 \end{bmatrix} + \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_1 & \gamma_0 & \gamma_1 \\ \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix}.$$

Pisarenko's method is based on observations on the eigenvalues and eigenvectors of

the first matrix above, which happens to be non–negative definite [34]. Its eigenvalues are 0, $2sin^2\lambda$, and $1 + 2cos^2\lambda$. The first eigenvalue of 0 has a corresponding eigenvector of $\psi = [1 \quad -2cos\lambda \quad 1]^T$ which can be written as a polynomial in z like $1 - 2z\ cos\lambda + z^2$. The zeros of this polynomial are $e^{\pm i\lambda}$. The algorithm, which suggests itself from these observations, is summarized in Figure (2.17).

Pisarenko's algorithm detects frequencies from multiple sinusoids and the dimension n of the autocovariance matrix is one more than twice the number of sinusoids to be detected present in the signal. So, for the example above, 2 sinusoids will be picked up. Additionally, if the eigenvector ψ above is also an eigenvector of the autocovariance matrix, Pisarenko's method yields consistent estimators for frequencies. For this to be the case, the sequence $x(t)$, usually considered—but need not be—the noise $e(t)$, is a stationary white Gaussian independent and identically distributed process with zero mean and σ^2 variance. That the consistency can be guaranteed in the limit is a particularly fortunate result with important practical implications (for details see Quinn & Hannan (2003) [61]). Even though the sample $y(t)$ is always zero–corrected (mean subtracted from it) before the sample autocovariance sequence is calculated, in this method it has no asymptotic effect. The estimator was originally developed to detect direction of arrival of a signal (bearing) in echolocation simulations where multiple collinear receptors collect time–series data.

2.14 *MU*ltiple *SI*gnal *C*haracterization (MUSIC)

Schmidt (1981) [68], and Schmidt (1986) [69] improved on Pisarenko's method, again leveraging the eigenvector structure of the sample autocovariance matrix. If Pisarenko's method used an autocovariance matrix of order $n_{PIS} \geq 2\alpha + 1$, where α is the number of sinusoids present in the signal to be detected, MUSIC uses $n_{MUSIC} \geq 2\alpha$. In other words, MUSIC is a more general method where $n_{PIS} = n_{MUSIC} + 1$.

In a logic similar to the one in section (2.13), let

$$\begin{aligned} e(\lambda) &= [\ 1 \quad e^{i\lambda} \quad \dots \quad e^{i(n_{MUSIC})\lambda}\ ] \\ &= [\ 1 \quad e^{i\lambda} \quad \dots \quad e^{i(K-1)\lambda}\ ] \end{aligned}$$

and let $\hat{P}_k$ be the normalized eigenvector of the autocovariance matrix $\Gamma_{(K-1)}$. Then the local minimizers of

$$\sum_{k=2r+1}^{K} |e^*(\lambda)\ \hat{P}_k|^2 \tag{2.64}$$

are the MUSIC frequency estimators. The reciprocal of Equation (2.64) is the celebrated MUSIC spectrum, with peaks denoting estimated frequencies. Figure (2.18) shows the MUSIC spectrum of our sample snippet data set used throughout this chapter.

Note that the heights of the peaks do not possess any physical meaning, since Equation (2.64) is rid of all amplitude information.

[34] Same as positive semidefinite, i.e., any matrix A which satisfies $\alpha^* A\alpha \geq 0$, where $(\cdot)^*$ denotes the conjugate transpose for complex or just the transpose for real matrices. If $\geq$ is replaced with $>$, A is positive definite, in that is definitely positive, no chance of being non–negative, which includes zero.

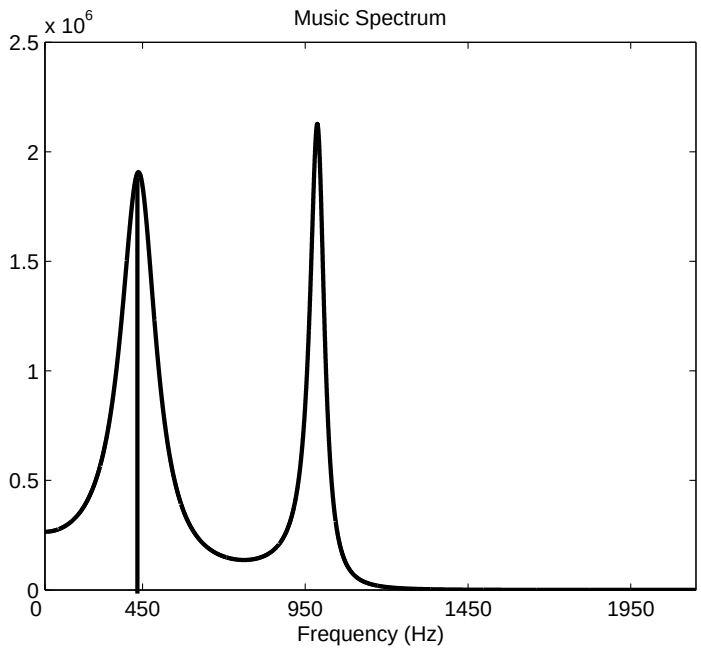


Figure 2.18 – Music Spectrum of the trisyllable "me–nos–pros" data sample used throughout this chapter for illustration.

2.15 Periodogram

In classic spectral estimation theory (see, for example, Stoica & Moses (2005) [80]), the *periodogram*, so named for its ability to uncover "hidden periodicities," (a term coined by Schuster A. (1898) [71]) is often derived directly from the fundamental definition of nonparametric power spectral density functions. For a derivation, discussion on its (high) variance, and possible modifications of it the reader is referred to Stoica and Moses, 2005 [80].

Another interesting approach on deriving the periodogram within the framework of the *general linear model* (that is more familiar to the statistician as opposed to the usual method that should resonate more with the engineer) is given in Quinn & Hannan (2003) [61]. It is common ground that all statistical models that observe linearity and additivity fall within this framework of the general linear model (see, for example, Tabachnick & Fidell (2007) [84] Chapter 17). This approach will be briefly outlined here primarily because it ties the periodogram with a host of other statistical methods used within *and* outside the natural sciences.

Consider a model that uses sinusoids as additive independent variables to predict the single dependent variable $y(t)$, with some white Gaussian noise added to it:

$$y(t) = \mu + \alpha\ cos(\omega t) + \beta\ sin(\omega t) + \varepsilon(t),$$

where μ is the overall grand mean of the series [35] , $\omega = 2\pi f$ is the angular frequency, and $\varepsilon(t)$ is noise. When ω is fixed (and for our purposes it is relatively unvarying) this is simply a univariate (in the sense of one criterion) linear regression model with sinusoidal co–variates [36].

Since this is a linear model, estimators can be given using the usual least squares method. It is

$$\begin{bmatrix} \hat{\mu} \\ \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} N & \sum_{t=0}^{N-1} cos(\omega t) & \sum_{t=0}^{N-1} sin(\omega t) \\ \sum_{t=0}^{N-1} cos(\omega t) & \sum_{t=0}^{N-1} cos^2(\omega t) & \sum_{t=0}^{N-1} sin(\omega t)cos(\omega t) \\ \sum_{t=0}^{N-1} sin(\omega t) & \sum_{t=0}^{N-1} sin(\omega t)cos(\omega t) & \sum_{t=0}^{N-1} sin^2(\omega t) \end{bmatrix} \begin{bmatrix} \sum_{t=0}^{N-1} y(t) \\ \sum_{t=0}^{N-1} y(t)cos(\omega t) \\ \sum_{t=0}^{N-1} y(t)sin(\omega t) \end{bmatrix}$$

or more succinctly

$$\begin{bmatrix} \hat{\mu} \\ \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = D^{-1}(\omega)\ E(\omega).$$

The sum of squares of the residuals is then given by

$$SS(\omega) = \sum_{t=0}^{N-1} y^2(t) - E^T(\omega)\ D^{-1}(\omega)\ E(\omega)$$

and upon maximizing this quantity with respect to the frequency ω and simplifying using

$$\sum_{t=0}^{N-1} = e^{i\omega t} = \begin{cases} \frac{e^{i\omega N}-1}{e^{i\omega}-1} & \text{if } e^{i\omega} \neq 1, \\ N & \text{if } e^{i\omega} = 1 \end{cases}$$

the regression sum of squares becomes

$$\begin{aligned} \hat{p}(\omega) &= \frac{2}{N}\left[\sum_{t=0}^{N-1} y(t)cos(\omega t)\right]^2 + \left[\sum_{t=0}^{N-1} y(t)sin(\omega t)\right]^2 \\ &= \frac{2}{N}\left|\sum_{t=0}^{N-1} y(t)\ e^{i\omega t}\right|^2, \end{aligned} \tag{2.65}$$

[35] I never explicitly talked about mean–correcting the series before obtaining statistics like the autocovariance sequences or fitting models like ARMA(n,m), but this is the term that is subtracted from the data during the mean–correction stage. This grand mean, or "DC" term, is the one that shows as the annoying spike at the origin of spectra or makes the models asymptotically unstable. Its removal may have substantial effects on nonparametric but more often than not on parametric modeling. On non mean–corrected data, an ARMA model *with* parameters is equivalent (in the limit) to a mean–corrected ARMA model *without* parameters, especially when it comes to model efficiency. This topic is large enough for a dissertation to be dedicated to it alone and it will not be discussed in this paper.

[36] Quinn does not discuss the role of *multicollinearity* (how much more new prediction is added by adding the predictors in a step–wise, hierarchical fashion) in his book, but in my humble opinion, this is a topic that deserves some attention. A sine is but a shifted cosine, and there is bound to be significant overlap in the variance explained by one with respect to the other.

the well–known periodogram, which is in effect the sample–scaled square modulus of the discrete Fourier transform (of Equation (2.2)) of the series $\{y(t)\ ;\ t = 0, 1, 2, \ldots, N-1\}$.

Figure (2.19) shows the periodogram for the trisyllable sound snippet used throughout this chapter. The frequency axis was scaled to 2000 Hz for better visibility.

2.16 Quinn & Fernandes Filtered Periodogram—$\kappa_N(\lambda)$

The periodogram in Equation (2.65) is not robust to initial ω_0 estimation errors; if the fundamental frequency is erroneously estimated when $\hat{p}(\omega)$ is maximized via a usual method (line Newton's), the plot of $\hat{p}(\omega)$ over ω would exhibit sidelobes next to the fundamental.

Quinn & Fernandes (1991) [62], whose work was also used in section (2.12), windowed the periodogram, thus creating a smoother version of it that does not depend on N (or T in the original notation), has the same asymptotic behavior (central limit theorem) as the periodogram itself, and is more robust to initial inaccuracies of the fundamental. The result is better estimation with a smoothed, cleaner plot.

The filtered periodogram can be expressed through convolution with a window $\mu(\omega)$

$$\kappa_T(\lambda) = \int_{-\pi}^{\pi} \hat{p}(\omega)\ \mu(\lambda - \omega)\ d\omega \tag{2.66}$$

where

$$\mu(\omega) = \sum_{k=1}^{\infty} \frac{cos(k\omega)}{k} = -\frac{1}{2}\ log\left[4\ sin^2\left(\frac{\omega}{2}\right)\right]\ ; \qquad \omega \neq 0. \tag{2.67}$$

Figure (2.19) shows the $\kappa'_T(\omega)$ estimates and the filtered periodogram for the trisyllable sound snippet used throughout this chapter. The frequency axis was scaled to 2000 Hz for better visibility.

2.17 Quadratic Interpolation and Rife & Vincent Estimator

The Cooley–Tuckey (1965) [19] fast Fourier transform Radix 2 algorithm reduces the complex multiplications from an order of $\mathcal{O}(N^2)$ to an order of $\mathcal{O}(NlogN)$, but still both real and imaginary parts are kept in storage. When large amounts of data need to be stored, analyzed, and reported through displays (sonograms, spectrograms, etc.) or tabulated tables, it would be efficient to find a way to store half of the complex Fourier coefficients, i.e., their moduli. Quadratic interpolators try to fit a parabola through the actual empirical curve using three or more points on the actual curve.

Let the complex coefficients be symbolized by $w(\omega_j)$ and the modulus of this sequence be $|w(\omega_j)|^2$. Also note that the Fourier coefficients are calculated at discrete Fourier frequencies

$$\left\{2\pi\frac{j}{N};\ j = 0, 1, 2, \ldots, N-1\right\}.$$

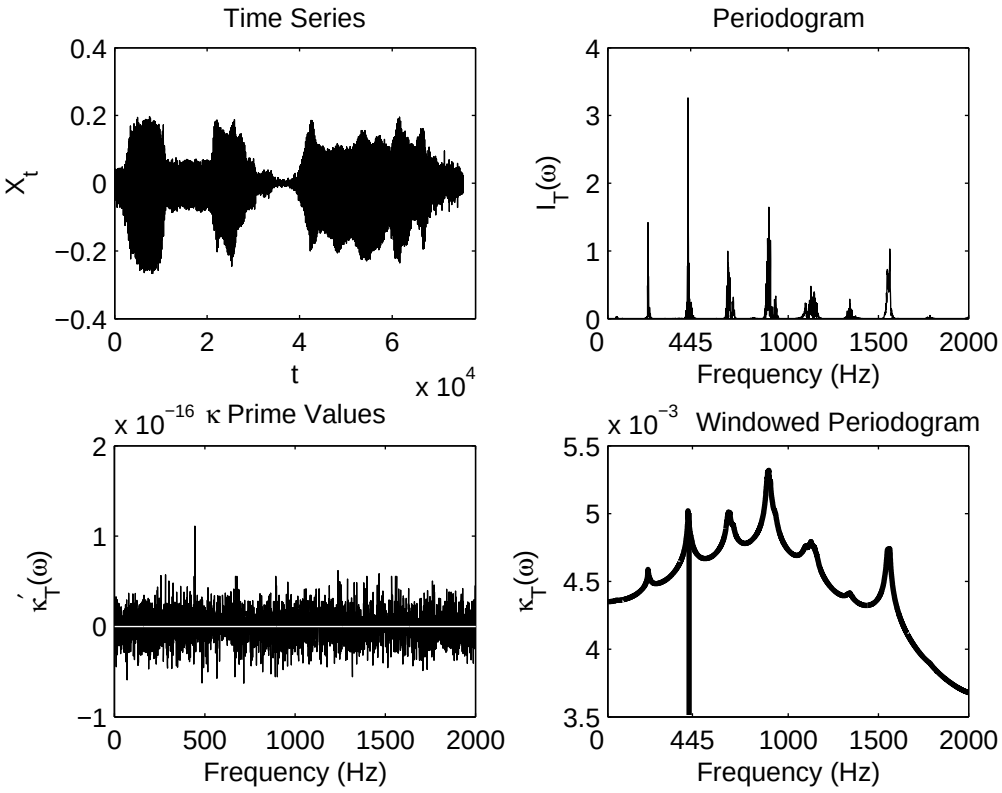


Figure 2.19 – The signal (X_t), the periodogram $(I_T(\omega))$, the $\kappa'_T(\omega)$ estimates (a normalized, non–smoothed version of the filtered periodogram estimates), and the filtered periodogram $[\kappa_T(\omega)]$ of the trisyllable "me–nos–pros" data sample used throughout this chapter for illustration.

Let us now pick three of these coefficients, the second one being at the origin of the index j and fit a quadratic through them, or in symbols

$$\left(j, \left| w\left(2\pi \frac{\hat{m}_N + j}{N} \right) \right|^2 \right); \qquad j = -1, 0, 1 \tag{2.68}$$

where $\hat{m}_N$ is the local [37] maximizer of $\mathcal{F}\left(2\pi\frac{m}{N}\right)$, where $\mathcal{F}(\cdot)$ denotes the Fourier transform in general. Equation (2.68) is known as a *quadratic interpolator*, and it can be shown that its bias is no better than than of the periodogram maximizer itself, in the order of $\mathcal{O}(N^{-1})$; its consistency is no better than that of the periodogram either.

One method that uses this kind of interpolation is that of Rife and Vincent (1970) [66]. Useful and extensive discussions on the analysis, application, interpretation, and statistical behavior of this algorithm can be found in Quinn (1997) [60] and Quinn & Hannan (2001) [61]. Consider the usual regression sinusoidal model

$$X_t = \mu + \alpha cos(\omega t + \phi) + \varepsilon_t \; ; \qquad t = 0, 1, 2, \ldots, N-1$$

[37] Local for the frame, but since this is a parabola, also global in the general sense.

Rife & Vincent Algorithm

1. Initialize $k_N = |Y_j^2|$, for $1 \leq j \leq [(N-1)/2]$.
2. If $|Y_{k_N+1}|^2 > |Y_{k_N-1}|^2$ then make $\hat{\alpha} = 1$ and -1 otherwise.
3. The estimator of ω is $\hat{\omega} = 2\pi(k_N + \hat{\delta})/N$, where

$$\hat{\delta} = \frac{\hat{\alpha} R_N}{1 + R_N}$$
$$R_N = \left| \frac{Y_{k_N+\hat{\alpha}}}{Y_{k_N}} \right|$$

Figure 2.20 – Rife & Vincent Algorithm.

where the parameter to be estimated is the frequency ω. Define two Fourier transforms, one for the model criterion and one for the model noise, which is white Gaussian as usual:

$$Y_j = \sum_{t=0}^{N-1} X_t \; e^{-i2\pi jt/N}$$
$$U_j = \sum_{t=0}^{N-1} \varepsilon_t \; e^{-i2\pi jt/N}.$$

Figure (2.20) uses the two definitions above to summarize the algorithm.

2.18 Conclusions

This Chapter presented a brief anthology of methods and algorithms found in the classical and more recent literature that either estimate the fundamental frequency or give useful quantitative insight into an otherwise vastly qualitative subject—sound perception.

Ten fundamental frequency algorithms were presented: (1) Phase vocoder, (2) McAulay–Quatieri, (3) Levinson–Durbin Algorithm, (4) YIN, (5) Quinn & Fernandes Estimator, (6) Pisarenko Frequency Estimator, (7) *MU*ltiple *SI*gnal *C*haracterization (MUSIC), (8) Periodogram, (9) Quinn & Fernandes Filtered Periodogram, and (10) Rife & Vincent Estimator.

Mathematical constructs and non–mathematical concepts that facilitate psychoacousitcal discussion include: (1) Frequency deviation, (2) Inharmonicity, (3) Precision and Accuracy, (4) Spectral centroid, (5) Normalized centroid versus root mean square amplitude, (6) Spectral irregularity, (7) inharmonic partials, and (8) Steady harmonics versus vibrato sounds on the singing voice.

The next Chapter deals with psychoacoustics, and its data–driven theoretical foundation on the pitch perception of musical tones. This will provide a basis of interpreting the fundamental frequency estimates tabulated in Chapter 4.

Chapter 3

Psychoacoustics

3.1 Introduction

This chapter addresses a fundamental question that is at the core of this dissertation: "For musically trained humans, what is the minimum distinguishable frequency difference." The previous chapter provided the theory of how to detect the fundamental frequency (a physical aspect) of sound. The next chapter will provide the experimental fundamental frequencies (still, a physical aspect) that were obtained from music chanted by Iakovos Nafpliotis and compare them to what music theory suggest chanters should be chanting. The question addressed in this chapter, therefore, links the two in the sense that any differences between theory and practice below the discernible frequency difference cut–off, have no real effect because the listener cannot distinguish them anyhow. The same argument goes for the performer. If the differences between the theory and practice are so minute that a trained musician cannot differentiate, clearly we cannot expect a performer to chant them. A similar argument could be made for differences between tone snippets across time as well.

Psychoacoustics, a child of psychophysics, is an interdisciplinary science at the intersection of physics and psychology. It is heavily empirical and data–driven, as opposed to some other branches of psychology that rely more on theory than experimentation. It has its roots in the early 19th century, save for the ancients, who touched on many of the topics discussed today, but not from the modern approach that we have become accustomed in this branch.

Relevant to our discussion on fundamental frequency versus pitch perception, psychoacoustics makes a clear distinction between the two: frequency lives in the space of physics, whereas pitch in the realm of humans' brains. The two are not the same by any stretch of the imagination, even though early physicists equated the two and even today some persist in doing so. Frequency is directly observable, pitch is indirectly measured (even with physiological experiments directly on the basilar membrane or ear bone conductivity). The variance of frequency observations is bound to be small (limited only by the instruments and calculation methods), whereas the variance of pitch is bound to be larger (not only from several measurements from one subject at one time or across time, but also inter–subject reliability becomes an issue as with any other perceptual measurement). The issue of low

data reliability leads to poor generalization of the results, that is to say, when in this dissertation inferences and conjectures are made based on pitch, those are all subject to high scrutiny and justified suspicion. In other words, even though one is forced to consider human perception when it comes to a topic such as the one explored in this dissertation, one *must* be aware of the plethora of drawbacks and pitfalls the concept of pitch inherently and naturally carry. With this word of caution in mind, we proceed carefully and based on data (rather than theory) to the best of the author's ability. Some basic perception theory, however, is presented briefly.

3.2 Theories of Pitch

What became collectively known as *theories of pitch* for some authors (for example, Rossing et al. (2002) [64], Stevens & Davis (1983) [79]), or *theories of hearing* for others (for example, Gulick (1971) [26]), refer to theories on how the ear physiologically resolves sound, that is to say, when air is excited, how do our brains perceive it as a sound.

It is customary to refer to the ear's ability to discriminate pitches in terms of the largest amount a frequency can deviate from itself and still be considered as the same tone. For example, if we take a tone of frequency f_{base} and we start frequency–modulating that tone up or down (call it f_{mod}), at some point a listener will identify f_{mod} as a different tone. The difference between these two (f_{mod} non–inclusive, of course) is what is known as the *just noticeable difference* (jnd) or in older literature as *difference limen*.

Basic physiology is needed to facilitate further discussion on theories of pitch. A mechanical sound wave enters the ear canal causing the eardrum to vibrate. The vibrations on the eardrum are conducted through the middle ear via the ossicles—three tiny bone structures, the last of which are the stapes. The stapes, in turn, oscillate the oval window which signify the beginning of the cochlea in the inner ear. The cochlea is filled with fluid, and the sound mechanical vibrations are now hydraulic mechanical vibrations. The cochlea has a membrane in it, called the basilar membrane, which takes these hydraulic pressures and transforms them into electrical pulses, firing with specific neurological patterns in what's known as the organ of Corti. The leap from the mechanical to the electrical happens by means of hair cells (celia) getting bent from the hydraulic pressure on the membrane which causes them to fire the electrical pulses. These pulses are then migrated into the brain via the auditory nerve, and electrochemical synapses are involved so that a sound perception is formed.

There are two classical theories of pitch, one known as the *frequency* theory and the other known as the *time* theory. There are a number of modern theories, as well. The next two subsections, briefly talk about both and their relevance to this dissertation.

3.2.1 Classical Theories of Pitch

The *frequency* (or *place*) theory of pitch has to do with where on the basilar membrane the excitation occurred (Gulick (1971) [26]). It is said to have originated from Helmholtz's

monumental work on cochlear experiments. The basilar membrane is divided into 24 regions, called *critical bands*, with each region being about 1.3 mm long and containing about 1300 neurons; each band acts as a unit for collecting sound data. The membrane itself is wider and loose at one end (close to the oval window) and narrow and stiff towards the other end (apex). Low frequency tones excite the wide portion of it and high frequency tones excite the narrow portion of it. The critical bands themselves have different frequency discrimination limits as well. The bands close to the wide, loose part have wider limits, thus making pitch discrimination of low–frequency tones less accurate than bands that pick up higher pitch tones at the other end of the membrane which happen to have smaller frequency deviation limits around the band's center frequency point. In other words, jnd is a function of at least one physical factor dictated by physiology, the frequency of the tone[1]. This theory views the ear as a spectrum analyzer.

The *time* (or *periodicity*) theory of pitch wants a time–series analysis applied to the firing pattern of the electrophysiological impulses in the organ of Corti (Gulick (1971) [26]). Shouten was not convinced that the frequency theory explains well–known phenomena like the case of the *missing fundamental* [2], the ear's ability to resolve the fundamental frequency and the brain's ability to process that information and "know" what the pitch of a tone is, even when the fundamental frequency was physically intentionally left out of the musical complex tone[3] (Rossing et al. (2002) [64]). Shouten called this missing lower part of the spectrum the "residue," and in a series of monumental experiments (in the first half of the 1940's) convinced the scientific community that it is not merely the physiological *place* on the basilar membrane that matters, but there must be some way for the brain to process the electrophysiological impulses from the organ of Corti and further into the brain. This pattern analysis was done in *time*, hence the name of the theory.

This means that there must exist a centralized unit in the brain that processes the signal in both domains. This centralized unit must be selectively using either frequency (spectra) or time (autocorrelation) data depending on the situation. It has been suggested that lower tones are processed primarily by the time domain analyzer and that higher tones are processed by frequency analyzers, with checks and balances between the two (modern theory). Echolocation, for example, could have an effect on how the time–frequency analyzers weigh input/output from one another. An interesting field, called *auditory computation* has since risen dealing with the mathematical neuromodeling of acoustic nerve firing patterns, but which in general encompasses a wide variety of models, from sound to perception. A short account of this fascinating field will be provided below.

[1] An interesting psychoacoustics experiment would be to quantify vibrato for lower and higher tones. One would expect that since the ear is less fine–tuned at the lower frequency span, that vibrato would be wider there compared to higher tones (directly proportional to the bandwidths). I also always thought that another consequence of this physiology would be that singing voice would be prone to more vibrato at the lower spectrum compared to instruments like violin at the same low frequencies, because a violinist can use his finger to "correct" this paradoxical limitation of her brain, but a vocalist has only but his ear to correct his ear. The signal does not pass through another "self–correcting" loop.

[2] This phenomenon was demonstrated about 100 years before by Seebeck (1841), but the paper is in German which I unfortunately do not speak. Shouten's work is what made this case widely known in the world of psychophysics.

[3] Complex here refers to many (hopefully harmonic) partials, as opposed to a pure tone which is only one wave with nothing above it.

3.2.2 Modern Theories of Pitch

Logic would suggest that from smaller axioms, lemmas, or theorems a scientist should be looking for a higher–level, unifying, or universal theory that would encompass the smaller parts of knowledge units and generalize them in as much simplicity as it is possible. At the same time (a second criterion, if you will) these general theories should be based on observation. In optics, the once distinct doctrines of corpuscle (Newton) and wave (Huygens) light theory were unified to what now explains most reality as the dual nature of light. Professor A. Einstein moved towards this unifying direction, albeit unsuccessfully, by attempting a "unified field theory," a theory that would bring gravity, nuclear forces, and electromagnetic theory under one umbrella[4].

It is, therefore, very unclear why modern theories of pitch perception (mainly those originating in the area of cognitive psychology) fixated on theories that not only do not seem to follow the logical idea of parsimonious unification and generalization, but also do not explain all of the phenomena we observe in empirical data (Rossing et al. (2002) [64]).

Since psychoacoustics is a vehicle for understanding the results of this dissertation and not the main topic of it, modern pitch theories en masse will not be discussed. One, however, will be outlined, that of Moore (2003) [45]. It follows Occam's razor and explains most observations, compared to competing theories.

Loven (2009) [36] is probably one of the sources that simplify and tailor Moore's theory (Moore (2003) [45]) in a practical way that fits our needs. It is a three–layer theory with frequency, time, and adjustment as its three layers. The first two layers have two levels each and the single–level final is a consideration of other variables to adjust the different weights for the first two layers in addition to considering some new information. The theory is summarized in Figure (3.1).

3.3 Auditory Computation

We attempt to present oversimplified aspects of the so–called *auditory computation* in hopes that this will illustrate how classical signal processing methods are used in signal transmission models constructed by this relatively new branch of psychoacoustics (Hawkins (1996) [28]). Topics are restricted to the limited physiology presented so far.

We talked about how the frequency or place theory requires to model excitation patterns on the basilar membrane. The spatiotemporal patterns of displacement along this membrane, then, could be modeled via a convolution of the time signal and a linear impulse response of the "cochlear filter." The notation usually uses the subscripts s and t to denote spatial and temporal components. Thus, the displacements $y(t;s)$ at locus s and time t can be approximated by

$$y(t;s) = h(t;s) * x(t), \tag{3.1}$$

where $x(t)$ is the input signal and $h(t;s)$ is the impulse response of the cochlear filter.

[4]Professor Einstein's ideas on this topic can be found in a compilation of some of his speeches in *Out of My Later Years* (1950), Philosophical Library, New York.

Moore's (2003) Modern Theory of Pitch

Frequency Theory Layer

1. *Critical band mechanism.* Center frequencies of critical bandwidths are set equal to frequency components in the spectrum of the incoming signal.

2. *Neural transduction.* Critical bands along the basilar membrane (don't have to be adjacent) fire electric pulses for the duration of the signal's excitation on that place, with some bands being turned on or off depending how much energy has accumulated within each band. Think of it as a very dynamic situation where mechanical excitation happens, a critical cut–off is reached, a specific neurophysiological structure is fired, and then the band comes to rest. Bands engage in this dynamic on/off pattern continuously.

Time Theory Layer

3. *Spike rate analysis.* Time–series analysis of neural response by means of autocorrelation is performed on each pattern across critical bands. Each band codes its activity in neural memory.

4. *Spike rate comparison.* Coded firing rate patterns are analyzed as new ones continue to flow in, and a backwards–time model keeps updating registries of activities, but this time across bands.

Adjustment & Decision Theory Layer

5. *Adjust existing info and incorporate new.* If similar firing rate codings are found along the membrane, and this is most probably the case, the brain adjusts how much the frequency components should adjust to compensate for the time analysis' indecisiveness. Memory, cognition, experience, attention characteristics, etc. of the listener as well as stimuli–driven variables are all integrated here. This part of the model has higher order cognitive characteristics, like parallel processing and ability to learn and adapt continuously as data flows in. In other words, if the first two layers were pure analysis steps that drew upon time–frequency domain theory, the model is drawing from artificial intelligence and machine learning theory, in the sense that it teaches itself to become better with more data flow.

Figure 3.1 – *Moore's (2003) Modern Theory of Pitch.* The input to this model is a music tone and the output is pitch perception.

These mechanical waves need to be converted into electric firing impulse trains during the *transduction stage*. The bending of the celia by the hydraulic forces exerted upon them that are proportional to $y(t;s)$, allow ionic currents to flow into the hair cells through nonlinear chemical channels (celia coupling), which in turn generate receptor potentials. To model those potentials mathematically, let us define $c(t)$ as the impulse response of the *celia coupling stage*, so that its convolution with the spatiotemporal displacements $y(t;s)$ will yield the so–called output of the *fluid–celia coupling*, i.e., $y_c(t;s) = y(t;s) * c(t)$. Further, let $w(t)$ to be a temporal smoothing window acting as a low–pass filter as a result of the hair–cell membrane and some function $g(\cdot)$ be an instantaneous nonlinearity that can take on the form of any nonlinear function depending on the emphasis the model wants to assign to it. The receptor potentials are then

$$r(t;s) = g[y_c(t;s)] * w(t). \tag{3.2}$$

The above equation can be modified to reflect detailed experimental models, for example, and enhanced to include stochastic firing patters on the auditory nerve with linear or nonlinear adaptations.

3.4 Factors Affecting Pitch Perception

We hinted above that frequency is one factor that affects jnd's and, therefore, pitch perception. It is also by far the strongest predictor of pitch, even though there are no experimental evidence on its relative strength compared to other factors. The second most influential factor for pitch perception seems to be sound intensity, even though more recent experiments seem to diminish the once higher emphasis intensity was given with respect to pitch perception.

3.4.1 Frequency and Pitch Perception

Frequency is by far the most influential component when humans judge the pitch of a musical tone. An example of how strong of a predictor the fundamental frequency is in perceiving a tone's pitch is shown in Figure (3.2). The data is partially based[5] on Zwicker et al. (1957) [91] and is also cited in Rossing et al. (2002) [64].

The dashed line is what was given in the original data. It makes sense to see that at higher frequencies the jnd's (indicated here by Δf in units of Hertz) are slightly larger. For example, approximately speaking, at 300 Hz the jnd is 2.5 Hz, but at 800 Hz the jnd is 4 Hz. This is in accordance to the critical bands being more finely tuned at the higher end of the frequency spectrum. Of course, we need to think about it proportionally, in relative terms. In absolute terms, one would be tempted to think that at higher frequencies our ears are doing a worse job discriminating frequencies than at lower frequencies. After all, from about 200 Hz to 1,000 Hz, jnd's increase approximately linearly.

[5]By "partially" we mean that only the frequencies of interest to us were retained in this view. This would justify the use of another researchers' data under the "fair use" of copyright law, otherwise permissions had to be obtained to display them here.

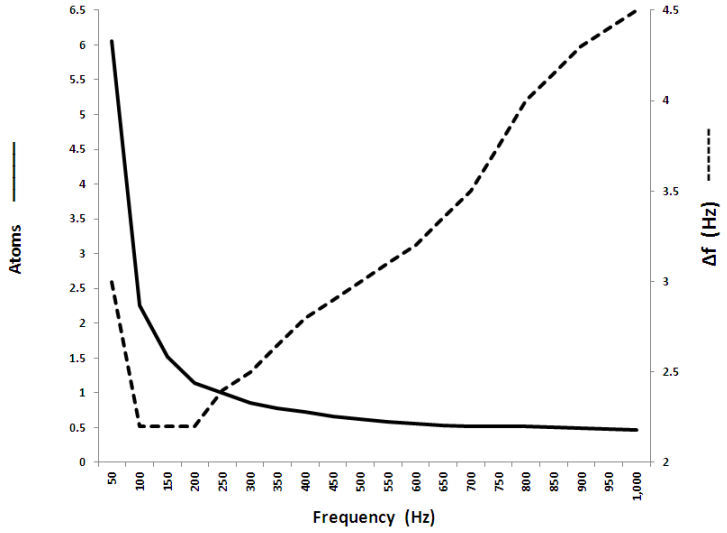


Figure 3.2 – *Pure tone ear frequency resolution—One Study.* Solid line shows human ear resolution (*just noticeable differences* or *difference limens*) of pure tones in units of atoms as a function of frequency at a constant sound level of 80 dB. Dashed line shows the same data, but as frequency differences in units of Hertz as a function of frequency. (*Source: Dashed line is based on partial data from Zwicker, Flottorp, and Stevens (1957) [91]. Solid line is a direct transformation of the dashed line.*)

We said before that we, humans, do not think in terms of jnd's or difference limens, but in terms of atoms or cents. We need to take the linearity out of this function my means of anti–logarithms. When we plot jnd's in terms of atoms, we used the usual formula of Equation (1.5) to plot the same data in a more friendly way. This is shown in Figure (3.2) as the solid line. We observe that our ability to discriminate pitch improves exponentially as a function of increased frequency.

This plot is used as an illustration of the effect of frequency on pitch perception and how a simple transformation renders the same data more understandable. A more comprehensive review will be given in a subsequent section of this chapter.

3.4.2 Intensity and Pitch Perception

Experiments by Stevens (1983) [79] as early as the mid–1930's found that as *intensity* ranged from 40 to 90 dB, a modest increase, pitch perception was altered by as much as 12%, two whole semitones. If this were true, the effects of orchestral dynamic changes would be detrimental to the average listener. It is now widely known that even for pure tones intensity has a much lesser effect, and for complex tones the effect is even smaller (Rossing et al. (2002) [64]). It is immaterial to explore intensity in detail, because it cannot be controlled for the purposes of our sample data. The intensity can be, of course, altered digitally for the sake of experimentation, but this would not help with our pitch detection objective. It is however, interesting to note that in large cathedrals where reverberation is very prominent, pipe organ music has reportedly changed pitch in a way that is proportionately inverse to

loudness. This is in accordance with what became known as *Stevens' rule*, who in his early experiments found that if intensity is increased, low–frequency tones seem to fall and high–frequency tones seem to rise in pitch. In the case of the pitch of pipe organ music, after loud chord ends (which signifies an abrupt drop in intensity), pitch seemed to have risen (Rossing et al. (2002) [64]).

3.4.3 Duration and Pitch Perception

If the signal *duration* is below 10 ms it is perceived as a click; even when the duration is up to 25 ms, the pitch perception is weak (Rossing et al. (2002) [64]—the data are partially based on experiments by Bürck, Kotowski, and Litchte (1935), but the paper is German, which I do not speak.). This implication should not affect our methodology of snippet concatenation, first because no snippet duration is even close to the 25 ms (the shorter ones are longer by about a factor of ten), and second, even if snippets were short, here we do not directly use them for pitch perception. Rather the concatenation of snippets constitute sound long enough for our detectors to reliably estimate the fundamental.

3.4.4 Other Factors Affecting Pitch Perception

Some subjects reported that some tones have an apparent "largeness" or "extensiveness" and this is termed by Stevens & Davis (1983) [79] as *volume*. Stevens also discusses another sound quality that might affect pitch, *density*, a feeling that a tone is "compact" and "tight." *Brightness* is another factor discussed in Stevens, which we attempted to quantify in section (2.6) by way of the spectral centroid. *Timbre*, a feeling that a sound has a certain "warmth" or "softness" to it, is yet another qualitative factor that some sound analysts attempt to quantify using special spectral envelops.

It should be mentioned, however, that none of the above factors affect pitch perception significantly enough to justify psychoacoustical investigation within the boundaries of this dissertation. In the next section we concentrate on frequency as the main factor affecting pitch discrimination.

3.5 Just Noticeable Difference in Psychoacoustic Literature

This section directly addresses the question we posed in the introduction to this Chapter, namely, "For musically trained humans, what is the minimum distinguishable frequency difference." Our quest to answer this question took us on a century–long journey, from Helmholtz (1863) to the 1970's. As scientific method matured and more educated questions were formed, technology was catching up providing the means to perform more refined experimentation. The interplay between experimental design and technological advances saturated in the 1970's, at least from my limited understanding of the issue, and the issue is now considered sufficiently resolved. Or at least, until a more precise definition of pitch discrimination is formulated and direct measurements on human brains is permissively non–invasive for scientist to pursue. I believe that only with this level of accuracy one will

be satisfied with a numerical answer to this question; of course, if that level of accuracy is achieved based on very refined methods, the definition of pitch discrimination itself—and its practicality—will become restrictive. It seems that with issues of this nature, as the empirical approaches the theoretical sufficiently closely, the trade–off between evidence and its practical use becomes harder to define.

3.5.1 Literature Review

In the previous paragraph it was stated that evidence in the 1970's are sufficient to answer the posed question. This doesn't mean that there is no more to be done to enhance our understanding of pitch discrimination; it means, however, that from post 1980's literature, to the best of the author's ability, no empirical evidence could be found to justify any significant addition to the already existing knowledge.

There are many experiments that could be performed on pitch discrimination, especially with complex tones in non–controlled settings. It is possible to analyze existing Byzantine music pieces and deduct valuable psychoacoustical results. Such analysis does not even require experimentally controlled data collection from subjects in the traditional way[6].

The following books were carefully reviewed and even though few of them cite or present for the first time evidence that could be useful to answering our question, most of them either cite the same older sources or paraphrase the results in one way or another (as we will do in this dissertation later).

1. O'Callaghan, C., and Nudds, M. (Editors) (2009). *Sounds and perception—New philosophical essays*, Oxford University Press.

2. Benson, D., J. (2008). *Music—A mathematical offering*, 3rd printing, Cambridge University Press.

3. Yost, W., A. (2007). *Fundamentals of hearing—An Introduction*, 5th Edition, Academic Press.

4. Lass, N. J. and Woodford, C., M. (2007). *Hearing Science Fundamentals*, Mosby Elsevier, St. Louis, MO.

5. Fastl, H., Zwicker, E. (2007). *Psychoacoustics—Facts and Models*, 3rd Edition, Springer, New York.

6. Plack et al. (2005). *Pitch—Neural coding and perception*, Springer Handbook of Auditory Research, New york.

7. Neuhoff, J. G. (Editor) (2004). *Ecological Psychoacoustics*, Elsevier Inc., of Elsevier Academic Press, London.

[6]Due to my background in Quantitative Psychology, I appreciate the statistical elegance experimental design affords. Generalizability of results is also a positive outcome. There are practical situations, however, that could justify observational (versus experimental) results.

8. Plomp, R. (2002). *The intelligent ear—On the nature of sound perception*, Lawrence Erlbaum Associates, Publishers, London.

9. Hall, D, E. (2002). *Musical Acoustics*, 3rd Edition, Brooks/Cole, California.

10. Howard, D., M. and Angus, J. (2001). *Acoustics and Psychoacoustics*, 2nd Edition, Focal Press, Woburn, MA.

11. Warren, R., M. (1999). *Auditory perception—A new analysis and synthesis*, Cambridge University Press.

Along with the above books, many journal articles were considered over the course of the last six years[7]. Other than those cited explicitly, none of them was used directly.

3.5.2 Pure vs Complex Tones

About complex tones we know much less compared to what we know about pure tones. This is a universally known and accepted fact that is usually mentioned only casually in most sources. Also commonplace seems to be the fact that for (musical or harmonic) complex tones pitch discrimination is much better than for pure tones. Plomp (1967) [58] (the same Plomp that published the book above, which provides discussions on his earlier findings) found that the first five harmonics of complex periodic sounds are most important in determining its pitch (actually for the range of frequencies we are interested in it is more like the second, third, and forth). A similar theoretical result from Moore et al. (1985) [47] was used in constructing the definition of weighted inharmonicity in Equation (2.11)[8]. The fact that complex tones yield lower jnd's is a fortunate one for us and it does make intuitive sense theoretically. The higher harmonics fall in the more finely tuned critical bands (whose bandwidths are narrower, closer to the apex of the membrane) and since the ear is back–fitting from the first five harmonics to form the perception of pitch, the perception is also more accurate as opposed to just using the one pure fundamental. This has special significance for us, because most of psychoacoustics data at our disposal are on pure tones.

Unfortunately, there is no data on pitch discrimination for complex tones usable for our purposes here. Moore et al. (2006) [46] worked on complex tone pitch discrimination, but the frequencies used were all above 2,000 Hz. Micheyl et al. (2010) [44] investigated pitch

[7]Pitch discrimination never went out of interest. A recent study (Dai, H. and Micheyl, C. (July 2011), Psychometric functions for pure–tone frequency discrimination, *J. Acoust. Soc. Am. 130* (1), 263–272), for example, concluded that linear fit on the $d' = \left(\frac{\Delta f/f}{\alpha}\right)$ versus Δf function is good enough, and previous papers that did not test for nonlinearities didn't really need to do so. This result actually directly supports my results in Figure (3.2). It is actually the reverse of what we did here: here we used exponents (or algorithms) to delineate the function, in this paper they use them to force–fit it as linear, with β being the slope of the line.

[8]In our signal we observe forty–nine (49) harmonics. We mentioned earlier that older recordings (before Alygizakis' CDs [1]) only a handful were present. This doesn't mean that it would not be possible to track the fundamental in the old recording. After all, the fundamental itself was there. But, we are missing out on the higher harmonics which convey information about the signal that were sporadically mentioned in Chapter 2.

discrimination based on the fundamental versus virtual pitch (the inclusion of higher harmonics, with or without the fundamental present), and concluded that difference limens for complex tones do approximate the fundamental perception well and that harmonic complex tones produced a more consistent sensation of pitch than inharmonic ones. An interesting approach was taken by Feth (1974) [23], which was unfortunately inconclusive due to limited data and (in my opinion) methodological limitations that might have hindered statistical power. As if the issue was put to rest from a psychophysical standpoint, others tried to tackle pitch of complex tones on a neurophysiological level, by measuring neural activity spike patterns in the auditory nerve (Cedolin (2005) [16]). Of course, it is possible to conduct a meta–analysis on extant literature and extrapolate some approximate values for complex tone jnd's customized to fit our needs in this dissertation, but the uncertainty of that kind of study would be greater than simply using the literature for pure tones as a benchmark.

Since we know complex tones make for better pitch discrimination relative to pure tones, we can use the pure tone results as a "worst case scenario" benchmark. Think about it as our own special "*pure tone higher bound*," a pun for the Cramér–Rao Lower Bound. Discrimination cannot get worse for complex tones that what it already is for pure tones, and pure tones are all we have to rely on. The jnd's cannot be larger, thus the "higher bound."

3.5.3 Experimental Results

Pitch discrimination psychometric curves from different independent experiments are reported in Weir et al. (1977) [90] and Moore (2003) [45]. They are shown in Figure (3.3) along with the data from Zwicker et al. (1957) [91] (also cited by Rossing et al. (2002) [64]), which we displayed in Figure (3.2). Two points to notice: (1) The same transformation of Figure (3.3) was applied in Figure (3.2) as well; and, (2) The results from Weir et al. (1977) [90] are widely cited even in very recent papers on pitch discrimination theoretical and mathematical modeling formulation. The fact that Moore (2003) [45] himself uses these results is an indication of its prominence[9].

Casual inspection of Figure (3.3) reveals the tremendous variation in the jnd's of the lower frequencies (250 Hz in the graph) and appreciable variance at the higher frequencies (1,000 Hz in graph). Maybe critical bands in the higher frequencies possessing more discriminatory power accounts partly for that. But even so, the variance[10] between experimental results is just too high for something that sounds so simplistically straight–forward as a jnd. The next subsection attempts to provide supporting information as to why this might be happening.

[9]This is a personal and therefore subjective opinion, that is why I put it in a footnote. Brian C. J. Moore (University of Cambridge, UK) is arguably the most prominent psychoacoustician alive and definitely a pioneer and leader in this field for the last 50 years. His book is packed with data–driven inferences, along with a wealth of actual data results drawn from his research, that again, has been dominant in this field for half a century now, along with the most reliable information from contemporary colleagues. Note how the lowest jnd curve is from his 1973 experiments. He continues to pursue the difficult task of complex tone pitch perception to this day. This is not the same Moore who co–authored Thomas Rossing's latest book [64].

[10]By variance here we do not mean the statistical variance, but the variability in practical terms. Older experiments (1931) exhibit a 19–cent jnd, which is about a fifth of a semitone.

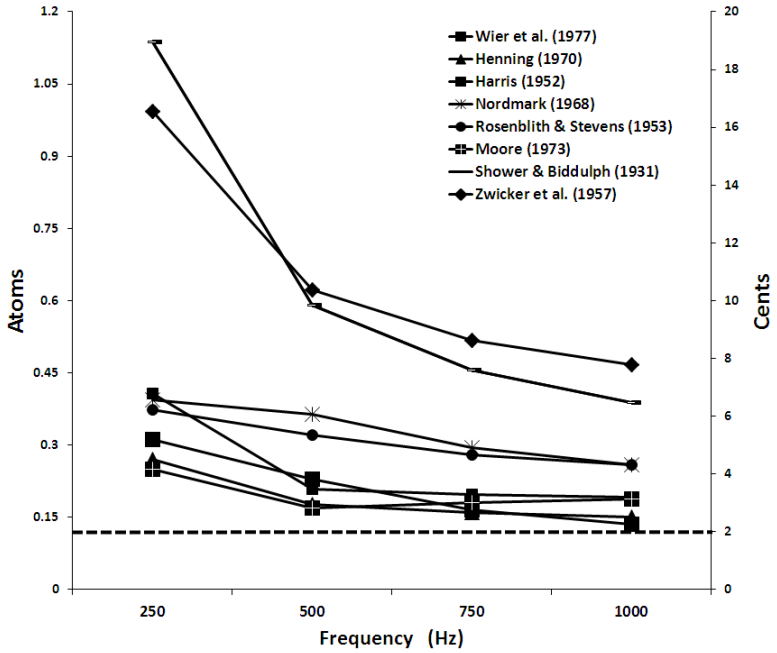


Figure 3.3 – *Pure tone ear frequency resolution—Meta–analysis.* Lines show human ear resolution (*just noticeable differences* or *difference limens*) of pure tones in both atoms and cents (see section (1.6) for more details) as a function of frequency at different sound levels that have been constant for each experiment. All data except for Zwicker, Flottorp, and Stevens (1957) [91] are taken from Moore (2003) [45] and Weir et al. (1977) [90]. The data is only partial (original graphs include a wider range of frequencies and more data points even within the 250–1,000 Hz range we display here and transformed to our needs (using Equation (1.5)—see text for explanation). The dashed horizontal line shows a single number jnd cut–off. However, a more reasonable measure would be frequency–adaptive, similar to the idea of critical bands of the frequency theory.

3.5.3.1 Factors Affecting Differences Among Pitch Discrimination Experiments

Differences between experimental results is a function of many parameters. Each experiment shown in Figure (3.3) used a different combination of these parameters. Some obvious ones are mentioned here, without claiming this to be a comprehensive meta–analytic review of psychoacoustical literature. Moore et at. (2003) [45], Gulick (1971) [26], and Plomp (2002) [57] provide many examples of psychoacoustical experiments with different parameters. For experimental design in general, Keppel & Wickens (2004) [33] is a good reference.

Please keep in mind that any mix of these parameters might have been employed by the experimenter. Each parameter has two or more levels that an experimental psychologist can vary. In general, the newer rather than the older literature seems to be more complex in the sense of combinations of these levels. This can be a good and a bad thing. It is an advantage in the sense that since pitch is a highly subjective construct, susceptible to effects of such different parameter levels, the more comprehensive the inclusion of those levels, the more variance in the perception one can account for experimentally (and statistically in the data

analysis stage). The main disadvantage is that the more parameter levels the experimenter decides to include in the design a priori, the more noise is introduced weakening the signal to be detected and analyzed—signal here is the perception of pitch and what "really" drives this experimental design model and noise is all the participant inputs to the experiment that add no significant value to our understanding of pitch perception. A secondary disadvantage is the lack of agreement between experimental design and the statistical method used to analyze data obtained by the design. Typically, statistics are dictated (enhanced or limited) by the design. This is especially true in social science in contrast to natural science. There is abundant literature on the differences between observational and experimental data, and there is a great interest in the corporate world to leverage the various powerful experimental methods (typically used in social and financial settings) using purely observational data. The reason for that is the fact that in the industry observational data is everywhere and typically experimental data is nowhere.

Experimental methods changed longitudinally, and even at time cross–sections there are no golden industry standards. The way the individual signals are randomized across trials and across subjects, the way the tones (temporal or spectral portion of tones) are masked intentionally, the signal–to–noise ratio and if this was randomized across trials and subject as well as if it was white or colored noise, smooth transition from tone to tone by phase–locking or sinusoidal transition (to avoid Gibbs oscillations which alter perception), if signals are presented sequentially or simultaneously with one tone constant and the other frequency–modulated, how the intensity level is defined and how the cut-offs of this level were constructed in regards to how many dB above intensity level were used, if intensity was constant or varied across different tones and how it was randomized, were tones produced mechanically or electronically, was the audition monaural or binaural, and if binaural was it with phase or time differences between the onset of the stimuli, was the experiment two–alternative forced–choice or not, are but a few factors that affect results. *Statistical methods* are also another class of factors contributing to the huge variability in the results among experiments. Not only poor experimental design can hinder statistical inference in general (to the point of results being unusable when inferences about the population are built on samples), but small samples of subjects and/or trials can influence statistical power. A far less sophisticated example is the arbitrary decision of where exactly a difference in frequency is concluded as the "real" jnd. Early experiments (including psychophysical giants like Stevens, for example) decided that a jnd is where 50% of the subjects can identify tonal difference. Later on it was decided that since 50% is really like a coin flip decision (completely by chance), a 75% level would be more appropriate since it is halfway between chance and perfect agreement among subjects[11]. *Inter–subject variability* as well as *intra–subject variability* are so obvious that they do not warrant explanation, but by no means less important. All these classes of factors are of basic (albeit, fundamental) level that a person with no formal training in psychophysics (such as myself) can understand and provide here.

[11]It is interesting to see how long it takes practitioners to get on–board with with statistical theory. At the time that Stevens, again, a giant of his time and a leading authority in psychoacoustics, only to be matched in stature by Brian Moore maybe, was conducting his landmark experiments, the decision to use a purely–by–chance "point of inflection" was unjustified. A decade or more before all this the famous feud between Pearson and Fisher (circa 1925) about hypothesis testing and levels of significance should have been a very good guideline to follow.

One conclusion became clear enough throughout this research: There is no one single jnd cut–off number. The dashed line in Figure (3.3) is of conceptual, not practical, significance. But, a jnd criterion is still needed in practice. The most reasonable option is to adopt the lowest jnd among all pitch discrimination experiments at each frequency (as opposed to an average, median, or some other point estimate) as our proposed jnd solution. The logic is twofold: (1) Experimental results are based on pure tones, which we know impede pitch discrimination ability; and, (2) the assumption that the experiments yielding higher jnd's (the ear is doing worse in discriminating) is due to experimental limitations (be it methodological, statistical, technological, or any other kind). The first one is clear. It becomes even more evident when we think about Plomp's (1967) [58] experiment, and that the second to the fifth harmonic are bound to be in the more fine tune place on the basilar membrane. The second one is an assumption that actually works against what we were set to show (that differences among theory and practice are insignificant since they cannot be discriminated). We have full knowledge of the restrictive nature of this choice. However, one may start with a low baseline benchmark and reassess it to more liberal standards, as it is the case in many practical application of statistics, like the subjective critical region of a normal distribution, for example, in hypothesis testing and even statistical inference.

3.6 Just Noticeable Difference Proposed Customizations

The results of Figure (3.3), as we discussed above, vary greatly. Our decision to take the lowest of the experimental jnd's as our benchmark was the most sensible decision considering the data, but it does make for an extremely narrow jnd to work with especially with singing voice signals.

We feel it was the most sensible decision because adopting a larger than the minimum experimental jnd would be completely out of place and arbitrary, because finer results are usually associated with more sophisticated experiments. This may or may not be the case, but it is hard to argue against this point[12]. Arbitrary selection of a greater jnd benchmark should be put under intense scrutiny, and justifiably so. One would need to look into the actual methods and analyses that make the different experiments vary so greatly. This was only superficially touched upon in Subsection (3.5.3.1) with the intension to demonstrate how involved such an analysis would be. A deep meta–analysis of this sort would be not only outside the scope of this dissertation, but also not as fruitful as one would think might be in justifying conclusively why an experiment yielding greater jnd's is more suitable for our very specific application at hand. Probably the best way to address a customized jnd for the singing voice of Nafpliotis is to perform an experiment using Nafpliotis' voice as the signal. With the new advances in voice analysis and synthesis (with phase vocoder being one of them), this is not at all out of reach in the future.

The reason such narrow jnd's make it especially hard to use when our input signal is human voice is partly due to the inherent difficulty to estimate fundamental frequencies for human voice since it is such a complex sound generating system. The issue becomes even

[12] I believe that this is not the case, but convincing a dissertation committee would be extremely hard.

more complicated if one takes into account how the first handful of harmonics are processed by the brain to decide on a pitch value.

We feel it necessary to propose two measures that may partially alleviate some of the limited applicability of such narrow experimental jnd's. The first, called perceptual confidence intervals, is an idea and it is deeply rooted in the concept of critical bands found in psychoacoustic theory. The second, called acceptable performance difference, is a metric based Nafpliotis' data. Both aim to provide more reasonable jnd's customized for this dissertation which are not arbitrary.

3.6.1 Perceptual Confidence Intervals

We often talk about *statistical confidence intervals* at some level, like 95% or 99%. The idea is that, given that our vector is normally distributed, and we have good reasons to believe it will be with large enough sample size due to the central limit theorem, we are "confident" a point estimate like the mean will fall within these intervals that much percent of the time.

We introduce the concept of *perceptual confidence intervals* at the level of x cents or atoms, or even $x\%$ of some $\Delta f/f$, limited exclusively and solely by experiment. Whereas the concept of statistical confidence intervals is purely statistical, perceptual confidence intervals form an empirical concept. In the former we let the data decide how much variance exists in the normal distribution, how narrow or wide it is; in the latter we let human perception decide how wide or narrow we let those perceptual intervals be. The wider you let them be, the more incapable we claim humans are in being "confident" that a fundamental frequency detected by some engineering algorithm really falls within those intervals. The opposite is also true. Figure (3.4) shows a schematic representation of this concept. The shaded areas represent the perceptual confidence intervals and are not drawn to scale. The scale of those shaded areas should be whatever Figure (3.3) wants them to be. For example, at the lower end, around D_4 at 295 Hz, the jnd could be about 4 cents, or 21% of an atom, really, a tiny 60% of 1 Hz. At the upper end close to D_5 (590 Hz) the jnd would be about 3 cents, or 17.5% atoms, or about 1 Hz. These intervals, as one can clearly see, are very strict. They are so strict, that their practicality is limited in two ways.

The first limitation of using such low perceptual confidence intervals is technical: atoms were passed down to us through Byzantine music theory in integer values. There was no need for anything more precise. Recall that in the explanation of Figure (2.1), we said "Perfect intervals are consonant, that is why no matter if you are in a Byzantine or a Western scale D_4 will always be 293.665 Hz with respect to the reference tone A_4." However, tones E_4 and B_4 can *never* form a perfect interval with any other tone (in Byzantine music; in Western they can). This is what the little black squares wanted to denote in the graph. For perfect intervals, the number in atoms is either exact or very close to the integer (depending on the temperament). For non–perfect, dissonant intervals, this may become a problem, since our jnd are so strict they range from about a fifth of such an atom down to 17.5%.

The second practical limitation has to do with the intended use of this jnd, in other words, for whom does this jnd apply? For example, is it for the performer himself, for trained

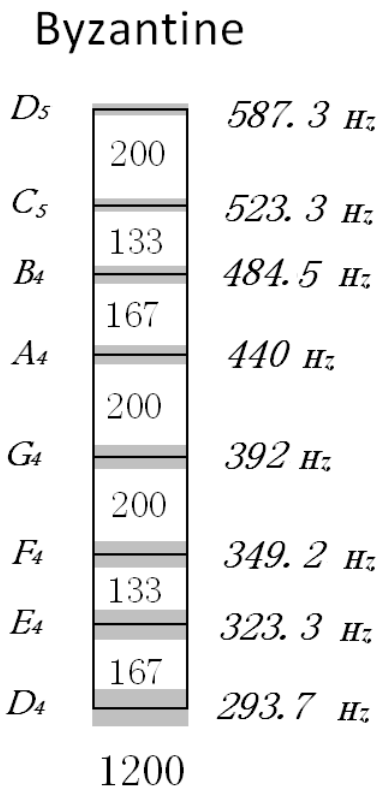


Figure 3.4 – *Perceptual Confidence Intervals.* Consistent with the idea of critical bands of the place theory in psychoacoustics, and borrowing elements from the statistical concept of statistical confidence intervals, perceptual confidence intervals give us a rough guideline of how confident we should be that a perceived frequency of a tone (the solid line separating the tonal boxes) is acceptably perceived as the same or at least not dissonant when compared to another tone differentiated by any frequency which falls within the shaded areas in the schematic above. Notice how the perceptual confidence intervals are a function of frequency (critical band concept).

chanters in Byzantine music, for trained vocalists in Western music, for instrumentalists, for people that were regulars at the cathedral where this person performed, or for the general public? It may sound as a trivial question, but it is not. Strict perceptual confidence intervals are more suitable for people trained in microtonal music cultures. Less restrictive intervals may be more realistic for people of no musical education[13].

[13]This is a subjective, anecdotal observation, that is why I put it in a footnote. I happened to have a conversation with a formally trained vocalist in Western music, a university professor and vocal coach of more than 30 years, who got involved with Byzantine music for the last 10. We were talking about scales and he performed what he believed to be the Byzantine chromatic scale, which happens to be a hard one to master. His intervals were clearly Westernized to me and I performed the scale his way and then the Byzantine way. He said that he can clearly hear the difference, but he would not be able to perform it in accordance to Byzantine standards. This example demonstrates a couple of situations where perceptual confidence intervals may need to be tailored for different cases. First of all, he could hear it, but not perform it. The jnd for *hearing* pitch discrimination, is not the same as the jnd for *performing* pitch discrimination. If such

3.6.2 Acceptable Performance Difference

During the course of our investigation on multiple musical performances by Nafpliotis, it quickly became evident that a tone one octave above is not exactly at a $2:1$ frequency ratio. Usually the tone one octave above, f_1, is slightly higher than double the fundamental, f_0, that is to say $f_1 > 2f_0$ instead of the expected $f_1 = 2f_0$. That the perception of the octave has a physical representation of a $2:1$ frequency ratio is universally accepted, not only in Western, but also in non–Western cultures (Burns & Ward (1978) [14]). This *octave equivalence* is what inspired music theorists around the globe to use only eight musical tone names and use numbers or primes to indicate their octave positions on the musical staff (like A_3, A_4, A_5, etc.). The pitch perception scale of *mel* (Stevens & Davis (1983) [79]), a psychometric function that shows how perceptual pitch varies as a function of frequency, was constructed on the premise of subjects modulating the frequency of a tone till it sounded half or double the pitch. However, it is also well known that under certain conditions *octave stretching* (Ward, (1954) [89]) or *pitch shift* (Smith et al., (1983) [75]) can alter the perception of octave equivalence.

Moreover, with singing voice especially, artistic license may grant the performer temporary permission to microtonally exaggerate an octave not only due to the performer's wish to project a vowel of low natural frequency over longer distances, but also due to the performer's wish to evoke certain feelings due to such exaggeration. What is noteworthy is that literature on octave stretching is usually based on *listening*, not performing subjects. It is, therefore, particularly interesting to find such an agreement between psychoacoustic results based on ear performance and their corresponding voice performance[14].

Whatever the cause of this octave stretching might be, it is found in Nafpliotis' data. The performer feels that this frequency deviation is within musical limits and it does not produce melodic disruption or dissonance. This may or may not be a phenomenon more prominent in singing voice signals, or even Byzantine music chant signals. Nevertheless, this frequency deviation is produced by the same performer within the same musical piece that is used to derive the frequencies of the musical scale. Whatever it is, is real and as valid as any other tone measurement. At the very minimum, we can assert with a certain degree of certainty that the performer is not appalled by this frequency deviation, no matter if it is intentional, unintentional, perceivable, or non–perceivable by him or the other listeners. We are not by any stretch of the imagination suggesting that this is the minimum this particular performer can perceive as a frequency deviation, that is his ear jnd. But we are making the claim that this amount of deviation is acceptable by the performer.

The next basic question should be "is this performing deviation more than the hearing jnd?" The answer to the above question is "yes." A typical example would be the base tone D_4 at about 302 Hz and D_5 not at 604 Hz, but at about 616 Hz[15]. In atoms this difference is about 2.0 and in cents it is about 34.0, a whole third of a semitone, which is substantially

subtleties exist within musically trained vocalists, you can imagine how wide the spectrum of possibilities is for untrained ears.

[14] Also interesting would be an experiment using musical instruments which can perform mictrotonal intervals, like the violin, and see if this ear–voice affinity is also valid for ear–instrument.

[15] Here the typical formula for octave stretching is used, namely, $\Omega = \frac{f_2 - 2f_1}{2f_1}$.

larger than all of the aural jnd's of Figure (3.3) let alone our low benchmark of about 3 cents or a fifth of an atom around 600 Hz. Remember the variability or difference in the experimental results were as high as about a fifth of a semitone and as high as this might seem, the discrepancy between Nafpiotis' performable deviation and aural jnd's is larger. Ironically, Nafpliotis' performable deviation is closer to older, coarser aural jnd measurements. It is tempting to think that maybe the aural jnd is not the best benchmark for comparing Nafpliotis' frequency deviations against. Maybe Nafpliotis' own allowable frequency deviations should be a benchmark for all Byzantine music chanters, since he is the most acclaimed Byzantine music performer caught on tape.

We propose a new term, the *acceptable performance difference* (apd), directly based on the idea of octave stretch as in Ward (1954) [89], and defined as

$$\Omega = \frac{f_2 - 2f_1}{2f_1}. \tag{3.3}$$

Due to the practical reasons outlined in this section, we decided to use this measure as our criterion for deciding if a frequency is deemed as acceptably different of not, in accordance with the idea of perceptual confidence intervals presented in Figure (3.4). Results in Chapter 4 will be interpreted using this definition.

It is important to note that octave stretching might very well be intentional, as we noted above, due to artistic license or due to any other reason. The *acceptable performance difference* may or may not apply to all kinds of intervals. As we will see in Chapter 4, the results suggest that for solid intervals, like a perfect fifth, Nafpliotis' performance is in agreement with the theoretical suggestions (41.9 versus 42 atoms, respectively, for one music piece analyzed, for example). However, similar stretching (which would lessen the agreement between theoretical and experimental) has been found in other intervals than the octave, like in perfect fifths or complete tetrachords. Some times an experienced listener can identify the performer's intent to overshoot the interval and this effect does evoke certain emotions. Think about a dramatic theatrical performer. In an attempt to convey feelings of distress the performer may utilize different ways of projecting the voice, like timber, intensity, voice cracking, loud pronunciation of proclamations like "alas" with intentional loose vowel elocution, etc. Many of these effects could be utilized by a Byzantine music vocalist as well, and some of them may affect the frequency directly (performer doing it intentionally) or indirectly.

3.7 Conclusions

This Chapter explored psychoacoustics literature on pitch discrimination to address the question "For musically trained humans, what is the minimum distinguishable frequency difference." Figure (3.3) gives the psychometric curves for jnd's in atoms and cents over the most predictive factor of pitch, the frequency. Since most reliable experiments that are relevant to our frequency range are based on pure tones, the issue of complex tones rendering themselves to better discrimination was used to make a case for using the most reasonable approach in adopting a jnd benchmark.

The restrictive nature of the lowest experimental aural jnd and the lack of a convincing explanation in it and of itself for using any other more liberal jnd, led to the definition of more customized measures of what should be an acceptable frequency deviation, namely the concept of acceptable performance difference. Within this framework, the concepts of perceptual confidence intervals and acceptable performance difference were introduced in parallel. All of these will be discussed in connection with results presented in Chapter 4 as to illustrate the practical advantage of apd's over aural ind's.

There is no single jnd or even acceptability metric adopted in this dissertation as a benchmark, simply because there is very little research on complex tones and specifically on singing voice (and none on Byzantine music chant). To rush into selecting one metric over another it would simply be naive. Much research remains to be done on the issue.

The next Chapter presents the fundamental frequencies of the tones of the Byzantine diatonic scale and provides discussion on theory versus practice.

Chapter 4

Results and Discussion

4.1 Introduction

The main focus of this Chapter is the tabulated presentation the results, i.e., the fundamental frequencies of the diatonic scale of Byzantine music as performed by Iakovos Nafpliotis. It starts out with some additional methodological information, beyond that given in Chapter 1, on the data collection, data preparation, and algorithm implementation. A general discussion on algorithm performance follows the actual results and some discussion on pitch discrimination based on information presented in Chapter 3 is also provided. The dissertation concludes after a general discussion on experimental versus theoretical scale intervals and suggested future research.

4.2 Methodology

A brief insight into the methodology was given in Chapter 1. Here we give more detail, even though some overlap exists.

Our samples come exclusively from the 5–CD publication of Professor Alygizakis (2008) [1], even though more than one music piece was analyzed. The old vinyl recordings were masterfully digitized and provide the most comprehensive as well as best preserved audio collection of Nafpliotis' music ever made available to the general public.

4.2.1 Data Preparation

Prior to collecting the data, the usual exploratory data analysis was performed. Even though the author of this dissertation has had experience with analyzing Byzantine music data from Nafpliotis' tape recordings that have been circulating in Byzantine music circles, some issues that existed in the older recording still exist now with Alygizakis' CD's. This is not necessarily a disadvantaged as we will see soon, however.

For example, in older tape recordings the frequency of the music piece (and therefore any tone of the musical scale to be tracked) gradually, but not linearly, decays over time.

This is easily verified by comparing the frequencies of a tone snippet from the beginning of the performance and another snippet of the same tone at the end (both of sufficient length, see notes on sampling below). Depending on the length of the piece, the frequency difference can be substantial. For example, for a piece about 3 minutes long a drop in frequency of about 15–20 Hz can be observed.

Of course the reason for this downward frequency shift over time in a non–linear fashion is due to the mechanical means by which the sound was originally recorded. Since the gramophone was not electrically powered, but rather by means of a mechanical spring, the force exerted by the spring to rotate the original record to be imprinted with Nafpliotis' voice was dissipating non–uniformly (at the beginning the force dissipates less, and the more the spring unwinds, proportion–wise, the force becomes weaker and weaker). This will cause the angular frequency to slow down in a manner proportional to the force dampening and the result is the drop in frequency[1].

We mentioned earlier that this anomaly in the data is not necessarily a disadvantage. Aligizakis' decision *not* to remedy this frequency drop is actually beneficial for our purposes. There are a number of methods that will stretch or compress the signal to achieve pitch shifts[2], but extreme care has to be taken as to the how the corrections will be applied, hopefully by reverse engineering the effect of the non–linear force dissipation to obtain the exact function. At any rate, it is better to have non–pitch–corrected data to extract pitches from instead of pitch–corrected data, even if accurate documentation of the corrections are available. No pitch correction was applied for the purposes of the data used in this dissertation.

We mentioned above that not only one, but multiple pieces of music were analyzed. The reason for that is partly this variable frequency shift. In general terms, any music piece will not have noticeable frequency dissipation rate within the first minute or so (frequency drops are in the order of a tenth or a hundredth of 1 Hz). We decided to analyze multiple pieces using only that initial segment of the recordings whose spectra are more stable. Other reasons for adopting this method is comparing consistency of the same performer across performances of the same scale. Remember that some performances were years apart in time. Nafpliotis' is extremely consistent across time[3].

Calculating and plotting frequency distributions for fundamental frequencies similar to the ones shown in Figure (2.4), is another exploratory data analysis way of becoming more

[1]Prior to acquiring Aligizakis's CD's I was always hopeful that this downward frequency shift would be due to the low mechanical quality of the *second* gramophone that later was used to transfer the signal to the magnetic tape. Unfortunately, however, it was the original recording which shifted downward; the shift still exists.

[2]One method of pitch correction is our familiar phase vocoder, specifically its synthesizing capabilities, which is also implemented in Beauchamp's SNDAN software. Not only can one adjust the hop–in and hop–out sizes to time–stretch, but additional control on tempo is also allowed.

[3]Frequency differences for each tone across different music pieces were on average less than 1 Hz. Ironically, Nafpliotis was much more accurate in preserving interval values over time (preserving for 60 years what his teachers had taught him) than the cutting–edge sound technology of his time (the gramophone couldn't preserve it for more than a minute). I suspect he still is more accurate than the cutting-edge psychoacoustics knowledge of my time. This limits technology as well. Because no matter how accurately we can estimate the fundamental frequency, pitch still seems to be an elusive concept.

familiar with one's data. Distributions from different tones, different time blocks within one piece, across different estimators, and across music pieces are all useful in learning more about frequency point estimates and their variances.

There are at least two kinds of frequency variances for this particular sample space[4], for any single tone: (1) frequency variance between snippets, and (2) frequency variance within one snippet. The first kind indicates how precise Nafpliotis' performance was (as opposed to how accurate he is, which requires comparison with the theoretical frequencies; see Section (2.3.3) for more details). The second would indicate merely the magnitude of vibrato Nafpliotis used in that snippet, similar to the examples we gave in Figure (2.4) for the trisyllable "me–nos–pros." A typical between–snippet standard deviation is about ± 4 Hz and a typical within–snippet standard deviation is about the same[5]. The fact that the between–snippet variation is about the same as the within–snippet variation in the data should provide some (weak) reassurance about the selection of the right snippets to represent the right tones, since as we pointed out in Chapter 1, this is still a subjective process.

4.2.2 Data Collection

Our data collection method is simple, but time–consuming. It basically consists of semi–subjectively deciding which snippet of music represents the corresponding tone one needs to extract or estimate the fundamental frequency from, saving it as a separate uncompressed sound file of the same high fidelity as the original CD signals, and then concatenating all of those snippets together in MATLAB®[6].

Above the term "semi–subjectively" was used, and even though "semi" implies about half of this process is subjective, it may be more than that. However, it is not entirely subjective. To facilitate a more practical discussion on what is meant here by subjectivity in data collection, let us ask the following question: "When the data collector selects a specific snippet of sound to be part of the space of a specific tone, how does he know that that tone really is part of the space?" In other words, are we creating a data collection process that will inevitably render our analysis adversely selected against? The data collector is the judge of what snippet makes it into the sample space. There is, therefore, an asymmetry of information between the space and the snippet. To know if a snippet should be in the sample, one needs to know its spectral content, but the content is not available until after the snippet was selected into the sample and analyzed. If we go about this problem the reverse way, that is collect the snippet, analyze it, and reassess if it should be in the sample or not, then we are running the danger of using the spectral content as an integral part of the data collection process, and this is an adverse effect, because the content is what the unknown to be estimated. In other words, if we reverse the process, we decrease information asymmetry,

[4]Since the data population (or universe) is the entirety of Alygizakis' CD's, formally speaking the sample space is the music piece of which the sound sample snippets originated, and samples are the actual concatenated snippets. Here I use the more friendly term "piece" instead.

[5]Using about a third of the standard deviation and then halving it since it is theoretically half above and half below the mean. If the vibrato was produced by a violin, say, instead of voice, then we would not need to half the standard deviation.

[6]MATLAB® is a registered trademark of MathWorks, Inc.

but at the expense of collection bias—only the snippets that *should* be in the sample *are* in the sample. The the accuracy, not the precision, of the results are compromised.

For example, the collector is in the decision process of whether a snippet that was just heard should be included in the sample space for estimating the frequency of tone F_4=349.22 Hz. The collector knows subjectively that this must be the tone F_4 since Nafpliotis' voice clearly is on that tone. Assume the tone is a clean vowel with no vocal embellishments, no rapid frequency changes, no glissando between tones, no vibrato, and no abrupt energy or amplitude fluctuations in the time domain. The collector "cuts" the snippet and includes it in a space with another 35 snippets. The the concatenated signal is analyzed. The collector sees that this tone is substantially lower than the average, let us say for example, about 346 Hz. Can the collector remove that snippet or not? If he does, he is using spectral information to decide what the spectrum should be. Equivalently, imagine we devised an algorithm that collects snippets for F_4=349.22 Hz, similar to musical instrument tuning software, that a note is played and the software detects its fundamental value and suggests adjustments. But when an instrument is tuned, the tuner starts out with no spectral content and wants to align his spectrum with the one suggested by the software. In existing pieces, the spectrum is there and a decision should be made on whether a tone belongs within the range of acceptable frequencies. In other words, even if an algorithm for data collection was created, it would not at all be likely to perform the collection better than a human collector, because the parameters of the algorithm are still input by the human.

One could argue that we can use a music score to see exactly when the performer's voice "passes through" the tone of interest and use every one of those tones in the resulting concatenated signal. This is practically impossible for several reasons.

First, there is no music score to accurately represent what Nafpliotis is chanting. He was, of course, taught the music using a music score, but the kind of music scores Nafpliotis was using (unlike some more analytical newer versions) were not accounting for all the detailed embellishments of the voice. They were more of a rough sketch of melodies and the chanter had the artistic freedom to deviate as long as his deviations were traditional, not arbitrarily[7]. One could devise an experiment where multiple chanters with experience in transcribing music to paper would do so and if the agreement between those was significant, an agreed–upon music manuscript could be used for this purpose.

Second, even if an accurate manuscript existed, not every time the voice passes through the tone of interest a reliable enough snippet is produced. Many of the snippets are contaminated with consonant sounds like "m" or "n", or fricatives like "s." Moreover, about half the time a tone is touched by the voice, the time duration is too short. A snippet cannot be infinitely small, for mathematical reasons. Since the Discrete Fourier Transform is at the

[7]This is another interesting topic, i.e., music memory. The older notation, the one Nafpliotis was taught partially but did not use throughout his career, was almost purely retained by memory. A symbol could mean an entire music line (called *thesis*) that could take several lines of today's notation to be written out. With the more simplified music notation that Nafpliotis was using, an accurate enough melodic line was written out (by means of notational *characters of quantity*) along with the addition of a plethora of signs to indicate not only by how much to go up or down the scale, but *how* to perform the line once there (notational *characters of quality*). The notation we are using today is the same as the one Nafpliotis used, but more detailed. So over the years, notation reduced the amount of memory storage needed.

heart of many fundamental frequency estimators presented in Chapter 2, let us remember how sampling in the time and frequency domain works.

Assume a snippet containing a clear vowel of the tone F_4=349.22 Hz (see Figure (2.1)) of our diatonic Byzantine music scale was selected to make it into the final sample. Its duration is $T_s = 0.5$ sec. The sampling frequency of the music sample is a conventional $f_s = 44,100$ Hz. The time resolution is therefore $\Delta t = 1/f_s \approx 2.267 \times 10^{-5}$ sec. The length of the snippet sequence in samples (or bins) is $M = f_s \cdot T_s = 22,050$. If we choose a window length of $N = 2^{11} = 2048$ samples[8], the snippet will be segmented in $\frac{M}{N} \approx 11$ frames. The window "fundamental" period, then, would be $T_N = \Delta t \cdot N \approx 0.0464$ sec. The frequency resolution is $\Delta f = 1/(N\Delta t) = T_N^{-1} \approx 21.53$ Hz, which means that the amplitude spectrum of the DFT will show amplitudes or energies at frequency bins separated by roughly 21.5 Hz each. The first DFT frequency bin will show how much energy is contained in the snippet that has 0 Hz (DC term), the second frequency bin (let us say $n = 2$ or at the 2/2048–th position) will show how much energy there exist in the signal at the frequency 21.5 Hz, the third at 43 Hz and so on and so forth. This will go on till $n = N/2$ and then the frequencies for which energy is shown for will descend down to zero, that is to say, only values up to $f_s/2 = 22,050$ Hz[9], the Nyquist frequency, can potentially be represented spaced by a resolution of 21.5 Hz. If our tone to resolve was a pure tone of 43 Hz, for example, the DFT amplitude spectrum would should a spike at the third frequency bin (or frequency band), assuming of course that our frequency resolution was 21.5 Hz exactly, not just rounded to that number like we did here. If a tone's frequency to be detected, however, does not coincide exactly with any of the 1024 different frequencies represented in our N–point DFT output, then the energy of that tone will be diffused over the entire window sample, what we referred to in Section (2.3.4.1) as spectral leakage, the effects of which we hope to attenuate (or taper) by windowing in the time domain before taking the transform.

The tone whose frequency we hope to estimate is F_4=349.22 Hz. The 16–th frequency bin would show the energy of a tone that happens to be at exactly 344 Hz and the 17–th frequency bin would show the energy of a tone that happens to have be at exactly 365.5 Hz. Our tone is, unfortunately, falling in between these two bins and thus we will have it diffused across the entire window. However, it is closer to the 16–th bin, off only by 5.22 Hz. This is about half of the worst case scenario which would have been a difference of $\Delta f/2$=10.75 Hz. But still, the inaccuracy of this hypothetical scenario is 1.56 atoms or 26.1 cents.

We would like to be able to increase the frequency resolution substantially. By inspection of the formula

$$\Delta f = \frac{1}{N\Delta t} = \frac{1}{T_N} = \frac{f_s}{N} \tag{4.1}$$

there are not many parameters to manipulate. One choice would be to decrease the sampling frequency, f_s, but that would limit how many partials would be included in our spectrum, since half of this value is the cut–off for the maximum frequency to be included. This is not detrimental for our purposes. We are tracking fundamental frequency, not the 49–th frequency. Even for psychoacoustic operation specific to pitch (not other phenomena) the

[8]Modern FFT algorithms do not require the window length to be a power of two, but choosing it to be so cannot hurt computational efficiency.

[9]That is an adequate maximum frequency since our ears cannot resolve anything more than that.

Table 4.1 – Window length N in sample points, frequency resolution Δf in Hertz, and window duration T_N in seconds. The following quantities are assumed constant: sampling frequency $f_s = 44,100$ Hz; signal duration (individual snippet or concatenated snippets) $T_s = 0.5$ seconds. The table illustrates the frequency–time resolution trade–off as a function of window length, ceteris paribus. Since to accurately estimate the fundamental frequency of a tone one complete fundamental window period must exist in the data, the duration of the sequence must be longer than the duration of the window, $T_s > T_N = 0.5$ sec. If only powers of two are used as the FFT length, the best we can resolve in frequency for a snippet of this duration is $2.69/2 \approx 1.34$ Hz. Please note that this is the resolution that two peaks can be resolved, not how well we can pinpoint where a peak is. Zero padding would fill in the spectrum and potentially help with locating the location of the peak more accurately.

N [samples]	Δf [Hz]	T_N [seconds]
2048	21.53	0.046
4096	10.77	0.093
8192	5.38	0.186
16,384	2.69	0.372
32,768	1.34	0.746
65,536	0.67	1.492

first handful of partials is enough. But, in general, we want to avoid band–passing our signal by downsampling.

Another thing we could be doing is increase the window length N. This will help our resolution indeed, but only if we collect more data samples so our snippet becomes meaningfully larger, that is to say, adding new information to the spectrum. Zero–padding (appending zeros at the right of the sequence) will not increase the frequency resolution at all. It will make the DFT approximate the DTFT better, but the increase is artificial, no new intelligence is added by adding more zeros. In other words, if we zero pad in the time domain, in the Fourier domain there will be more frequency bins on the exact same transform, but the width of the transform will not be narrower. The resolution is essentially the same.

If it wasn't for the uncertainty principle (see Section (2.4.2)) the solution would be easy: infinitesimally increase the window size until the resolution in the frequency domain approaches a continuous function. But then the resolution in the time domain would be horrible. Not only the time resolution is useful for practical reasons, like for example calculating between–snippet variance (because one needs to be able to see when things happen in time, like when a new snippet starts and the old one ended), but as a rule, one need to keep a complete fundamental period in the snippet window if one wants to be comfortable with the precision of the fundamental frequency estimate. As a consequence, the length of the window (in units of time) cannot exceed the length of the sequence, i.e., $T_N \leq T_s$. This is why the data collector cannot include a snippet of very small time duration.

In our example above the window length we chose allows for 11 fundamental periods, the same number as frames. This is because the ratio of the length of the entire concatenated

signal to the window length is the same in units of samples or units of time. So we can use this trade–off between time and frequency resolution to make the best out of an imperfect situation. As we saw above in Equation (4.1), time and frequency resolution are inversely proportional to one another with the window length as a factor to be adjusted (we also pointed out that downsampling is possible, but not a usual option in practice). If we increase N by powers of two (at least initially), we see from Table (4.1) that the $T_s > T_N = 0.5$ sec is at window size 16,384 samples, which provides a frequency resolution of about $2.69/2 \approx 1.34$ Hz. Since powers of two are not required with more modern FFT algorithms, we still have some leeway to fine–tune our frequency–time resolution. A window length of 22,000 samples yields a frequency resolution of $2.00/2 \approx 1.00$ Hz with a window duration of 0.498 sec which is theoretically still lower than the snippet length of 0.5 sec. Remember this is for illustration purposes. In practice, it is preferable to include two or three periods, not barely one and this check was performed on every concatenated signal that was analyzed.

It is important to note that Equation (4.1) assumes that the discrete transform's dimensions are as shown in Equation (2.2). Different notational conversions in textbooks sometimes assume that either $\Delta f = 1$ and $\Delta t = 1/N$, or $\Delta f = 1/N$ and $\Delta t = 1$, hence the $1/N$ factor in front of the forward or backward transform. But dimensional analysis dictates that for the units to be correct, if Equation (4.1) is used to calculate the above relationships, then Equation (2.2) must be used to calculate the DFT. This does not mean that different program packages that utilize equivalent DFT definitions will not produce the same results. It means that Equation (2.2) assumes Equation (4.1) for units to come out correctly.

4.2.3 Algorithm Implementation

The phase vocoder and McAuley-Quatieri algorithms were implemented in SNDAN (see Section (2.5) for details). In many occasions output data from SNDAN were exported and analyzed in SAS®[10]. The rest of the algorithms whose results are included here were implemented in MATLAB®. Program code for the latter was either written by the author of this dissertation, found at the appendices of books or articles that have been cited in Chapter 2 at the respective places where the mathematics of the different algorithms were used, or were found and used directly or modified from the MATLAB® Central File Exchange Web site which can be found here http://www.mathworks.com/matlabcentral/fileexchange/.

Indicatively, information on SNDAN algorithm implementation was found partly in Beauchamp (2007)[11] [2], and for the rest of the algorithms in Quinn & Hannan (2001) [61], and Stoica & Moses (2005) [80]. Again, this is not an abridged reference citation.

4.3 Results

This section presents the results obtained by implementing the following eleven algorithms: (1) Phase vocoder, (2) McAulay–Quatieri, (3) Levinson–Durbin Algorithm, (4) YIN,

[10] SAS® is a registered trademark of SAS Institute Inc.

[11] I would like to thank Professor James W. Beauchamp for helping me with initial set–up of his software and file structure back in 2005.

(5) Quinn & Fernandes Estimator, (6) Pisarenko Frequency Estimator, (7) *MU*ltiple *SI*gnal *C*haracterization (MUSIC), (8) Periodogram, (9) Quinn & Fernandes Filtered Periodogram, (10) Rife & Vincent Estimator, and (11) the Fourier transform itself.

Subsection (4.3.1) presents results of all eleven algorithms for sample data collected from one musical piece for comparison. Subsection (4.3.2) gives a comprehensive table including results of this dissertation, results from Tsiappoutas (2002) [85], a thesis leading to this dissertation, and theoretical suggestions of what the frequencies of the diatonic scale should be according to the two major theoretical schools of thought in Byzantine Music, i.e., that of the Patriarchal Byzantine Music Committee (1883) [52] and that of Chrysanthos from Madytos (1832) [18].

4.3.1 Algorithm Precision

Consistent with our definition of precision in Subsection (2.3.3), this subsection presents fundamental frequency estimates for all eleven (11) pitch detection algorithms presented in Chapter 2.

Strictly speaking, the accuracy of the estimates cannot be determined in absolute frequency terms. For example, the tone A_4 seems to be at a frequency of 449.6 Hz, and there is no industry standard to compare that against. In relative terms, however, upon taking ratios between tone frequencies (or atoms or cents, for that matter), one can readily make comparisons between scales that were extracted from data of different pieces of music. This aspect will be explored more in the next section, where frequencies are shifted so that their point of reference, tone A_4, is at the same frequency of 440 Hz.

Algorithm performance in terms of computational efficiency was hinted throughout Chapter 2. Algorithm performance in terms of accuracy cannot be determined, strictly speaking, since, as we noted above, there is no standard to compare each one against. For example, if the fundamental frequency of the signal were known a priori, one could devise metrics to show distances between the standard and each estimate. In our case, however, f_0 is the unknown parameter to be estimated.

The data in Table (4.2) are from one music piece only, which is entitled, "erhomenos o Kyrios," which roughly translates to "when the Lord was coming." Casual inspection of the frequency estimates themselves shows very precise measurements, within 80% of one Hertz. We believe this estimation precision to be sufficient for our purposes. Remember that a Δf of one Hertz around the lower range of the tone frequencies, say 300 Hz, is about 0.34 of an atom or about 5.7 cents. At about 600 Hz this difference is about 0.17 of an atom or about 2.88 cents.

We see here, and we will keep this in mind throughout the presentation of the results in this Chapter, that algorithm precision is limited by our perception. This is a very justifiable benchmark, just like the well accepted $f_s = 44.1$ kHz which is imposed by our inability to perceive any frequencies higher than half the sampling rate.

		ALGORITHMS											Average	St. Dev.	Confidence Intervals	
		FT	PV	MQ	LD	YIN	QF	PIS	MUSIC	PER	QFFP	RF			Lower 99%	Higher 99%
TONES	D_5	604.3	604.0	604.5	605.3	604.6	604.5	606.0	604.1	604.0	604.3	605.2	604.6	0.62	604.0	605.2
	C_4	537.9	538.7	538.4	538.4	538.3	539.0	537.9	538.2	537.8	538.7	537.9	538.3	0.38	537.9	538.6
	B_4	499.3	499.6	500.0	498.5	498.1	499.7	499.1	499.3	497.6	499.3	497.9	498.9	0.80	498.2	499.7
	A_4	450.1	449.0	450.0	449.3	448.1	450.0	450.4	449.5	449.6	449.5	450.5	449.6	0.68	449.0	450.3
	G_4	398.1	397.1	398.4	398.3	399.1	398.1	398.5	397.7	397.7	397.3	397.6	398.0	0.59	397.4	398.6
	F_4	350.8	351.7	351.6	352.0	351.3	352.6	351.2	351.9	352.4	352.4	351.1	351.7	0.60	351.2	352.3
	E_4	329.6	330.9	330.4	330.7	330.7	330.4	330.0	330.1	329.5	329.6	329.5	330.1	0.52	329.6	330.6
	D_4	298.2	300.3	300.6	299.3	300.2	300.3	298.7	300.6	299.1	299.6	299.6	299.7	0.80	298.9	300.4

Table 4.2 – *Algorithm Precision.* Results of frequencies in Hertz for the eleven (11) algorithms presented in Chapter 2 for comparison. The averages, typically taken to be the the final frequency for each tone of the diatonic scale, have a standard deviation of no more than 0.80 Hz, well below the just noticeable difference even by conservative standards. The Confidence Intervals for the mean at a 99% level are also shown.
Algorithm abbreviations: FT—Fourier transform; PV—Phase vocoder; MQ—McAulay–Quatieri; LD—Levinson–Durbin Algorithm; QF—Quinn & Fernandes Estimator; PIS—Pisarenko Frequency Estimator; MUSIC—*MU*ltiple *SI*gnal *C*haracterization; PER—Periodogram; QFFP—Quinn & Fernandes Filtered Periodogram; RF—Rife & Vincent Estimator.

4.3.2 Tabulated Results

The results are tabulated in Table (4.4). At the highest level, the table is sectioned in three parts: (1) *Frequencies in Hertz*. These are the experimentally extracted frequencies shown in Table (4.2), center–shifted so that the tone A_4 is at 440 Hz. This enables us to use tone A_4 as a reference tone to ascend or descend along the scale space using atoms, cents, or frequency ratios. The reasoning behind our decision to center–shift is mostly practical: (1a) most of us are accustomed to thinking of A_4 at 440 Hz as a yardstick that makes tones of perfect intervals fall at exactly the same locations as Western music intervals do. Shifting origins just makes comparisons more intuitive; and (1b) it makes it easier to compare frequencies in cases where one chanter is contrasted with another. In this case, for example, Nafpliotis with Stanitsas can be compared even at intervals that are not perfect; (2) *Atoms*. These are given in both theoretical scales used here, i.e., that of the Patriarchal Byzantine Music Committee (1883) [52] (labeled as "Committee") and that of Chrysanthos from Madytos (1832) [18] (labeled as "Chrysanth"); (3) *Cents*. These are basically a mere transformation of the values in (2) solely to facilitate understanding of the table results for readers not used to the Byzantine units of atoms.

The acceptable performance differences (apd's) shown in Table (4.4) deviate significantly from the jnd's proposed by psychoacoustical experiments presented in Figure (3.3). The reasons for this decision are multiple, as presented in Chapter 3. Most importantly it was the restrictive nature of the narrowness of the most conservative findings among experimentalists. However, the pattern of the curves in Figure (3.3) were used in conjunction with our two proposed customizations, specifically (1) the *acceptable performance differences* presented in Subsection (3.6.2), which incorporates the notion of (2) *Perceptual Confidence Intervals* presented in Subsection (3.6.1). For example, using the formula given by Ward (1954) [89], namely, $\Omega = (f_2 - f_1)/(2f_1)$, and the fact that in this particular piece Nafpliotis is stretching the octave by about 5 Hz, we can extrapolate that value downward one octave using the lowest two values from Figure (3.3) as a guide[12].

Table (4.4) compares two theoretical intervals on top of the two experimental ones. A measure of how the two theoretical intervals compare to each other is not easy to construct, but perhaps the use of just intonation intervals would shed light on some of the intervals in Byzantine music that happen to be perfect, like a perfect fifth and a perfect fourth. The same comparisons could be made with major (and minor) thirds, but to a lesser extend. Perfect intervals are universally understood as fundamental building blocks of scales for many cultures. They are extremely fundamental to the harmonic structure of the scale. Major and minor intervals, on the other hand, are also consonant, but may not exist purely as independent and self–sufficient scale building units for other cultures. Table (4.3) compares the two theoretical scales used in this dissertation, namely that of the Patriarchal Byzantine Music Committee (1883) [52] (labeled as "Committee") and that of Chrysanthos from Madytos (1832) [18] (labeled as "Chrysanth").

[12] The two lowest values in this case were that of Moore (1973) at the lower end (about 0.24 Chrysanthian atoms) and that of Harris (1952) at the higher end (about 0.13 Chrysanthian atoms). A linear downward relationship was assumed, for simplicity.

		Ratio	Committee		Chrysanth		Normalized Percent Differences	
			Just	Actual	Just	Actual	Committee	Chrysanth
Just Intervals	Fifth	3:2	42.12	42	39.77	40	-0.286%	0.575%
	Fourth	4:3	29.88	30	28.22	28	0.400%	-0.786%
	Third	5:4	23.18	24	21.89	24	3.417%	8.792%

Table 4.3 – *Just Intonation vs Theoretical Scales.* The Pythagorian perfect fifth and fourth as well as the major third interval are contrasted to two Byzantine music theoretical guidelines. The better agreement between just intonation and Byzantine music as compared to the less accurate representation of the major third may be indicative of not only the fundamental nature of the perfect intervals, but also the theorists' intent to represent those perfect intervals as accurately as possible. Percent differences are normalized row–wise, i.e., each with respect to its corresponding interval (not a whole scale omnibus normalization), and it should be interpreted as such.

Above it was mentioned that multiple pieces were analyzed. The results in Table (4.4) are exclusively from the sound track entitled "erhomenos o Kyrios," even though results from at least one other track ("esose laon") yield very close intervals.

4.4 Discussion

There are several points of interest in the results presented in Table (4.4). Most notably, how well experimental results approach the theoretical suggestions. Before we discuss this aspect of the results, however, some discussion on the theoretical aspect of intervals alone should prepare us better appreciate the subtle differences between experimentation and theory.

4.4.1 Theoretical vs. Theoretical

The need for establishing just noticeable differences in pitch discrimination was eminent throughout this research. As seen in Table (4.2), the fundamental frequency estimates are precise *enough* for us to feel comfortable using their output as valid data to compare experiment to theory. Pinpointing how much is sufficiently *enough*, has been typically seen (throughout this dissertation) as a function of our perception: if one cannot perceive a frequency difference, clearly the performer is not expected to perform it and the listener is not expected to distinguish it, even if it was produced my electronic means.

During this research, however, we have presented data on pure tones (that usually yield worse pitch discrimination as compared to complex tones) that were limiting our ability to make any meaningful judgments for the application at hand of the singing voice. The great variability in the psychoacoustics literature led us to believe that maybe only some experiment methods were applicable to our specific needs, but we also commented that the justification of using one experiment over another would be a subjective point to argue

Tone	Frequencies (Hz)					Atoms					Cents				
	Experimental		Theoretical		APD	Experimental		Theoretical		APD*	Experimental		Theoretical		APD
	Nafpliotis	Stanitsas	Committee	Chrysanth		Nafpliotis	Stanitsas	Committee	Chrysanth		Nafpliotis	Stanitsas	Committee	Chrysanth	
D_5	591.7	580.0	587.3	585.3	5.0										
						12.1	12.1	12.0	12.0	0.9	201.1	202.4	200.0	211.8	15.0
C_4	526.8	516.0	523.3	517.9	5.7										
						7.9	6.3	8.0	7.0	1.1	131.7	105.5	133.3	123.5	17.5
B_4	488.2	485.5	484.5	482.3	6.4										
						10.8	10.2	10.0	9.0	1.2	180.0	170.4	166.7	158.8	20.0
A_4	**440.0**	**440.0**	**440.0**	**440.0**	7.1										
						12.7	12.5	12.0	12.0	1.4	211.1	208.8	200.0	211.8	22.5
G_4	389.5	390.0	392.0	389.3	7.9										
						12.8	14.3	12.0	12.0	1.5	214.1	237.5	200.0	211.8	25.0
F_4	344.2	340.0	349.2	344.5	8.6										
						6.6	6.3	8.0	7.0	1.7	109.5	105.0	133.3	123.5	27.5
E_4	323.1	320.0	323.3	320.8	9.3										
						10.1	10.2	10.0	9.0	1.8	167.5	170.4	166.7	158.8	30.0
D_4	293.3	290.0	293.7	292.7	10.0										
TOTAL						**72.9**	**72.0**	**72.0**	**68.0**		**1215.0**	**1200.0**	**1200.0**	**1200.0**	

Table 4.4 – *Fundamental frequency estimations.* Frequencies of the diatonic scale of Byzantine music based on Iakovos Nafpliotis' performance of "erhomenos o Kyrios" are the average of the eleven (11) different performance algorithms whose individual results are presented in Table (4.2). The frequencies have been shifted so that the reference tone A_4 is at 440 Hz. The frequency estimates based on Thrasivoulos Stanitsas' performance of the slow version of "Kyrie, ekekraksa" are also shown here for comparison. The latter frequencies are taken from Tsiappoutas et al. (2004) [85]. The metric used to indicate acceptable pitch deviation in performance shown here (apd's) is based on the proposed customization called *acceptable performance differences* presented in Subsection (3.6.2) which incorporates the notion of *Perceptual Confidence Intervals* presented in Subsection (3.6.1). Columns labeled as "Theoretical" provide the two most established Byzantine music theory guidelines of what a chanter should be performing. These theoretical guidelines come from Patriarchal Byzantine Music Committee (1883) [52] and Chrysanthos from Madytos (1832) [18]. Values in atoms and cents are also provided, to accommodate the reader familiar to each one of those inter–interval quantification systems. Throughout the Table, "Theoretical" refers to what it should be according to music theory and "Experimental" to what it is in reality. This distinction also holds for atoms and cents.

* indicates that the apd's for atoms were calculated based on Chrysanth's theoretical scale of 72 atoms.

for or against. Consequently, the author of this dissertation proposed a departure from psychoacoustical experimentation towards other measures of pitch discrimination that are not based on psychoacoustics per se, but on a hybrid of Nafpliotis' data and experimental patterns. In this section we see yet another source of possible jnd's derivation: pure music theory, with no connection to human experiments or deduction from human performance.

Table (4.3) presents some clear asymphony among just intonation intervals, the committee's suggested intervals, and Chysanthian intervals. We alluded earlier to the fact that atoms were rounded up to the nearest integer and that could be a source of slight disagreement. If the theoreticians were not concerned about giving distinctions finer than one atom, then maybe one atom was intended to be a theoretical jnd. Of course, the more atoms one includes in the scale, the finer the distinctions would be made, so each situation should be viewed in a normalized fashion, or in some ratio representation that makes each unit relevant to its total. This argument is not very convincing, in our opinion. It seems that practicality was the main concern of theoreticians.

4.4.1.1 Just Intonation vs Byzantine Intervals

Just intervals are considered the clearest theoretical definition for perfect intervals. From Table (4.3) we see that in general the Committee's intervals are closer to just intonation. This can readily be seen by the magnitudes of the normalized percent differences (signs indicate over– or under–shoot of the theoretical as compared to the just intonation intervals).

This could be an intentional goal of the Committee or a result of using their perception of tonal intervals during the implementation of the special instrument they have devised and used to determine their ratios. In the case of the first assumption, the closer proximity between the Committee and the just is less interesting for the purposes of this dissertation than the second assumption.

4.4.1.2 Committee vs Chrysanth

Even beyond single–atom disagreements, we see that for a major third Chrysanth's scale falls short from the just interval for more than one atom, about 2.11 atoms. The Committee's interval is closer to just major third, even though both of them exaggerate the third with their full blown "ditone[13]." In terms of percentage points, the Chrysanthian scale wants the third about 8.8% greater than the just and the Committee's scale wants it only about 3.4% greater.

The disagreement between the two theoretical scales could be a stand–alone jnd. The reasoning behind this proposal is the following: If two of the most influential interval theories are in disagreement about a standard interval like the major third, it would imply that either one of them was very wrong about their suggestions, or a difference of more than 5% of the major third is not that significant to argue over.

[13]This fact in it and of itself is noteworthy. I always thought of the Patriarchal style *exaggerating* the major third and from what we see in the experimental results, it is.

In Byzantine music circles (except for very specialized discussions that are not part of the every day practicing chanter's activity), such differences are not really discussed. We cannot help but to conclude that accuracy of the theoretical scales were not the primary priority of theoreticians. The exact performance of intervals is based on verbal teaching, not in the numbers displayed on a page. A teacher might point to a number and remind the student to adjust some of the intervals during lessons, but the aural input is valued much more than the visual.

4.4.2 Experimental vs. Theoretical

Experimental results as shown in Table (4.4) are in good agreement with the theoretical. Probably the only exception would be the A_4–F_4 major third, which is exaggerated by both chanters, even though Nafpliotis is closer to the theoretical guidelines than Stanitsas is. More specifically, the theoretical suggestion is 24 atoms, Stanitsas performs this third at an astonishing 26.8 atoms whereas Nafpliotis yields a 25.5–atom major third (both measurements are in Committee's atoms).

The second noticeable difference between the two chanters would be the A_4–C_4 interval, which in the Western scale it would have been a minor third, but in the Byzantine scale it is what the two theoretical guidelines give us (see Table (4.4)). A suggested 18–atom interval, Stanitsas undermines it by about 1.5 atoms at 16.5 atoms, and Nafpliotis exaggerates it by 70% of an atom, i.e., at 18.7 atoms. This may have something to do with the fact that Nafpliotis stretches his octave slightly. Of course, anything more said on this issue would be purely subjective.

In the lower tetrachord, the interval E_4–F_4 is diminished to accommodate the stretched major third.

4.4.3 Nafpliotis vs Stanitsas

In Tsiappoutas et al. (2004) [85]) it is noted that Stanitsas' performance of F_4 is substantially lower than the theoretical. We see evidence now that Nafpliotis is not so low as Stanitsas in this respect. However, they both exaggerate this major third to some extent.

In the second tetrachord, Nafpliotis seems to be more robust in the lower than in the upper end. His upper end is exaggerated slightly by about one Committee atom, or about 5 Hz, an octave stretch that may be the result of either perceptual pitch shift or intentional artistic license. The interval A_4–B_4 is for both chanters over the theoretical suggestion, but Stanitsas' first step into the tetrachord is not as decisive. Even greater differences are exhibited in the midst of the second tetrachord, with interval B_4–C_4 showing Nafpliotis' clear intentions to either undermine the pure C_4–D_5 tone, or overstretch it. He does the latter.

In the first tetrachord upwards we observe a very good agreement between the two chanters, at the expense of theory. The interval E_4–F_4 is consistently lower than both theoretical suggestions, with both chanters being within 0.3 atoms of each other. This could be (and probably is) a side–effect of the over–stretched major third above. The interval D_4–E_4 is also very stable among chanters and in very good agreement with theory as well.

The interval A_4–D_4, the perfect fifth, which theoretically in the just intonation space should be at a clear 1.5 factor, Nafpliotis achieves an astonishingly accurate 1.50017. This fact alone should make us feel very comfortable about the performer and the algorithms. The two perfect fourths, one in each tetrachord, are very well established as well, with the lower one being at a factor of 1.327 and the overstretched upper one being in a factor of 1.344. The upper perfect fifth is also overstretched, at 1.519.

4.4.4 Thesis vs Dissertation

In general, the work done for the dissertation research was both broader and deeper than the work done for the thesis research. First, where only the Fourier transform was utilized in the thesis research, there were ten (10) additional algorithms used in the dissertation research. The assurance we achieved by the close results of the several algorithms is very comforting. Any *one* algorithm output would be always suspicious, but such general agreement between algorithms cannot happen by chance.

There were several psychoacoustical findings, comments, theory, practical results that were part of this dissertation, but probably outside of the scope and comprehensiveness of a thesis work. Even though these additional bits and pieces were not directly related to the detection of fundamental frequencies, they did offer insight into the qualitative aspects of Nafpliotis's voice through quantitative findings.

4.5 Conclusions

Comparisons (and some conclusions) among theoretical scales have been provided earlier. Theoretical comparisons are insightful for understanding what a person performs in practice, because they reveal the suggested foundations upon which a piece should be performed. On the other hand, empirical comparisons, especially from renowned chanters like Nafpliotis, are not the foundation of theory, but it is arguable if they should be. In other words, there is an obvious one–way process–flow between theoretical and empirical, but the reverse is not necessarily true, unless the empirical is performed by such an authority on the subject that outweighs the theoretical. To support this notion, one could also argue that the empirical is what survived through the centuries in reality, and the theoretical might have been off due to many factors, one of them being the technological limitations of the times. Whatever framework conclusions are placed in, comparisons among music–theoretical suggestions for the same genre could have been made without the contribution of this dissertation, which concentrates on setting the foundations for reliably and precisely estimating the diatonic scale empirically. Therefore, purely theoretical comparisons will be left to other, more capable researchers, like ethnomusicologists. In this section we concentrate on the empirical conclusions, with direct references to the theoretical comparisons.

There are five (5) moving parts to the following conclusions. The empirical information on the diatonic scale is due to (1) Nafpliotis and (2) Stanitsas; the theoretical information on the same scale is due to (3) Chrysanth, (4) Patriarchal Committee, and (5) just intonation. It was found that breaking down the conclusions by interval of interest is the most insightful way to present. The following subsections do this. The final subsections point out some high–level, general conclusions.

4.5.1 Octave

The octave is universally accepted as the most perfect of all intervals, save for the unison itself. It is the easiest to hear and to perform. Most psychophysical experiments have had subjects double or half the frequency of a modulated tone, indicating the octave above or below. It is this apparent and universal perfection of the octave that makes it especially difficult to interpret Nafpliotis' tendency to overstretch it by fifteen (15) cents[14].

Inconsistencies between the experimental and theoretical such as these, justify our decision to use eleven (11) different algorithms for the extraction of the diatonic interval fundamentals. Otherwise it might seem an apparent data glitch or algorithmic oversight. Special care was taken to rule out most controllable possibilities for such mistakes, both in data collection and in algorithm implementation. The algorithm consistency alone should provide comfort in these results.

Methodology and algorithms aside, one could conjecture that this inconsistency between theory and practice is due to the performer's poor delivery of the performance or inadequate musical education. This justifies our decision to choose Nafpliotis as our universe space. Anyone with moderate understanding of who the performer is should find it very easy to dismiss such arguments in a moment's thought. But let us not impose on the reader's ability to judge the sample, let us form another question that could help answer if the inconsistency is due to poor performance. If this were true, then other intervals which are universally accepted as fundamental building blocks of scales would inevitably be underperformed, as, for example the perfect fifth.

4.5.2 Perfect Fifth

Nafpliotis performs the perfect fifth interval A_4–D_4 astonishingly accurately[15] with reference to the just intonation suggestion. The just intonation suggestion for a perfect fifth is about 702 cents; the Committee's suggestion is 700 cents; and Chrysanth wants his perfect fifth to be 705.9 cents. Nafpliotis chooses to perform it closest to the just, with 702.2 cents. In regards to just intonation, the difference between theoretical and experimental is minute. It is about twenty (20) times below the sensory just noticeable difference of about 4 cents (that was deemed by us too narrow to have any practical meaning) and 150 times below the acceptable performance difference of 30 cents.

We will operate on the safe principle that such accurate results do not happen by chance. A poor performer is not likely to achieve such perfection by accident. In reference to the previous section, we can then ask, "was the octave overstretch erroneous?" We can relatively safely answer "No" based on the following argument: it is highly unlikely that Nafpliotis performs the perfect fifth, which is harder to perform than the octave, with such accuracy and overstretches (consistently, the octave was not once understretched across pieces) the octave by such a gross amount.

[14]This fact was not only true to this specific music piece analyzed for this dissertation. Nafpliotis seems to overstretch his octaves even by more than 15 cents in other diatonic pieces as well, not presented here

[15]Here we are allowed to use the word "accurate" instead of "precise," because we refer to intervals, which are relevant to each other's frequencies, not absolute frequencies.

The validity of Nafpliotis' performance is an important point to establish, because this is the basis of our proposed acceptable performance difference measure. Had Nafpliotis' performance been proven unreliable, which it has not, then methodology and algorithmic precision should be also questioned.

Stanitsas overstretches his fifth significantly at 721.7 cents, even though it is below the acceptable performance difference of 30 cents[16]. Clearly, the two chanters (an experimental vs experimental comparison) have perceptible yet musically acceptable differences.

4.5.3 Perfect Fourth

The perfect fourth is not so consistent with theory. The Committee (as well as Western music) wants this interval at 500 cents; Chysanth at 494.1 cents. Nafpliotis performs is at 491.1 cents and Stanitsas at 512.9 cents. The difference between the two chanters (which is the widest range possible among all theoretical and experimental values, that is to say, the most extreme values relevant to the fourth) is 21.8 cents, more than two thirds of the apd. Again, acceptable musically[17], but different significantly for the trained ear.

Again, we see Nafpliotis closer to Chrysanthian and just tonality. Just intonation, long considered as the preferred suggestion for all vocalists (Western and Byzantine alike) wants the fourth shorter than the Western. Chrysanth minimizes it even more. And Nafpliotis places it even shorter. Stanitsas is directionally opposite of all theoretical trends (that span from 500 cents to 494.1 cents).

4.5.4 Major Third

The major third is the only interval exhibiting such huge difference between the just and Byzantine suggestions. The just intonation gives this interval a depth of 386 cents, whereas the Committee (and also Western) want it at 400 cents and Chrysanth is again overstretching it at 423.6 cents. Nafpliotis is remarkably close to the Chrysanthian suggestion, while Stanitsas exaggerates the already stretched Chrysanthian interval.

This is the only noteworthy time that Nafpliotis departs significantly from just and abides more with Byzantine, specifically Chrysanth. This may be one of the trademark intervals for the Byzantine diatonic scale, so uniquely situated by Chrysanth.

[16]Remember, this is the acceptable difference. It implies that differences greater than this threshold are not acceptable, based on Napliotis' overstretched octave.

[17]Maybe a clarification is in order on the term "musically acceptable." We do not mean it in the sense that in Byzantine music values below the apd are acceptable. To the contrary, in Byzantine music, a music based on microtonal intervals, a much smaller value is what is considered traditional. Musically acceptable here is meant in regards to producing dissonance. This point in it and of itself is worth another whole dissertation. The determination of what is traditionally unacceptable even though it is consonant is a very interesting topic. Traditional chanters like Nafpliotis are not forgiving of the slightest deviation from the norm. Yet, other traditionalists, like Stanitsas (of lesser stature, to be sure) do perform intervals differently.

4.5.5 Second

The interval G_4–F_4, a major second, is overstretched by Stanitsas by almost an entire apd. The Committee and Chrysanth assign 200 and 211.8 cents, respectively. Nafpliotis performs it at 214.1 cents, in good agreement with Chrysanth, and Stanitsas at 237.5 cents.

This was one of the major findings of Tsiappoutas et al (2004) [**?**], in connection with Stanitsas' overstretched major third. As we see now clearly, this *is* the reason why his third is stretched as much as it is, due to this second, not the one above it. A very peculiar departure from reasonable expectations, indeed.

4.5.6 Second Tetrachord

In the second tetrachord Nafpliotis ascends very decisively and assertively from the outlet. His B_4–A_4 is a surprising 180 cents, far above his own E_4–D_4 of 167.5 cents, which is in good terms with the Committee (166.7 cents) and less with Chrysanth (158.8 cents). His second interval into the upper tetrachord is also in good agreement with the Committee and a good deal higher than his corresponding second interval of the lower tetrachord.

This tonal behavior (at least the ascending side of the scale) is sometimes described as "real diatonic," in the sense that the exaggeration is in accordance with the diatonic character of the scale.

Additionally, this kind of very aggressive entrance into the second tetrachord might be preparing the grounds for the octave overstretch that we discussed in an earlier subsection.

4.5.7 General Conclusions

In general terms, Nafpliotis seems to be more just than Byzantine most of the times, and he is Byzantine, Chrysanth is his preferred structure, at least in the first tetrachord of the diatonic scale. The major third is one of the very prominent exceptions, which renders this interval uniquely Byzantine in theoretical terms.

Stanitsas' tendency to exaggerate the major third interval A_4–F_4 is still evident in Nafpliotis' results, but to a lesser extent. Nafliotis ascends in a more robust and diatonic manner in the second tetrachord than Stanitsas.

Nafpliotis' over–stretched octave has given rise to a newly defined just noticeable difference definition, which is a hybrid of psychoacoustical experimental results and Nafpliotis' data. The concepts of *acceptable performance differences* (Subsection (3.6.2)) and *Perceptual Confidence Intervals* (Subsection (3.6.1)) were both used as a final metric of acceptable performance deviation. Other ways of presenting or conceptualizing jnd's or acceptability measures suitable for the application at hand were explored along the way.

Overall, the author of this dissertation is satisfied with the agreement among different algorithms in detecting and estimating the fundamental frequency of acoustical signals from Nafpliotis' recordings. The psychoacoustics aspect of providing a clear–cut acceptable frequency deviation, however, is not as straightforward as the author had hoped it would be when this effort started several years ago.

4.6 Future Research

Suggested future research has to do more with the psychoacoustics portion of this dissertation and the inclusion of more diverse data rather than the pure detection and estimation of signals.

Frequency estimation is an established field with many different tools to choose from for different applications. For the estimation of frequencies from a singing voice acoustical signal, several have been utilized in this dissertation, and certainly the list of the algorithms included was not exhaustive. Of course, there is always the possibility of future research on more customized algorithms than the ones currently known, but we feel that the precision of the estimates we achieved as presented in Table (4.2) is satisfactory for our case. Any better precision is limited by our perception.

When it comes to establishing an authoritative apd's specifically tailored for the singing voice for microtonal intervals, there is no black and white answer. Several ways of defining such metrics were proposed in this dissertation research, but it would be outside the scope of this paper to explore this aspect more. Probably the most promising way for establishing an apd specific to our situation would be a combination of psychoacoustics literature and data directly based on Nafpliotis' performance. For example, one could construct tones from resynthesized signals from Nafpliotis' voice and have subjects not only distinguish aural jnd, but also reproduce performable apd's.

One distinction we made specific to jnds was between aural and performable. A series of experiments could explore differences between subjects that have musical training in Byzantine music vs not musically trained. A combination of pure and complex tones based on resynthesized signals from Nafpliotis' voice could shed more light on the issue of pure vs complex tones jnd, both aural and performable.

Furthermore, the octave overstretch of 15 cents found in this piece will not be the same across all pieces by Nafpliotis, or across chanters. Maybe a normalized value can be formalized for future benchmarking.

Another interesting aspect to explore is if this octave overstretch was intentionally or unintentionally done. A factor into all these might be the architectural acoustics of the cathedral which echo the voice in complex ways. Maybe chanters in cathedrals and opera singers do not behave similarly when it comes to octave stretches.

The groundwork done in this dissertation lends itself as a foundation for a host of data analyses from chanters of different schools of thought (see discussion relevant to progressivists and traditionalists in Section (1.1.2). Performers representative of one school of though can be analyzed to establish an acceptable variation in fundamental frequencies within one group, and that variance could be compared to variation within another school of thought. Of course, direct mean differences could also be compared and analyzed statistically.

Bibliography

[1] Alygizakis, A. E. (2008). 78 RPM Orpheon–Odeon [1914–1926]. Byzantine Music. *The Protopsaltis of the Holy Great Church of Christ, Iakovos Nafpliotis*. [CD]. Book + 5 CDs. Kalan Music.

[2] Beauchamp, J. W. (Eds.) (2007). *Analysis, Synthesis, and Perception of Musical Sound—The Sound of Music*. Springer: Modern Acoustics and Signal Processing.

[3] Beauchamp, J. W. and Fornango, J. P. (1966). Transient Analysis of Harmonic Musical Tones by Digital Computer. *31st Convention of the Audio Eng. Soc., New York, Audio Eng. Soc.* Preprint No. 479.

[4] Beauchamp, J. W. (1969). *A Computer System for Time–Variant Harmonic Analysis and Synthesis of Musical Tones, in Music by Computers*. H. von Foerster and J. W. Beauchamp, Eds. (J. Wiley & Sons, New York), pp. 19–62.

[5] Beauchamp, J. W. (1975). Analysis and Synthesis of Cornet Tones Using Nonlinear Inter–harmonic Relationships. *J. Audio Eng. Soc., 23*(10), 778–795.

[6] Beauchamp, J. W.(1993). Unix Workstation Software for Analysis, Graphics, Modification, and Synthesis of Musical Sounds. *94th Convention of the Audio Eng. Soc., Berlin, Audio Eng. Soc.* Preprint No. 3479.

[7] Beauchamp, J. W., Maher, R. C., and Brown, R. (1993). Detection of Musical Pitch from Recorded Solo Performances. *94th Convention of the Audio Eng. Soc., Berlin, Audio Eng. Soc.* Preprint No. 3541.

[8] Beauchamp, J. W. and Horner, A. (1995). Wavetable Interpolation Synthesis Based on Time–Variant Spectral Analysis of Musical Sounds. *98th Convention of the Audio Eng. Soc., Paris, Audio Eng. Soc.* Preprint No. 3960.

[9] Boersma, P. (1993). Accurate short–term analysis of the fundamental frequency and the harmonics–to–noise ratio of a sampled sound. *Proc. Institute of Phonetic Sciences 17* (Amsterdam), 97–110.

[10] Benade, A. (1976). *Fundamentals of Musical Acoustics*. Oxford University Press, New York.

[11] Benson, D. (2006). *Music: A Mathematical Offering*. Cambridge University Press, London, UK. An updated 2008 web version is also freely distributed here: http://www.maths.abdn.ac.uk/~bensondj/html/maths-music.html.

[12] Bracewell, R. N. (2000). *The Fourier Transform and Its Applications*. McGraw–Hill Higher Education—International Series, Singapore.

[13] Brown, J. C. (1992). Musical fundamental frequency tracking using a pattern recognition method. *J. Acoust. Soc. Am. 92*(3), 1394–1402.

[14] Burns, E. M., Ward, W. D. (1978). Categorical perception - Phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. *J. Acoust. Soc. Am. 63*, 456-468.

[15] Cano, P. (1998). Fundamental frequency estimation in the SMS analysis. *Proc. 1998 Digital Audio Effects Workshop* (ZMFX98).

[16] Cedolin, L. and Delgutte, B. (2005). Pitch of Complex Tones: Rate-Place and Interspike Interval Representations. *J Neurophysiol, 94*: 347–362. in the Auditory Nerve

[17] Chen, K. (2001). Pitch–Synchronous Overlap–Add Musical Resynthesis with Variable Time–Scaling Set by a Human Conductor. Unpublished masters thesis, Univ. of Illinois at Urbana–Champaign, Urbana, IL.

[18] Chrysanthos from Madytos. (1832) *The Grand Theoretical Book of Byzantine Music*. Michele Weis Press.

[19] Cooley, J. W. and Tuckey, J. W. (1965). An algorithm of the machine calculation of complex Fourier series. *Math. Comp. 19*, 297–301.

[20] de Cheveigné, A. (1998). Cancellation model of pitch perception. *J. Acoust. Soc. Am. 103*, 1261–1271.

[21] de Cheveigné, A. and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am. 111*(4), 1917–1930.

[22] Doval, B. and Rodet, X. (1991). Estimation of fundamental frequency of musical sound signals. *Proc. 1991 IEEE Conf. on Acoustics, Speech, and Signal Processing* (ICASSP–91), Toronto (IEEE, New York), pp. 3657–3660.

[23] Feth, L., L. (1974). Frequency discrimination of complex periodic tones. *Perception & Psychophysics, 15*(2), 375–378.

[24] Fitz, K., Walker, W., and Haken, L. (1992). Extending the McAulay–Quatieri analysis for synthesis with a limited number of oscillators. *Proc. 1992 Int. Computer Music Conf.*, San Jose, CA (Int. Computer Music Assoc., San Francisco), pp. 381–382.

[25] Fletcher, H. (1964). Normal vibration frequencies of a stiff piano string. *J. Acoust. Soc. Am. 36*(1), 203–209.

[26] Gulick, W., L. (1971). *Hearing—Physiology and Psychophysics*. Oxford University Press, New York.

[27] Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE 66*(1), 51–83.

[28] Hawkins, H., L., McMuller, T., A., Popper, A., N., Fay, R., R. (1996). *Autodory Computation*, Springer Handbook of Auditory Research, New York.

[29] Hess, W. (1983). *Pitch Determination of Speech Signals*. (Springer–Verlag, New York).

[30] Horner, A., Beauchamp, J., and So, R. (2004). Detection of Random Alterations to Time–Varying Musical Instrument Spectra. *J. Acoust. Soc. Am. 166*(3), 1800–1810.

[31] Kay, S. M. (1998). *Fundamentals of Statistical Signal Processing—Detection Theory*, Prentice Hall Signal Processing Series.

[32] Kay, S. M. (1993). *Fundamentals of Statistical Signal Processing—Estimation Theory*, Prentice Hall Signal Processing Series.

[33] Keppel, G. and Wickens, T., D. (2004). *Design and analysis—A researcher's Handbook*, 4–th Edition, Pearson Prentice Hall, New Jersey.

[34] Kinsler, L. E., Frey, A. R., Coppens, A. B. and Sanders, J. V. (1999). *Fundamentals of Acoustics*, 4–th Edition, John Wiley & Sons.

[35] Lattard, J. (1993). Influence of inharmonicity on the tuning of a piano–Measurements and mathematical simulation. *J. Acoust. Soc. Am. 94*(1), 46–53.

[36] Loven, F. (2009). *Introduction to normal auditory perception*, Delmar Cengage Learning, New York.

[37] Luce, D. A. (1963). Physical Correlates of Non–Percussive Musical Instruments, PhD dissertation, Massachusetts Institute of Technology, Cambridge, M. A.

[38] Luce, D. A. (1975). Dynamic Spectrum Changes of Orchestral Instruments. *J. Audio. Eng. Soc. 23*(7), 565–568.

[39] Lyons, R. G. (2009). *Understanding Digital Signal Processing*, Pearson Education, Inc., New Jersey.

[40] Maher, R. C. (1989). An approach for the separation of voices in composite musical signals, unpublished Ph.D. dissertation, University of Illinois at Urbana–Champaign, Urbana, IL.

[41] Maher, R. C. and Beauchamp, J. W. (1994). Fundamental frequency estimation of musical signals using a two–way mismatch procedure. *J. Acoust. Soc. Am. 95*(4), 2254–2263.

[42] McAdams, S., Beauchamp, J. W., and Meneguzzi, S. (1999). Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *J. Acoust. Soc. Am. 105*(2), 882–897.

[43] McAulay, R. J., and Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans, on Acoustics, Speech, and Signal Processing 34*(4), 744– 754.

[44] Micheyl, C., Divis, K., Wrobleski, D., M., and Oxenham, A., J. (2010). Does fundamental–frequency discrimination measure virtual pitch discrimination? *J. Acoust. Soc. Am. 128*(4), 1930–1942.

[45] Moore, B. C. J. (2003). *Introduction to the psychology of hearing*, 5–th Edition, Academic Press, London.

[46] Moore, B. C. J., Glasberg, B. R., and Flanagan, H., J. (2006). Frequency discrimination of complex tones; assessing the role of component resolvability and temporal fine structure. *J. Acoust. Soc. Am. 119* (1), 480–490.

[47] Moore, B. C. J., Glasberg, B. R., and Peters, R. W. (1985). Relative dominance of individual partials in determining the pitch of complex tones. *J. Acoust. Soc. Am. 77*(5), 1853–1860.

[48] Moorer, J. A. (1974). The optimum comb method of pitch period analysis of continuous digitized speech. *IEEE Trans, on Acoustics, Speech and Signal Processing ASSP–22*(5), 330–338.

[49] Moorer, J. A. (1975). *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer*, Report No. STAN–M–3, Center for Computer Research in Music and Acoustics (CCRMA), Dept. of Music, Stanford, CA.

[50] Noll, A. M. (1967). Cepstrum pitch determination. *J. Acoust. Soc. Am. 41*(2), 293–309.

[51] Nuttall, A. H. (1981). Some windows with very good sidelobe behavior. *IEEE Trans, on Acoustics, Speech, and Signal Processing ASSP–29*(1), 84–91.

[52] Patriarchal Byzantine Music Committee (1883). *Fundamental Teaching of Ecclesiastical Music*. Greek Orthodox Ecumenical Patriarchate Press.

[53] Panagiotopoulos, D. (1981). *Theory and Practice of Ecclesiastical Byzantine Music*, Soter, Athens, Greece.

[54] Papoulis, A. (1962). *The Fourier Integral and its Applications*, McGraw-Hill Companies.

[55] Pisarenko, V. F. (1973). The retrieval of harmonics from a covariance function. *Geophys. J. R. Astr. Soc. 10*, 347–366.

[56] Piszczalski, M., and Galler, B. A. (1979). Predicting musical pitch from component frequency ratios. *J. Acoust. Soc. Am. 66*(3), 710–720.

[57] Plomp, R. (2002). *The intelligent ear—On the nature of sound perception*, Lawrence Erlbaum Associates, Publishers, London.

[58] Plomp, R. (1967). Pitch of complex tones. *J. Acoust. Soc. Am., 41*: 1526.

[59] Pressnitzer, D., Patterson, R. D., and Krumbholz, K. (2001). The lower limit of melodic pitch. *J. Acoust. Soc. Am, 109*, 2074–2084.

[60] Quinn, B. G. (1997). Estimation of frequency, amplitude, and phase from the DFT of a time series, *IEEE Trans. of Signal Proc. 45*, 3, 814–817.

[61] Quinn, B. G., and Hannan, E. J. (2001). *The Estimation and Tracking of Frequency*. Cambridge University Press—Cambridge Series in Statistical and Probabilistic Mathematics.

[62] Quinn, B. G., and Fernandes J. M. (1991). A fast efficient technique for the estimation of frequency, *Biometrika 78*, 65–74.

[63] Roads, C. (1996). *The Computer Music Tutorial* (MIT Press, Cambridge, MA).

[64] Rossing, T., D., Moore, F., R. and Wheeler, P., A. (2002). *The Science of Sound*, 3–rd Edition, Pearson Education Inc., publishing as Addison Wesley.

[65] Ritsma, R. J. (1962). Existence region of the tonal residue. *J. Acoust. Soc. Am, 34*, 1224–1229.

[66] Rife, D. C. and Vincent, G. A. (1970). Use of the discrete Fourier transform in the measurement of frequencies and levels of tones. *Bell. Syst. Tech. J. 49*, 197–228.

[67] Sakai, H. (1984). Statistical Analysis of Pisarenko's method for sinusoidal frequency estimation. *IEEE Trans. on ASSP 32*, 95–101.

[68] Schmidt, R. O. (1981). A signal subspace approach to multiple emmiter location and spectral estimation, Stanford University Ph.D. Thesis.

[69] Schmidt, R. O. (1986). Multiple emmiter location and signal parameter estimation. *IEEE Trans. Antennas Propag. 34*, 276–280.

[70] Schroeder, M. R. (1999). *Computer Speech: Recognition, Compression, Synthesis*, Springer, New York.

[71] Schuster, A. (1898). On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Terrestrial Magnetism and Atmospheric Electricity, 3*, 13-41.

[72] Serra, X. (1989). *A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition*, Report No. STAN–M–58, Center for Computer Research in Music and Acoustics (CCRMA), Dept. of Music, Stanford, CA.

[73] Sethares, W., A. (2007). *Rhythms and Transforms*, First Edition, Springer, New York.

[74] Shumway, R., H., and Stoffer, D., S. (2006). *Time Series Analysis and Its Applications—with R Examples*, Second Edition, Springer Texts in Statistics, New York.

[75] Smith, A. T., May, J. G., and Lyman, R. R. (1983). Pitch shifts contingent on the modulation frequency of an adaptation tone. *J. Acoust. Soc. Am. 73*, 2, 691–693.

[76] Smith, J. O., and Gossett, P. (1984). A flexible sampling–rate conversion method. *Proc. 1984 IEEE Conf. on Acoustics Speech, and Signal Processing* (ICASSP–84), San Diego (IEEE, New York), pp. 19.4.1–19.4.2.

[77] Smith, J. O., and Serra, X. (1987). PARSHL: An analysis/synthesis program for non–harmonic sounds based on a sinusoidal representation. *Proc. 1987 Int. Computer Music Conf.*, Urbana, IL (Int. Computer Music Assoc., San Francisco), pp. 290–297.

[78] Sirong, W. and Clark, M. (1967a). Synthesis of Wind–instrument Tones. *J. Acoust. Soc. Am. 41*(1), 39–52.

[79] Stevens, S., S. and Davis, H. (1983). *Hearing—Its psychology and physiology*, Published by the American Institute of Physics for the Acoustical Society of America (First printing 1938).

[80] Stoica, P. and Moses, R. (2005). *Spectral Analysis of Signals*. Pearson—Prentice Hall, Upper Saddle River, New Jersey.

[81] Strong, W. and Clark, M. (1967b). Perturbations of Synthetic Orchestral Wind–instrument Tones. *J. Acoust. Soc. Am. 41*(2), 277–285.

[82] Sundberg, J. (1974). Articulatory interpretation of the 'singing formant'. *J. Acoust. Soc. Am. 55*(4), 838–844.

[83] Surmani, A., Surmani, K. and Manus, M. (2004). *Alfred's essentials of music theory: A complete self–study course for all musicians*, Alfred Publishing Company, Van Nuys, CA.

[84] Tabachnick, B., G. and Fidell, L., S. (2007). *Using Multivariate Statistics*, Fifth Edition, Pearson International Edition, Boston.

[85] Tsiappoutas, K. M. (2004). Byzantine music intervals: An experimental signal processing approach. Unpublished Masters Thesis, Univ. of New Orleans, New Orleans, LA.

[86] Tsiappoutas, K. M., Ioup, G. E., Ioup, J. (2006). Frequency tracking of ecclesiastical Byzantine music frequency intervals. *J. Acoust. Soc. Am., 119*, 3440.

[87] Tsiappoutas, K. M., Ioup, G. E., Ioup, J. (2004). Measurement and analysis of Byzantine chant frequencies and frequency intervals. *J. Acoust. Soc. Am., 116*, 2581.

[88] Wallesz, E. (1961). *A History of Byzantine Music and Hymnography*. Clarendon, London.

[89] Ward, W.D. (1954). Subjective musical pitch. *J. Acoust. Soc. Am. 26*, 369-380

[90] Weir, C., Jesteadt, W., and Green, D.M. (1977). Frequency discrimination as a function of frequency and sensation level. *J. Acoust. Soc. Am. 61*, 178–184.

[91] Zwicker, E., Flottorp, G. and Stevens, S., S. (1957). Critical bandwidth in loudness summation. *J. Acoust. Soc. Am. 29*: 548.

Vita

Kyriakos Michael Tsiappoutas was born in Cyprus in 1977. He holds a Bachelor's degree from the University of New Orleans in Psychology with a minor in Physics (2002), a Masters degree from the University of New Orleans in Applied Physics (2004), and a Masters degree in Quantitative Psychology (statistics for social science) from Illinois State University (2007). He received lessons in Byzantine music under Hatzisavvas Psaltis.

He is married to Professor Elisabeta Pana and has a son, Alexander Tsiappoutas, his greatest achievement. He is currently a *Statistical Research Analyst* with State Farm Insurance Companies, Bloomington, Illinois.