Physically Informed Subtraction of a String's Resonances from Monophonic, Discretely Attacked Tones : a Phase Vocoder Approach

by

Matthieu Hodgkinson

NUI MAYNOOTH
Ollscoil na hÉireann Má Nuad

A thesis presented in fulfilment of the requirements for the Degree of
Doctor of Philosophy

Supervisor: Dr. Joseph Timoney

Department of Computer Science
Faculty of Science and Engineering
National University of Ireland, Maynooth
Maynooth, Co.Kildare, Ireland

May, 2012

"I do not define time, space, place, and motion, as being well known to all."

Isaac Newton

Declaration

This thesis has not been submitted in whole or in part to this or any other university for any other degree and is, except where otherwise stated, the original work of the author.

Signed: ____

Matthieu Hodgkinson

Abstract

A method for the subtraction of a string's oscillations from monophonic, plucked- or hit-string tones is presented. The remainder of the subtraction is the response of the instrument's body to the excitation, and potentially other sources, such as faint vibrations of other strings, background noises or recording artifacts. In some respects, this method is similar to a stochastic-deterministic decomposition based on Sinusoidal Modeling Synthesis [MQ86, IS87]. However, our method targets string partials expressly, according to a physical model of the string's vibrations described in this thesis. Also, the method sits on a Phase Vocoder scheme. This approach has the essential advantage that the subtraction of the partials can take place "instantly", on a frame-by-frame basis, avoiding the necessity of tracking the partials and therefore availing of the possibility of a real-time implementation. The subtraction takes place in the frequency domain, and a method is presented whereby the computational cost of this process can be reduced through the reduction of a partial's frequency-domain data to its main lobe. In each frame of the Phase Vocoder, the string is encoded as a set of partials, completely described by four constants of frequency, phase, magnitude and exponential decay. These parameters are obtained with a novel method, the Complex Exponential Phase Magnitude Evolution (CSPME), which is a generalisation of the CSPE [SG06] to signals with exponential envelopes and which surpasses the finite resolution of the Discrete Fourier Transform. The encoding obtained is an intuitive representation of the string, suitable to musical processing.

Acknowledgments

Tout d'abord, je voudrais exprimer une reconnaissance profonde envers ma mère et mon père, qui n'ont jamais cessé de me montrer et de me donner leur support. Mais je voudrais aussi adresser une note particulière à André, qui en quelque sorte a tenu pour moi, sans que je m'en rende compte d'abord, un rôle de mécène au cours de mes années d'université. J'en profite aussi pour rendre hommage à un grand ami, Nicolas Perrin, qui nous a quitté Novembre dernier, et qui, j'en suis sûr, aurait été très fier de dire que, ça y'est, Hodgkinson, il est docteur. Et moi Nico, j'aurais été fier de pouvoir te rendre fier comme ça.

Back in Ireland, I would like to thank my supervisor Dr. Joseph Timoney, for the guidance he gave me throughout my PhD. And beyond my supervisor, there is the entire department of Computer Science of NUI Maynooth, with its Head, lecturers, technicians and secretaries, who have all been admirably helpful. I guess you get this nowhere like in Ireland.

Foreword

The sensation of sound occurs when our tympanic membrane is set in a vibrational motion of appropriate amplitude and frequency. The vibration of our tympanic membrane is normally the effect of an analogue oscillation in the ambient air pressure, and thereby it can be said that the medium of sound is air. However, sound can also be experienced underwater, so water can also be the medium of sound. In fact, even solids transmit sound: shake a light bulb beside your ear, and even though it is hermetically enclosed in glass, you will hear the filament shaking.

Sound has a number of media. In the 20th century, the domestication of electricity as a medium for sound has revolutionised our experience of music. Now sound can be stored, processed, and even generated. Electronic machines with keyboard interfaces use mathematical algorithms to create tones previously unheard. In the second half of the past century, these *synthesizers* become so popular as to grow in a family of instruments of their own. Most notably, the success of the Moog synthesizer, commercialised first in the 1960s [PT04], is monumental. The Doors, the Beatles, Pink Floyd, to name just a few, are all names this instrument contributed to the success of.

Technically speaking, the Moog uses a paradigm for the synthesis of musical signals called *subtractive synthesis*, whose elementary principle is to generate a harmonic or near-harmonic waveform of rich spectral properties, and filter it thereafter in inventive manners. Other paradigms used in the 60s, 70s and later include additive synthesis and FM (Frequency Modulation) synthesis [Cho73]. During this time, computers become more powerful and affordable, and digital sound synthesis emerges. A remarkable development of that time in sound synthesis is the birth of *physical modeling*, which aims at emulating in a computer the physical mechanisms that lead to the production of sound waves. Hiller and Ruiz take a Finite Difference (FD) approach to approximate the solutions to the Partial Differential Equations (PDEs) derived from the physical analysis of vibrating bodies [HR71a, HR71b]. In the early 80s, Karplus and Strong introduce a digital system which, with a simple delay line and averaging filter arranged in a feedback loop, produce tones whose resemblance with string tones is uncanny. The method inspires Julius Smith to develop the theory of *Digital Waveguide Synthesis* [III92], which models d'Alembert's (1717-1783) solution to the *wave equation* in delay lines, connected in a digital loop enriched with various types of filters to emulate the various phenomena undergone by the waves during their propagation: frequency-dependent dissipation of energy, dispersion, and so on.

The developments of this already successful sound synthesis paradigm did nevertheless not stop here. In string tones, the instrument's body was problematic in the sense that it could not practically be modeled as a 3-dimensional waveguide, meanwhile contributing – in some cases significantly – to the timbre of the instrument. This problem was worked around with the advent of Commuted Waveguide Synthesis (CWGS), which stores the response of the body in a wavetable, and uses it as the input to the digital waveguide [KVJ93, Smi93]. The method shows a potential to producing virtual instrumental parts hardly distinguishable from acoustic recordings, and this, in real time. Examples of the method can be found at the Web address http://www.acoustics.hut.fi/~vpv/ (latest access: September 10th, 2011).

Commuted Waveguide Synthesis requires the indirect response of the instrument's body to the excitation of the string. This response is obtained through a preliminary process commonly known as *excitation extraction*. A string instrument's note is recorded, and all sinusoidal components that are not part of the body's response are canceled. The residual is a burst of energy, very short in some cases, slightly longer when the body shows reverberant qualities, but rarely exceeding a second. In the cases where the string is materially flexible enough to show negligible inharmonicity, when the harmonics are not too numerous and when the body does not show prominent resonances, using the inverse of a string model [KVJ93] or a simple Sinusoidal Modeling approach [MQ86, IS87] can yield satisfying results. However, when some or none of these conditions are met, an unwanted residual of the string's resonances threatens to remain. Moreover, the interpolative nature of Sinusoidal Modeling Synthesis prevents the real-time processing of the input.

The intent of this thesis was initially to devise an automated method for excitation extraction. A Phase-Vocoder approach proved convenient, and in addition brought the possibility of a real-time implementation within reach. This possibility raised the question "What could real-time add to excitation extraction?". Live musical effects was the answer. But live conditions are different to the studio conditions where tones suitable for subsequent use in digital waveguides should be recorded. Thereout emerged the paradigm of *string extraction.*

Publications

The following is an ordered list of the publications of the author:

1. **Matthieu Hodgkinson**, Jian Wang, Joseph Timoney, and Victor Lazzarini "Handling Inharmonic Series with Median-Adjustive Trajectories", *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09)*, Como, Italy, 2009.

2. **Matthieu Hodgkinson**, Joseph Timoney, and Victor Lazzarini "A Model of Partial Tracks for Tension-Modulated, Steel-String Guitar Tones", *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, 2010.

3. **Matthieu Hodgkinson** "Exponential-Plus-Constant Fitting based on Fourier Analysis", *JIM2011 – 17èmes Journées d'Informatique Musicale*, Saint-Étienne, France, 2011.

Contents

List of Figures

List of Tables

Chapter 0

Introduction

This thesis proposes a novel method to extract the oscillatory components issued from a string's vibration in monophonic, plucked or hit string tones. The method operates within a Phase Vocoder time-frequency representation of the input tone, where a vertical process is repeated on a frame-by-frame basis, independently of the state of previous or following frames. This process consists of three steps: the detection and identification of the string's partials; the measurement in frequency, phase, magnitude and exponential envelope of these partials; and their frequency-domain re-synthesis and subtraction. Upon completion, the tone is decomposed in a "string" part, that consists of a set of partials that are completely determined by the above-mentioned parameters, and the rest of the tone, which includes stochastic elements and often sinusoidal elements as well, such as resonances of the instrument's body or faint vibrations of other strings. Conceptually, this process is reminiscent of *excitation extraction* [LDS07], but here the term *string extraction* is preferred, for reasons that are going to be developed in the first section of this introductory chapter. Following this conceptual clarification, we will delineate the range of tones that are suitable inputs to the method, suggest applications of the method, and finally present the plan of the main body of this thesis.

0.1 String extraction: conceptual definition

String extraction in a way can be regarded as some sort of sound source separation, where there are two entities to separate, one simple, and one complex. The simple entity is the "string" entity, and the complex entity is "all the rest". We cannot readily give a specific name to the latter, because itself may be decomposed into several sub-entities: the response of the instrument's body to the excitation, some vibrations of the other strings, a recording noise floor, ambient noises, and so on. The reason why we nevertheless group all these components into one entity is because our aim is simply to extract the string, and it does not matter what this rest is – so long as it is not so invasive that it compromises the working of our method, which looks for a prominent time-frequency structure to the tone that it can associate to a string model.

The signal processing paradigm closest to ours would presently be *excitation extraction*. This paradigm was probably motivated by the advent of Commuted Waveguide Synthesis (CWGS) [KVJ93, Smi93] in the early 90s, although it is conceptually close to the decomposition of a signal into deterministic and stochastic parts, which stems back to the mid-80s [RPK86, MQ86, IS87]. CWGS is a sound synthesis method for plucked or hit string instruments where the string is modeled as a system of filters and the response of the instrument's body is used to excite the model. It is thereby that the response of the body to the excitation is itself seen, in CWGS, as an excitation. "Body response" and "excitation" may therefore be used interchangeably.

To obtain this excitation, a standard deterministic-stochastic decomposition may be used. However, there is a potential risk in this approach that resonances of the body are mistaken for string partials, and mistakenly taken away from the instrument's body response. To avoid this, a model of the string can be used as a guide to deciding whether a partial belongs with the string or not. But then another risk arises, that sympathetic vibrations of other strings are seen as part of the response of the body, which they are not either.[1] In our opinion, so long as a method cannot distinguish between body resonances and sympathetic vibrations, an automatic method for excitation extraction cannot be devised unless the assumption is made that the other strings of the instrument are muted. Even then, stochastic energy that is not part of the body response, such as the noise floor of the recording or accidental ambient noises, might remain in the excitation. Another assumption therefore has to be made is that the recording is of such quality that any sound components that are part of neither the excitation nor the targeted string are inaudible.

[1] Unless muted, other strings are likely to respond indirectly to the excitation of the target string and to vibrate sympathetically, because of the transmission of vibrations through the bridge of the instrument.

Such optimal constraints can only be satisfied in carefully arranged studio conditions. These constraints are relaxed in the context of string extraction, extending the range of applications of the paradigm. But before the potential applications are discussed, a moment should be spent to determine what types of input are suitable to string extraction, both conceptually and technically.

0.2 Suitable inputs

Before we begin discussing what inputs are suitable inputs, the distinction should be made between inputs that are suitable at a *conceptual* level – from which it makes sense to remove the vibrations of one or more strings – and those suitable at a *technical* level – from which is is technically *possible* to remove the vibrations of the string(s). An example of an input suitable at both levels would be a pluck of a guitar string isolated in time. An example of input suitable at a conceptual level, but not at a technical level, could be a piano piece, because it is polyphonic, and our method currently does not support polyphonic input. The distinction between conceptual and technical suitability is thereby easy to make: a input, even if conceptually suitable, will only be technically suitable if the technical means are built into the method to deal with its complexity. What is the condition for conceptual suitability, on the other hand, is not evident, and should be briefly discussed here.

The rumble of a train, one would probably agree, is not suitable to string extraction. But what about a bowed violin melody? Separating the string from the indirect response of the body to the bow and any other sort of non-string sounds (such as the breathing of the violinist) makes sense, and surely could have numerous applications. However, we are going to draw the line around *discretely attacked* string tones. If this line were not drawn, then it could be argued that sung tones are also suitable, and speech tones, and so on, which would then make the string extraction paradigm a reduction of source-filter modeling. In a way, it is, but as the reader will see by the reading of Chapter 1, it is also an "augmentation" of source-filter modeling, because string extraction relies on a string's physical model, and the time-frequency data collected during the process is given meaning through its association to this model. A plucking position, for example, can be inferred from the notches found in a string's comb-like spectrum [TI01], or a gain spectrum can be derived from the measurement of the decay rates of the partials [KVJ93, VHKJ96], which may turn out to be typical of a Spanish guitar or a plucked double bass. These are all timbre attributes that make the specificity of discretely-attacked string tones, to the point that, even played in isolation of the body response, the sinusoidal structure of a string still very much sounds like a string. Conceptually suitable inputs are therefore inputs whose time-frequency characteristics are inherent to the physical model that will be described in Chapter 1: plucked or hit string tones, simply.

So ideally, we would present in this thesis a method that is flexible enough to be able to deal with polyphonic parts of plucked string instruments. However, the intent was initially to devise an automated method for excitation extraction for subsequent use in digital waveguides, and as such, was only meant for monophonic tones – the conceptual generalisation to *string* extraction only came at an advanced stage of the thesis' genesis. As for all sinusoidal analysis methods based on the FFT, the good resolution of the partials is assumed in our method, and the difficulty of overlapping partials may therefore require an entirely different approach, which is beyond the scope of this thesis. Now the question must be asked as to what types of inputs our method is technically capable of dealing with. Only after light has been shed on this point can current applications can be discussed.

Chapter 4 tests the string extraction method described throughout this thesis. All these tests were run upon monophonic string tones, each recording featuring one note only. The time structure of the physical model developed in this thesis is an approximation whereby the string's vibrations are nil until time $t = 0$, when they are instantly set into sinusoidal motion. In our processing, this is viewed as a unit-step windowing of the sinusoidal motion of the partials, which otherwise would have been vibrating ever since $t = -\infty$. The tone stops when all the vibrational energy is dissipated, when the string is muted, or when the attack is renewed, either on a same note or a different note. The muting of a note or its interruption by the plucking of an other note can also be modeled by a product with a time-reversed unit-step function. Our point here is that, even within a melodic phrase, a note can be taken out of its context to reproduce the testing conditions of Chapter 4. The idea is illustrated in Figure 1.[2]

Finally, some string instruments such as the piano, but also the dulcimer

[2] In this figure, the unit-step windowing looks like rectangular windowing, but in our processing mindset, we look at tones at the time scale of a STFT window, where it is unlikely that the analysed tone is both attacked and muted. In this case, a unit-step windowing expression seems more appropriate.

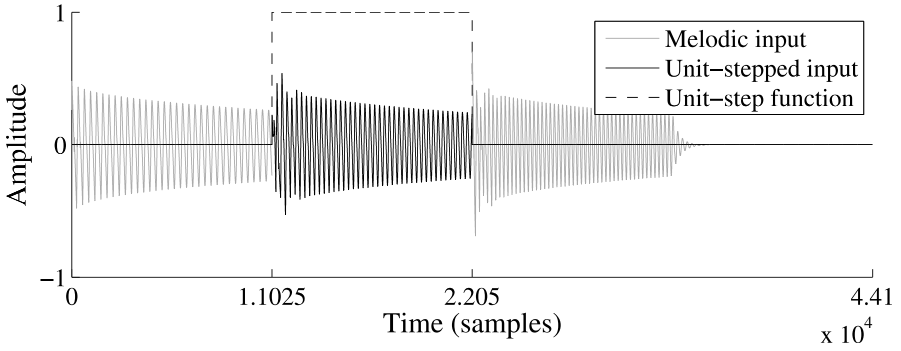

Figure 1: Isolating a note from a melody with unit-step windowing.

or the harpsichord feature *courses* of strings, arrangements of two or more strings which vibrate together in the production of one same note. There are several reasons for such facture, mainly a gain of loudness, but also an increase of the "depth" of the sound, produced by strings that are very nearly, but not exactly, tuned to unison. This is problematic for our method, which has been devised in this thesis to deal with the pseudo-harmonic series of one string only. In our collection of piano and harpsichord tones, the information was not readily available whether tones were the contribution of a single string or of courses of string. This consideration could therefore not be made on a tone-by-tone basis, but it may account for some of the sense of pitch that remained after string extraction in our least successful examples. With regard to the extent of this thesis, we will consider that tones issued from courses of strings are also eligible, only we will assume perfect unison tuning. Without this assumption, a number of highly regarded instruments such as the piano or the harpsichord would be excluded, while our method still gives satisfying results for the greater part of their range.

0.3 Applications of the present method for string extraction

In summary of the previous section, the inputs suitable for our method of string extraction are monophonic sequences of discretely-attacked string tones. These sequences may reduce to one note only, or may be entire melodies. Albeit restrained by the monophonic limitation, our method can already be the starting point for a range of musical effects. In the view of this type of application, the method has been developed as much as possible to facilitate a real-time implementation. A Phase Vocoder approach – this paradigm will be presented in detail in Chapter 3 – has therefore been preferred to a Sinusoidal Modeling Synthesis approach [MQ86, IS87], where a real-time implementation is compromised by the fact that it uses interpolation between measurement points to cancel the partials. This means that, at frame a, it will have to wait for frame b – "much" later – to be processed before interpolation and cancelation takes place. In our approach, the cancelation takes place on a frame-by-frame basis. This does not reduce the latency to zero, as the buffering of a few fundamental periods of the tone is still required for the frequency-domain analysis and processing to take place, but it reduces it substantially. How audible and inconvenient this minimal latency is to the real-time effects made achievable by the method, and how it can be worked around, is yet an open question.

The applications that gain from a real-time implementation are essentially musical applications. The subtractive nature of our string extraction method, as well as the underlying physical model and the original techniques employed, render many "traditional" effects very accessible, and also bring about original, physics-based effects. For example, a pitch-shifter/harmoniser the likes of which is found in [LD99], but readily benefitting from the detection and cancelation of the partials of the string extraction process, seems now fairly straightforward to implement, with the added guarantee that potential sympathetic vibrations or body resonances, themselves, remain at their original pitch. In combination with a pitch-shifter could come an "inharmonicity modifier": the frequency of a string partial is not only determined by the fundamental frequency, but also by the Inharmonicity Coefficient (IC). A method original to this thesis, the Median-Adjustive Trajectories (MAT), integrates the estimation of this coefficient in the peak detection. Upon string extraction, the IC can be modified, virtually altering the physicality of the string.

Another effect that the various innovations of this thesis may offer is a real-time "sustain stretcher". In contrast with a standard time stretcher, a sustain stretcher would modify the decay rates of the partials without slowing down or accelerating the frequency glide of the partials typical of tension-modulated tones [HTL10]. This is achievable in real time thanks to the CSPME, a generalisation of the CSPE [SG06] to exponential-amplitude signals. This novel method returns the decay rate of partials on the basis of *a single frame*, allowing for a modification "on the spot" of the decay rate, and hence of the sustain, of a string's partials.

More advanced, ambitious effects can also be considered. For example, a quadratic model with a zeroth-order and a second-order coefficient that dictate the frequency-dependent trend of decay rate of string partials is developed in [CA93]. Our fits of Section 1.2.3 show that the second-order coefficient tends to be bigger for nylon strings than for steel strings. This could inspire a morphing effect achieved by the interpolation of the coefficients, conferring progressively to an acoustic guitar a Spanish-guitar like decay rate spectrum.

This listing of applications is, of course, not exhaustive. In fact, the method's range of applications extends to processing useful for analysis, typically done offline. For instance, it was already said that the character and the fine detail of the response of the body only becomes audible after the string has been extracted. For example, an electric guitar's body response is very short and dry, while a harpsichord's is relatively long and reverberant. Being able to proceed to such decomposition is therefore of interest to the student and researcher acoustician alike, in that it offers a privileged insight into the composition of string tones. Similarly, recording artifacts may only become obvious post-processing, which makes of our method an interesting tool for the quality assessment of a tone. It has been found, for instance, that the string extraction process often exposed background noises in samples of professional standard.

0.4 Organisation of the thesis

This thesis is articulated in four chapters. First, an analytical model for the vibrations of the string will be developed. The role of this chapter will primarily be to answer the question: “What are we looking for?”. Where, in frequency, are the string partials to be found, what kind of frequency-dependent magnitude distribution is to be expected, how do string partials evolve in time, etc., are all questions that can inspire appropriate detection, measurement and cancelation methods. Another important aspect of this chapter is the physics-based musical applications that it can inspire. In this regard, the reader will find that some features of the physical model developed in Chapter 1 are not found in the string extraction method *per se*. The fundamental frequency and inharmonicity coefficient are, for example, essential to the detection and identification of the partials, but the method is ignorant of the attack-point-dependent comb-like shape of the spectrum, to whose explanation Section 1.1.3 is dedicated altogether. The design of a musical effect that consists of virtually displacing the position of the attack along the string, however, would benefit of all there is to know about this phenomenon.

The second chapter will deal with the low-level topic of physical analysis. The first half of this chapter is dedicated to the topic of analysis windows. Such an extensive development will be found justified *a posteriori* in the description of our method, which relies heavily on the properties of these windows, both in the time-domain (e.g. the constant-sum properties of cosine windows, indispensable to a transparent Phase Vocoder scheme) and the frequency-domain (e.g. the expansion of a partial's spectrum from the four parameters listed previously is only possible with an analytical expression for the analysis window's spectrum). Following this, a transition will be operated from the continuous-time domain (convenient for the development of analytical expressions) to the discrete-time domain (where our method's processing will take place), with a discussion on the relevant specificities of discrete signals. This second chapter will close with the selection of a method, among the well-established and more original ones, for the estimation of the parameters of the partials, all-important for their good cancelation.

The third chapter will give a description of the method. A short formulation of the Phase Vocoder scheme will be given that is a reduction of the general formulation [Por81] to a constant-rate scheme and whose notation has been made consistent with the notation of the rest of this thesis. The focus will then be put on steady-state frames, and after that, on frames that overlap with the onset of the sound. In both cases, a frequency-domain method for the cancelation of the partials will be introduced. At that stage, there will remain to approach and solve the problem of the detection of the peaks, as well as their identification, in terms of partial number, but also regarding whether they belong to a transverse series or a longitudinal series of phantom partials.

Chapter 4 presents the results of the method as tested on a variety of instruments of contrasting character. First, some successful results will be used to support a discussion that helps pin-pointing the concept of string extraction, this time with the help of visual and sonorous examples. Then the various methods that altogether make up the method will be examined for their individual contribution. This part of the thesis will be interesting for the understanding of our method, its strengths and its limitations, but it will also give a listing of the various innovative techniques found in this thesis which, albeit inspired by the string extraction problem at hand, may find applications in other fields of sound and signal processing as well.

The conclusion to this thesis will give a recapitulation on its aim and look back at the role of each chapter in reaching it. This conclusion will also be an opportunity to outline the various contributions of this thesis to the broader field of audio processing. Directions for future work, both short- and long-term, will finally be given.

Chapter 1

Development of a physical model

Introduction

The goal of this chapter is to derive a time-frequency model of the vibrations of the string. On the one hand, the reality that we attempt to describe is of infinite complexity. On the other hand, our model must be finite. Hence some guideline must first be established as to which aspects of the real string must figure in our model, and which it is superfluous to include.

Ideally, this guideline should be perceptibility. Not all changes in atmospheric pressure can be detected by the ear. This applies very well to the atmospheric disturbance caused by the vibration of a nearby instrumental string. As the application of string extraction is ultimately the modification of the aural quality of the string – in other words, sound effects – it is superfluous to manipulate any feature of the sonic structure at hand which, both before and after processing, is imperceptible. Such feature could be an object or a group of objects, like the highest partials of a piano string, or an aspect of objects, like the glide in the fundamental frequency of a piano string, that is inherent to the string's decay in vibrations, but is indeed inaudible.

However, even the condition that everything that is audible should be modeled is difficult to meet. To the best of our knowledge, and a much as the scope of a doctoral thesis chapter allows, we will strive to satisfy this condition. Yet at times, due to their complexity, some questions may remain to be resolved, and *ad hoc* solutions may be used instead. At other times, the disregard of some object of modest perceptible impact may bring great simplifications in our model. Hence, suitable trade-offs may become apparent towards the end of the chapter.

The development of the model will begin with the derivation of the well-known Wave Equation. Fixed boundary conditions will then be introduced, and stand as a supporting body. Optionally, plucking or hitting initial conditions can be drawn to specify a time-zero state of the string. Then can we start to refine the model, adding the phenomenon of damping, caused by air friction and internal friction, and also, the element of stiffness to our string.

By this time we are in possession of a rigorous physical model, exclusively derived from textbooks, and whose description resides in a solution that satisfies a Partial Differential Equation (PDE). The augmentation of the model with time-varying fundamental frequency and Inharmonicity Coefficient (IC), partly realised through empirical formulations, will be the departing point

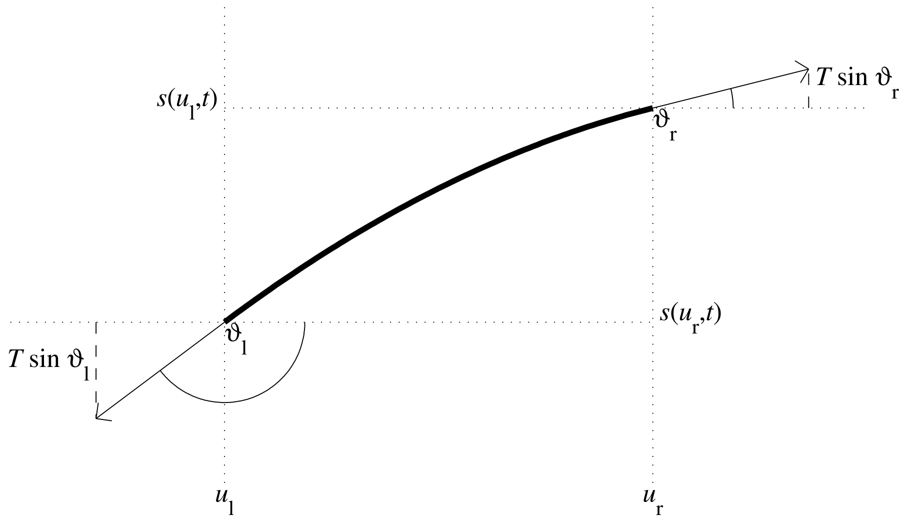

Figure 1.1: Vertical forces onto string segment

from strict consistency in physical analysis. From there, the inclusion of longitudinal partials will be considered, based on the most recent literature.

1.1 The Wave Equation for transverse vibrations

1.1.1 Derivation

The following derivation was drawn selectively upon three reference textbooks: [FR91], [Rai00] and [Ste96]. We derive the wave equation upon the sketch of a short segment of string, shown in Figure 1.1. The string as a whole is looked upon as a function of space and time, $s(u,t)$, $u = [0, L]$, where L is the length of the string.

As we are concerned here with transverse vibrations, we only consider the vertical component of the force acting at either end of the segment. Then to reach the wave equation in its partial differential form, we reduce the length of the segment to something infinitesimally small.

The vertical component of the net force applied onto the string segment is the sum of the vertical component of the tension T at either end of the segment, $T \sin \vartheta_{\mathrm{r}} + T \sin \vartheta_{\mathrm{l}}$, as shown in Figure 1.1. This force can be equated with the mass of the segment, $\mu(u_{\mathrm{r}} - u_{\mathrm{l}})$, times its acceleration, grossly denoted a for now. μ is the linear density of the string, in kilogram per meter. To summarise,

$$T(\sin \vartheta_{\mathrm{r}} + \sin \vartheta_{\mathrm{l}}) = \mu(u_{\mathrm{r}} - u_{\mathrm{l}})a. \tag{1.1}$$

Before the Wave Equation in its final form can be reached, an approximation needs to be made: $\sin \vartheta \approx \tan \vartheta$, for small ϑ, which holds provided that the vertical displacement of the string inclination is small [FR91, Rai00, Ste96]. This approximation is important in that it allows us to reach a first-order derivative, considering that $\tan \vartheta = \partial s / \partial u$. We thus rewrite (1.1) as

$$T \left(\left. \frac{\partial s}{\partial u} \right|_{u_{\mathrm{r}}} - \left. \frac{\partial s}{\partial u} \right|_{u_{\mathrm{l}}} \right) / \mu(u_{\mathrm{r}} - u_{\mathrm{l}}) = a.$$

Now we can take the limit on each side as $u_{\mathrm{r}} - u_{\mathrm{l}} \to 0$, to get

$$c^2 \frac{\partial^2 s}{\partial u^2} = \frac{\partial^2 s}{\partial t^2}, \tag{1.2}$$

where $c = \sqrt{T/\mu}$ is the propagation speed of disturbances along the string, in metres per second, and multiplies the string's curvature, the term on the right-hand side of the equal sign being the vertical acceleration of the string.

(1.2) is the Wave Equation, found in various fields of physics. Associating it with our string, it says that the curvature of the string is proportional to its acceleration. Where the curvature is negative, the string has a concave, ∩-like shape, and in such place it makes sense for the acceleration to be negative too - or downwards. Also, the acceleration is proportional to the tension, which here is the only restoring force we are considering, and is inversely proportional to the mass density of the string, which opposes its inertia to acceleration.

1.1.2 Solution for strings fixed at both ends

The intent of this section is to find a general but explicit formulation for $s(u, t)$, that satisfies not only (1.2), but also two *boundary conditions*: that the displacement remains nil where the string is attached, i.e. $s(0, t) = 0$ and $s(L, t) = 0$.

We give here the general solution to the wave equation, attributed to D'Alembert (1717-1783) [FR91]:

$$s(u, t) = f(ct - u) + g(ct + u). \tag{1.3}$$

The first boundary condition, $s(0, t) = 0$, implies that $f(ct) + g(ct) = 0$ and thus $g(u) = -f(u)$. Substituting this result in (1.3), we get

$$s(u, t) = f(ct - u) - f(ct + u). \tag{1.4}$$

From the boundary condition at the other end of the string, where $u = L$, $s(L, t) = 0$ it can be deduced that $f(ct+u) = f(ct+2L+u) = f(c(t+2L/c)+u)$, and by extension, we get $s(u, t) = s(u+2L, t) = s(u, t+2L/c)$. This says that our function s is periodic in u in $2L$, and in t in $2L/c$.

The solution we are looking for is the product of two functions. The first is periodic in $2L$ and remains 0 for both $u = 0$ and $u = 2L$, and this is $\sin(\frac{\pi}{L}u)$. The second is periodic in $2L/c$; this is $\cos(\omega_0 t + \phi)$, where

$$\omega_0 = \pi c/L \tag{1.5}$$

and ϕ is an arbitrary phase constant. The solution can be further generalised if we multiply it by an amplitude constant A. Altogether, we get:

$$s(u, t) = A \sin\left(\frac{\pi}{L}u\right) \cos(\omega_0 t + \phi) \tag{1.6}$$

(1.6) satisfies the wave equation, as can be verified by substitution into (1.2).

A couple of more steps are needed to reach the most general formulation. First, multiplying the frequency of each component function by an integer $k \in \mathbb{N}$:

$$s(u, t) = A \sin\left(k\frac{\pi}{L}u\right) \cos(k\omega_0 t + \phi).$$

Finally, the sum of any number of such functions is also a valid solution:

$$s(u, t) = \sum_{k=1}^{\infty} A_k \sin\left(k\frac{\pi}{L}u\right) \cos(k\omega_0 t + \phi_k). \tag{1.7}$$

(1.7) is the most general solution to the wave equation for a string fixed at both ends. Although this model is too simplistic for convincing string sound synthesis, it already shows the harmonic nature of string tones, because the frequency $k\omega_0$ of each component is an integer multiple of the fundamental frequency ω_0. Yet the amplitude A_k and initial phase ϕ_k of each component remains undefined. The initial conditions, that is, the displacement and/or velocity state of the string at time $t = 0$, allow those to be determined. This is the purpose of the next section.

1.1.3 Plucked and hit strings

With regard to the problem of finding the harmonics in the spectra of acoustic tones, knowing the amplitude coefficient series A_k is useful, in ways that are going to be obvious as soon as we reach explicit expressions for it. Yet A_k is not the only unknown in the general solution to the wave equation of (1.7). Notwithstanding the intellectual pleasure found in reaching the complete mathematical expression of a theoretical string, the practical purpose of finding ϕ_k, in the context of this thesis, really is to make the finding of A_k possible.

The form of the series A_k and ϕ_k depend on the type of excitation. Instruments whose strings are supposed to vibrate freely are mostly played by plucking (guitar, harp, mandolin...) or hitting (piano, dulcimer, cymbalom...) the strings. Consistently we are going to focus here on those two forms of excitation, plucking and hitting.

In each case we will proceed as follows:

1. We express the string's initial state (i.e. either the displacement or the

velocity of the string at time $t = 0$) as a function of u, and derive the corresponding Fourier series.

2. We equate $s(u, 0)$ as found in (1.7) with the inverse Fourier series of the aforementioned function.

3. We restore the time variable in the equation and find A_k and ϕ_k.

A general formulation for the Fourier series and its inverse are given in Appendix (A.3) and (A.4). However, the following developments can be greatly simplified if we reduce (A.4) to a form that is closer to (1.7).

Let us consider $z(u)$, interchangeably denoting the displacement or the velocity of the string at time $t = 0$. z is periodic in $2L$, and, according to (A.4), can be expressed as

$$z(u) = \frac{1}{2L} \sum_{k=-\infty}^{\infty} Z[k] e^{jk\pi u/L}, \tag{1.8}$$

where $Z = \mathcal{FS}\{z\}$, the Fourier series of z.

We reduce (1.8) to a one-sided inverse Fourier series expression,

$$\begin{aligned} z(u) &= \frac{1}{2L} \sum_{k=-\infty}^{\infty} Z[k] e^{jk\pi u/L} \\ &= \frac{1}{2L} \sum_{k=1}^{\infty} R[k] e^{jk\pi u/L}, \end{aligned}$$

which is acceptable for the following reasons. First, and obviously, $s(u, t)$ is real, meaning that $Z[-k] = Z^*[k]$, and hence that the negative-frequency side is redundant. Also, $s(u, t)$ is odd about zero, as can be inferred from (1.4) (see section 1.1.2): $s(u,t) = f(ct-u) - f(ct+u) = -(f(ct+u) - f(ct-u)) = -s(-u,t)$. This has the effect of annihilating the real part of $Z[k]$, implying that $Z[-k] = -Z[k]$ and that $Z[0] = 0$.

This said, and if we get rid of the summation by focusing on one value of the frequency index k only, we can write that

$$R[k]e^{jk\pi u/L} = Z[k]e^{jk\pi u/L} + Z[-k]e^{-jk\pi u/L},$$

and hence,

$$\begin{aligned} R[k]e^{jk\pi u/L} &= Z[k]\left(e^{jk\pi u/L} - e^{-jk\pi u/L}\right) \\ &= j2Z[k]\sin\frac{k\pi u}{L}. \end{aligned}$$

For the derivation of the A_k and ϕ_k series, we can now use the equality

$$z(u) = j\frac{1}{L}\sum_{k=1}^{\infty} Z[k]\sin\frac{k\pi u}{L}. \tag{1.9}$$

Plucked string

We introduce the plucking of the string as a displacement state of the string at time $t = 0$. The displacement model here is simplistic, but the results obtained with it are surprisingly faithful to reality, as will be seen by the end of this section. More sophisticated models can be used for physical modeling-based synthesis, but in our case, where we use our results as mere guides for analysis, the added complexity would be superfluous.

At time 0, we give the string the triangle-like shape seen in Figure 1.2, where A is the displacement of the string at the point where it is plucked, and u_p, the plucking point itself, along the length of the string. (It is here necessary to the Fourier analysis to define the string over a complete period, $2L$, which is the reason for the extension on the negative u side.)

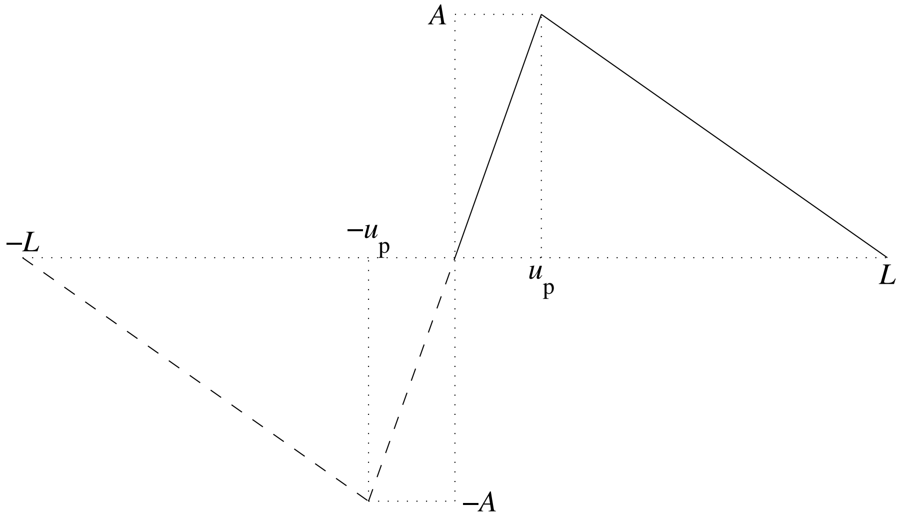

Figure 1.2: Simple model for pluck excitation

Mathematically formulated, the shape seen in (1.2) is

$$s(u,0) = \begin{cases} \frac{A}{u_\mathrm{p}-L}(u+L) & u \in [-L, -u_\mathrm{p}] \\ \frac{A}{u_\mathrm{p}}u & u \in [-u_\mathrm{p}, u_\mathrm{p}] \\ \frac{A}{u_\mathrm{p}-L}(u-L) & u \in [u_\mathrm{p}, L] \end{cases} , \tag{1.10}$$

and the corresponding Fourier series,

$$Y[k] = j\frac{A2L^3}{\pi^2}\frac{1}{(u_\mathrm{p}-L)u_\mathrm{p}}\frac{1}{k^2}\sin\frac{ku_\mathrm{p}\pi}{L}. \tag{1.11}$$

We now substitute (1.11) into (1.9), and equate with $s(u,0)$:

$$s(u,0) = \sum_{k=1}^{\infty} A_k \sin\frac{k\pi u}{L} \cos\phi_k = j\frac{1}{L}\sum_{k=1}^{\infty} Y[k] \sin\frac{k\pi u}{L}$$

Only one frequency term will be needed to find our unknowns, which we will denote $s_k(u,t)$:

$$s_k(u,0) = A_k \sin\frac{k\pi u}{L} \cos\phi_k = j\frac{1}{L} Y[k] \sin\frac{k\pi u}{L}.$$

This result has to be valid for any u. Let us get rid of the sine terms by setting $u = L/2k$:

$$s_k(L/2k,0) = A_k \cos\phi_k = j\frac{1}{L} Y[k].$$

We now need to restore the time variable to the middle term, and consistently, multiply the right-hand side with a sinusoidal term of identical frequency $k\omega_0$. The only such term to be 1 at time $t = 0$ is a cosine term, so we write

$$s_k(L/2k,t) = A_k \cos(k\omega_0 t + \phi_k) = j\frac{1}{L} Y[k] \cos k\omega_0 t.$$

As this has to hold for all t, it is necessary that $\phi_k = 0$, and hence,

$$A_k = j\frac{1}{L} Y[k]. \tag{1.12}$$

The result of this development is obtained by substitution of (1.11) into

(1.12):

$$A_k = -\frac{A2L^2}{\pi^2}\frac{1}{(u_\mathrm{p}-L)u_\mathrm{p}}\frac{1}{k^2}\sin\frac{ku_\mathrm{p}\pi}{L}. \tag{1.13}$$

There is a lot to say about (1.13). The general trend of the amplitude of the harmonics of a plucked string can be seen here to decay with the harmonic index k, but in subtle ways: the energy distribution may look like that of a sawtooth wave or a triangle wave, depending on the plucking position u_p.

To see this, consider A_k as the plucking position nears, say, the bridge (i.e. where $u = 0$):

$$\lim_{u_\mathrm{p}\to 0} A_k = A\frac{2}{\pi}\frac{1}{k}. \tag{1.14}$$

(1.14) is known to be the Fourier series of a sawtooth wave of amplitude A. In that case, the amplitude of harmonics is inversely proportional to their index.

Now if the string were plucked halfway ($u_\mathrm{p} = L/2$), (1.13) would become

$$A_k = A\frac{8}{\pi^2}\frac{1}{k^2}\sin\frac{k\pi}{2}, \tag{1.15}$$

which is known to be the Fourier series of a triangle wave of amplitude A. Here, the harmonics are inversely proportional to the square of their index. Confronting the two cases of (1.14) and (1.15), we can infer that the closest to one end the string is plucked, the greatest the harmonics' amplitude in the higher end of the spectrum, and the brightest the sound. Anyone with a minimum of experience in playing the guitar shall find this statement reflective of reality.

Another important aspect of (1.13) is the phenomenon of *nodes*. In (1.15) for example, it is very clear that, due to the $\sin \frac{k\pi}{2}$ term, all harmonics of even index $k = 2, 4, ...$ are going to be missing. More generally, for a string plucked an n^{th} of its length from the bridge, every harmonic whose index is a multiple of n is going to be missing. Figure (1.3) illustrates A_k for various plucking positions u_p, among others, those found in the triangle- and sawtooth-like cases discussed above.

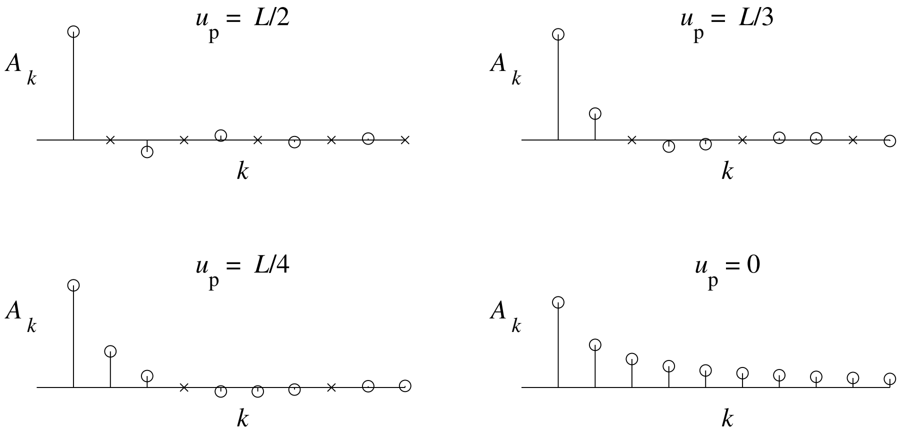

Figure 1.3: Coefficient series A_k for various plucking positions u_{p}

In relation to the problem of automatically finding the harmonics in the spectrum of acoustic tones, our result shows that, beyond a certain index, harmonics will become too faint to emerge from the noise floor inherent to any recording. Also, its is important to design an algorithm that accounts for nodes, i.e. places in the spectrum where a partial goes missing or is extremely small, in a frequency band otherwise featuring healthy harmonics.

A more refined plucking model may account for the non-zero length of the string segment which makes contact with the plectrum or the finger, smoothening the triangular shape of the string at time $t = 0.$, and hence low-pass filtering the amplitude series of (1.11) [TI01]. Indeed, it is known of guitar players that plucking the string with the meat of the finger produces a more mellow sound than with the nail. Because this physical fact has some perceptual bearing, it should be accounted for in synthesis models. However, the amplitude trend described in this section is sufficiently refined for the piece of processing proposed in this thesis.

Hit string

We picture our string again, this time, hit by a hammer at the instant $t = 0$. The displacement of the string at that time is nil, and so is its velocity, except for the hitting point u_{h}, where the velocity equals that of the hammer, v_{h}. Remember however that our string function is odd about zero, so its velocity at $-u_{\mathrm{h}} = -v_{\mathrm{h}}$. Those pieces of information together yield the construct

$$v(u,0) = \left(\delta(u - u_{\mathrm{h}}) - \delta(u + u_{\mathrm{h}})\right) v_{\mathrm{h}}, \tag{1.16}$$

illustrated in Figure 1.4.

The Fourier series V of (1.16) is easily found to be

$$V[k] = -j2Lv_{\mathrm{h}} \sin \frac{k\pi u_{\mathrm{h}}}{L}. \tag{1.17}$$

$v(u,t)$ is the time derivative of (1.7), i.e.

$$v(u,t) = -k\omega_0 A_k \sin \frac{k\pi u}{L} \sin(k\omega_0 t + \phi_k).$$

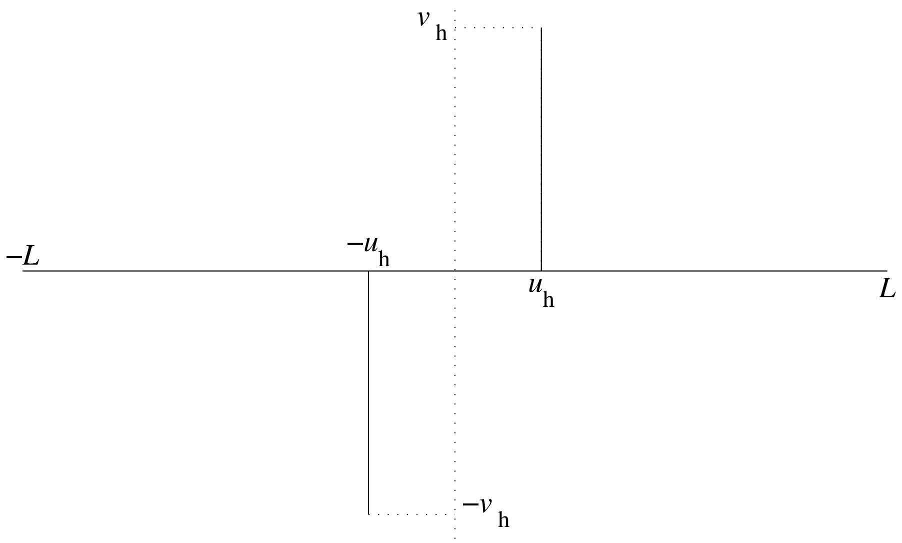

Figure 1.4: Simple model for hit excitation

As we did for the plucked-string case, we consider the k^{th} component of the velocity at $u = L/2k$ and $t = 0$, and we equate it with $V[k]$ to get the following:

$$v_k(L/2k, 0) = -k\omega_0 A_k \sin\phi_k = j\frac{1}{L}V[k].$$

We bring in the time variable, restoring t on the left-hand side, and on the right, multiplying by the only sinusoidal function of frequency $k\omega_0$ that is 1 when its argument is 0:

$$v_k(L/2k, t) = -k\omega_0 A_k \sin(k\omega_0 t + \phi_k) = j\frac{1}{L}V[k]\cos k\omega_0 t,$$

which makes it obvious that

$$\phi_k = \pi/2$$

and $-k\omega_0 A_k = jV[k]/L$, or

$$A_k = -v_{\mathrm{h}} \frac{2L}{\pi c} \frac{1}{k} \sin \frac{k\pi u_{\mathrm{h}}}{L}. \tag{1.18}$$

(1.18) tells us that the amplitude of the harmonics series is proportional to the velocity of the hammer. Individually, each harmonic's amplitude is inversely proportional to its index. Hit string tones are, on this basis, generally brighter than plucked string tones, whose harmonics' amplitude may be inversely proportional to the *square* of their index, for tones plucked near the middle of the string. Perceptually, this is a loss of 6 decibels per octave for hit strings, and between 6 and 12 decibels per octave for plucked strings, depending on the plucking position. Finally, (1.18) shows that hit-string amplitude coefficients have a sine-like behaviour, with nodes every L/u_{h} harmonics, which is identical to that found in plucked tones.

Yet again, the excitation model presented here is simplistic, albeit sufficient for our purpose. Refinements may include the non-zero width of the hammer and the string-hammer interaction beyond time 0. For an introductory discussion and further references, see [FR91].

1.2 Refinement of the model

The solution to the Wave Equation (equation (1.7) page 19) is for now too simplistic, and does not account for the time-dependent magnitude behavior of the harmonics witnessed in the analyses of recorded string tones. Especially, the energy of the solution waveform stays constant over time, never decays; this is because damping is yet absent from our model. In reality, the harmonics of freely-vibrating string tones feature a decay whose trend can be approximated with an exponential function.

Also, our series of partials is at the moment perfectly harmonic, as the frequency of the k^{th} harmonic component strictly equals k times the fundamental frequency ω_0. For most string tones, especially where the string is made of stiff material such as steel, the frequency deviation in the harmonics as the harmonic index increases tends to be such that our present model cannot even be used for such tasks as automated partial detection.

Until now we have been deriving our differential equations using graphical representations of the state of the string, working out the forces acting on its parts and later invoking Newton's Second Law to derive a partial differential equation. In contrast, we are now going to refine the basic wave equation by adding terms based on reasonable assumptions. References to literature, reasonable solutions and verification by analysis of recorded tones should be found to validate those initial assumptions.

1.2.1 Air friction

The first step taken here in the refinement of our physical model is to account for air friction. To do so, we equate the total acceleration of the string $\partial^2 s/\partial t^2$ not only with the wave propagation speed times the string's curvature, $c^2\partial^2 s/\partial u^2$, but also with an air resistance acceleration term. This term is assumed to be proportional to the velocity of the string by an unknown constant α, and opposite to the direction of the displacement:

$$\frac{\partial^2 s}{\partial t^2} = c^2 \frac{\partial^2 s}{\partial u^2} - \alpha \frac{\partial s}{\partial t}. \tag{1.19}$$

Our guess is that this air friction term is going to provoke an exponential-like decay of the harmonics, as witnessed in the analyses of recorded tones. We therefore multiply each harmonic with the term $e^{\gamma_k t}$, where γ_k is the decay rate of the k^{th} component. To simplify the writing, we consider one harmonic only $s_k(u,t)$ of the whole series $s(u,t)$,

$$s_k(u,t) = A_k \sin \frac{k\pi u}{L} \cos(\omega_k t + \phi_k) e^{\gamma_k t}, \tag{1.20}$$

where the angular frequency ω_k of the k^{th} mode of vibration is yet to be derived.

After derivation of the derivatives found in (1.19) for (1.20) and their substitution, we obtain the equality

$$\left(\gamma_k^2 - \omega_k^2 + \alpha\gamma_k + k^2\omega_0^2\right) \cos(\omega_k t + \phi_k) = \omega_k(\alpha + 2\gamma_k) \sin(\omega_k t + \phi_k),$$

ω_0 still being equal to $c\pi/L$. This can only be true if each side equals zero, and hence $\alpha = -2\gamma$ [1] and

$$\omega_k = \sqrt{k^2\omega_0^2 - \gamma^2}, \quad k = 1, 2, 3, ... \tag{1.21}$$

[1] This is consistent with the model of transverse motion of a piano string found in [CA93]

(consistently with the results in [FR91, p. 10], where the frequency of oscillation of a simple mass-spring system accounting for air friction is derived). α, like all partial differential equation coefficients in linear models, is necessarily a constant, and hence the same goes for γ, turning out to be independent on frequency (hence the disappearance of the subscript k).

Our wave equation becomes

$$\frac{\partial^2 s}{\partial t^2} = c^2 \frac{\partial^2 s}{\partial u^2} + 2\gamma \frac{\partial s}{\partial t}, \tag{1.22}$$

and our solution,

$$s(u,t) = e^{+\gamma t} \sum_{k=1}^{\infty} A_k \sin \frac{k\pi u}{L} \cos\left(\left(k^2 \omega_0^2 - \gamma^2\right)^{\frac{1}{2}} t + \phi_k \right). \tag{1.23}$$

Measurements on actual tones that verify the results obtained here are performed at the end of section 1.2.2.

1.2.2 Internal damping

The fact that the decay introduced by the air damping term is independent on frequency is not satisfying enough. In reality, higher harmonics tend to decay significantly quicker. To account for this phenomenon it is necessary to introduce an additional term to the acceleration equation, which emulates the *viscoelasticity* of the string's material [TR03, p. 45]. This additional term is proportional to the time derivative of the curvature of the string, i.e. this is a term which opposes to changes in curvature. With regard to the solution, we assume that this additional damping is going to cause the already decaying partials to be multiplied by yet another exponential, $e^{-b_{3,k}t}$, resulting each partial to have the decaying envelope $e^{\gamma_k t}$, where $\gamma_k = -b_1 - b_{3,k}$. (From now on b_1 will replace the γ found in (1.22) and (1.23), for convenience in the following development and for consistency with the literature [CA93].)

We thus express the acceleration of the string as

$$\frac{\partial^2 s}{\partial t^2} = c^2 \frac{\partial^2 s}{\partial u^2} - 2b_1 \frac{\partial s}{\partial t} + \alpha \frac{\partial^3 s}{\partial t \partial u^2}, \tag{1.24}$$

and the vibrational model, as

$$s_k(u,t) = A_k \sin \frac{k\pi u}{L} \cos(\omega_k t + \phi_k) e^{\gamma_k t}. \tag{1.25}$$

We repeat the process of the previous section, finding the derivatives for (1.25), substituting them in (1.24) and equating the cosine with the sine term to obtain that $b_{3,k} = k^2 b_3$, with $b_3 = \alpha\pi^2/2L^2$, and $\omega_k = \left(k^2\omega_0^2 - (b_1 + b_3 k^2)^2\right)^{\frac{1}{2}}$. The differential equation becomes

$$\frac{\partial^2 s}{\partial t^2} = c^2 \frac{\partial^2 s}{\partial u^2} - 2b_1 \frac{\partial s}{\partial t} + \frac{2L^2}{\pi^2} b_3 \frac{\partial^3 s}{\partial t \partial u^2},$$

and the solution,

$$s(u,t) = \sum_{k=1}^{\infty} A_k \sin \frac{k\pi u}{L} \cos\left(\left(k^2\omega_0^2 - \gamma_k^2\right)^{\frac{1}{2}} t + \phi_k\right) e^{\gamma_k t}, \tag{1.26}$$

where

$$\gamma_k = -b_1 - b_3 k^2. \tag{1.27}$$

The decay rate associated with the internal damping is now shown to be proportional to the square of the harmonic index, which seems consistent with reality *a priori*. Also, it is interesting to notice here that the frequency deviation is dependent on the decay rate, whether this decay rate is the effect of air damping, internal damping, or both.

1.2.3 Evaluation of the damping model

According to our model, the decay rate series of the harmonics of a given tone is a quadratic polynomial in k. We therefore experimented on tones issued from a Spanish and an acoustic guitar[2] to find b_1 and b_3 for a best fit in the least-square sense[3]. The interest in the experiment is to find whether this quadratic decay rate model is suitable in reality, as well as to get an impression for the order of magnitude of b_1 and b_3, and hence evaluate the importance of the inharmonicity due to damping.

The measurements and fits on an open E4 string tone (*treble E*) for the (nylon-string) Spanish and (steel-string) acoustic guitars are shown in figures 1.5 and 1.6, respectively. We observe that the internal damping coefficient b_3 is about four times greater for the Spanish guitar's nylon strings than for the acoustic guitar's steel strings, which is reminiscent of the statement found in [FR91, p. 51], and would explain why acoustic guitar tones sound brighter.

[2] These are professional quality recordings, obtained from Yellow Tools' sampler *Independence* (instruments: Acoustic Guitar Spanish; Ovation Piezo Guitar). http://www.yellowtools.com/ (latest access: October 20th, 2011)

[3] To obtain the decay rate series, the partials of the tones were tracked and their decay times were estimated with the usual technique which consists of taking the logarithm of the magnitude envelope, fitting therein a first-order polynomial in a least-square sense, and deriving the decay rate from the polynomial coefficients thus obtained.

However, the air-damping coefficient is greater for the acoustic guitar than for the Spanish guitar, while acoustic guitar treble E strings seem to have lesser radii and thus offer less surface for air resistance to occur. This feature is thus counter-intuitive, and no valid interpretation for this phenomenon can be found here.

On the other hand, it can be seen that the inharmonicity due to the introduction of the γ_k term in (1.26) is largely negligible, with a deviation of the order of the millionth of semitones in both cases, while informal tests run by the author on a musically educated audience seemed to indicate that pitch changes of less than a 40$^{\text{th}}$ of a semitone in a 1kHz pure tone were completely indiscernible. Consistently, this type of inharmonicity is rarely taken into account in the harmonic series models of string tones, only the inharmonicity due to string stiffness [FBS62, RLV07, HTL10].

1.2.4 Stiffness

The relative stiffness of strings is not responsible for a decay, but for a modification of the frequency series ω_k which, for metal strings, is known not to be negligible [FBS62]. Stiffness brings an additional contribution to the acceleration of the string that is proportional to the fourth space derivative:

$$\frac{\partial^2 s}{\partial t^2} = c^2 \frac{\partial^2 s}{\partial u^2} - 2b_1 \frac{\partial s}{\partial t} + \frac{2L^2}{\pi^2} b_3 \frac{\partial^3 s}{\partial t \partial u^2} - \frac{\pi}{4} \frac{r^4 E}{\mu} \frac{\partial^4 s}{\partial u^4}, \tag{1.28}$$

where r is the radius of the circle that describes the cross-section of the string, and E, the Young modulus of the string's material (a measure of stiffness, in newtons per unit area). The complex physical analysis leading

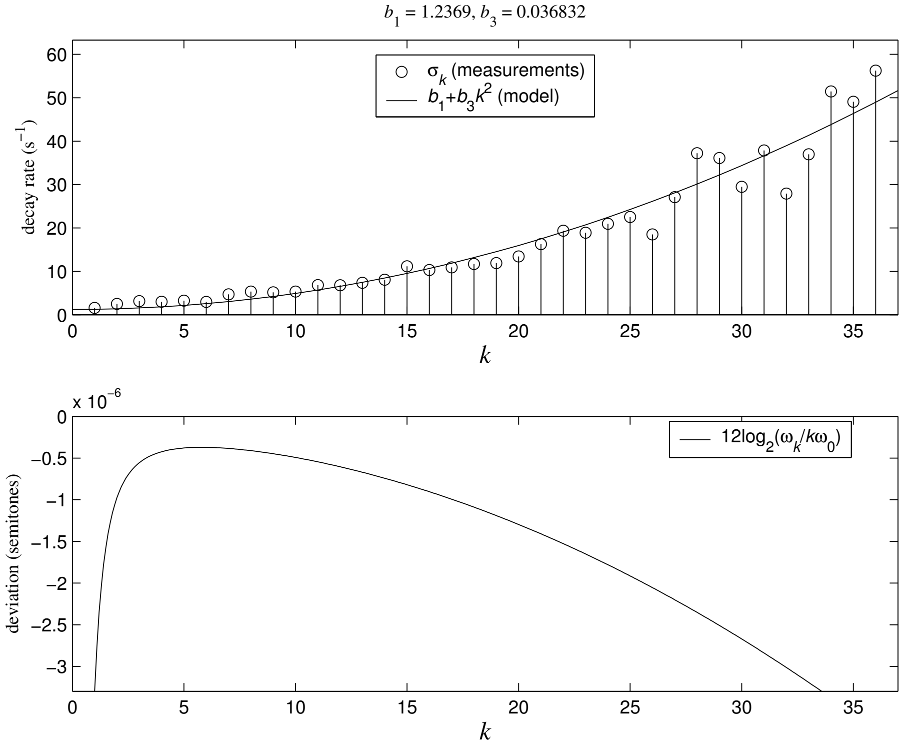

Figure 1.5: Harmonic-index-dependent decay rate (upper plot) and frequency deviation (lower plot) in semitones due to damping on a Spanish guitar open E4

to the obtention of this additional term can be found in [FR91] and [Rai00][4].

Based on a model for the k^{th} vibrational mode identical to (1.25) (Section 1.2.2), we substitute its derivatives into (1.31) and solve for ω_k, yielding

$$\omega_k = \sqrt{k^2 \omega_0^2 \left(1 + \beta k^2\right) - \gamma_k^2}, \tag{1.29}$$

[4] Considering that our object of study is strings exclusively, whose cross-section is well approximated with a circle, we replaced the radius of gyration κ found in [FR91] and [Rai00] with $r/2$.

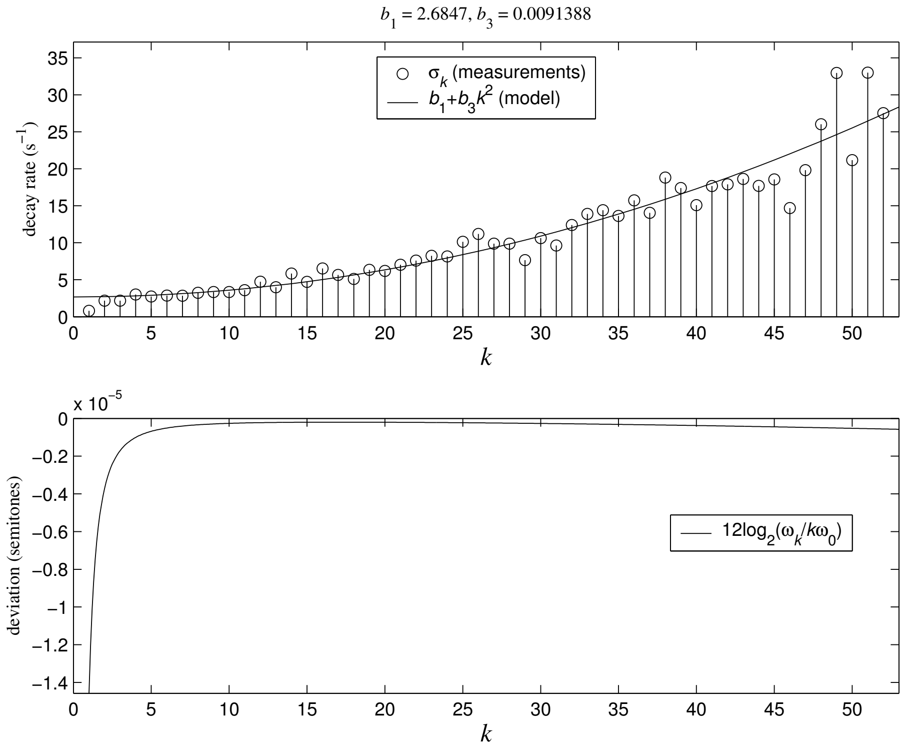

Figure 1.6: Harmonic-index-dependent decay rate (upper plot) and frequency deviation (lower plot) in semitones due to damping on an acoustic guitar open E4

where

$$\beta = \frac{\pi^2}{4}\frac{r^4 E}{T L^2} \tag{1.30}$$

and is known as the *inharmonicity coefficient* of the string. This coefficient can significantly affect the harmonic series ω_k, and for this reason pervades the literature [FBS62, Leh08, HTL10, FR91]. We might therefore like to see it appear in the acceleration equation,

$$\frac{\partial^2 s}{\partial t^2} = c^2\frac{\partial^2 s}{\partial u^2} - 2b_1\frac{\partial s}{\partial t} + \frac{2L^2}{\pi^2}b_3\frac{\partial^3 s}{\partial t \partial u^2} - c^2L^2\frac{\beta}{\pi^2}\frac{\partial^4 s}{\partial u^4}. \tag{1.31}$$

The frequency series of (1.29) satisfies our physical model, with

$$s(u,t) = \sum_{k=1}^{\infty} A_k \sin \frac{k\pi u}{L} \cos \left(\sqrt{k^2\omega_0^2 \left(1+\beta k^2\right) - \gamma_k^2} t + \phi_k \right) e^{-\left(b_1 + b_3 k^2\right)t} \tag{1.32}$$

for solution. However, the decay rate γ_k term is, as seen in the section 1.2.3, of negligible impact on the series. Hence the referential series that will be made use of in the rest of this thesis is a simpler model,

$$\omega_k = k\omega_0\sqrt{1+\beta k^2}, \quad k = 1, 2, 3, ... \tag{1.33}$$

When inharmonicity is negligible and $\beta \approx 0$, the series reduces to a harmonic model $\omega_k = k\omega_0$, in which case the frequency of the first harmonic equals the fundamental frequency, i.e. $\omega_1 = \omega_0$ – this is mathematically correct because our series only applies to values of k greater than 0, and ω_0 is not the "zeroth harmonic".

The frequency deviation in semitones of the frequency values given by (1.33) from a purely harmonic series $\omega_k = k\omega_0$ is measured in Figure 1.7, for the Spanish and acoustic guitar treble E tones whose partials' decay rates were measured in Figures 1.5 and 1.6, as well as for an open bass E of the same acoustic guitar.

An interesting question with regard to stiffness inharmonicity is whether it is audible or not. This depends on the degree of inharmonicity itself, of course, but also on the decay time of the partials affected by this inharmonicity; some may deviate by an interval that is largely audible, but may not last long enough to make an impression of "height" on the listener. Also, it is

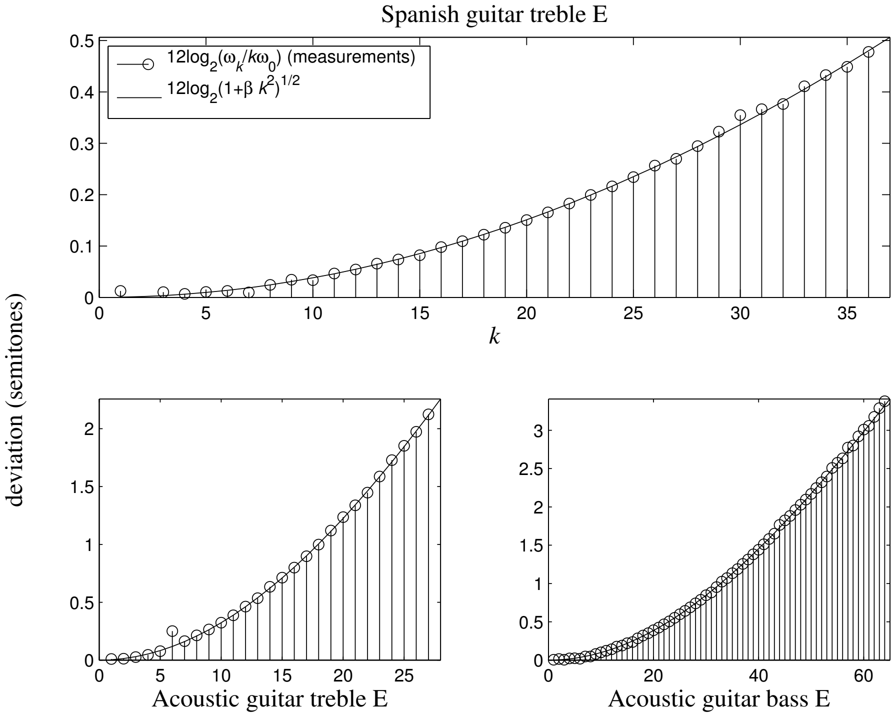

Figure 1.7: Frequency deviation in semitones due to stiffness on Spanish and acoustic guitar tones. The deviation of the partials, clearly measurable and visible, is well approximated with the inharmonic-series expression (1.33) (solid curve). Note : the musical interval in semitones between two frequencies f_2 and f_1 is calculated as $12\log_2 f_1/f_2$, hence the $12\log_2$ terms in the legend.

stated in [Moo04] that the sense of pitch was lost on subjects for pure tones of frequency greater than 5kHz, although differences in frequency were still recognised. In the case of the piano, the instrument showing the greatest degree of string-stiffness-related inharmonicity, it is well known that inharmonicity is a constituent part of the timbre. In [FBS62], the 16th partial of an upright piano's A0 (the lowest note on the keyboard, of fundamental frequency 27.5Hz) deviates by one semitone, which would bring it to a frequency of 466Hz instead of 440Hz. In general, the inharmonicity is reported to be audible and to give a characteristic "warmth" to the sound.

In the case of such instrument as the Spanish guitar, where the strings are made of nylon or plastic, either plain (three treble strings) or wound with bronze or copper wire (three bass strings), the inharmonicity seems to be inaudible. A synthetic model using digital waveguides, presented in [LEVK01], was implemented according to a harmonic model, and yields very satisfying results – examples can be accessed on the Internet following http://www.acoustics.hut.fi/demos/dafx2000-synth/ (last access: February 12th, 2011).

No perceptual tests are known of the author regarding the audibility of the inharmonicity of acoustic guitar tones. However, the measurements made and stored in Table 1.1 should help clarify this question. In this table each line corresponds to a selected harmonic, for which is specified the instrument and note it was measured from, its harmonic number, its frequency (in hertz), its deviation (in semitones), and its decay time (i.e. the time interval over which the partial loses 60 decibels). Those measurements include the 10th and 35th partials of E4 and E2 tones, respectively, in Spanish guitar, acoustic guitar and Steinway grand piano tones[5]. Those specific partials were chosen because, across all instruments: they showed sufficient sustain to be detected and measured; at such indexes inharmonicity has a potential effect; and the frequency of those all fall within or near the 3-4kHz region, which equal-loudness contours show to be the region where the human ear is most sensitive

[5] The piano tones were downloaded from http://theremin.music.uiowa.edu/MIS.piano.html (last access: Frbruary 13th, 2011).

to [FM33].

Instr.	Note	Har. No.	Freq. (Hz)	Dev. (semitones)	60dB Dec. Time (s)
Sp. guitar	E4	10	3,290	0.03	1.29
	E2	35	2,906	0.20	0.62
Ac. guitar	E4	10	3,320	0.09	2.12
	E2	35	3,104	1.13	1.30
Grd Piano	E4	10	3,384	0.4	4.91
	E2	35	3,080	1.12	5.71

Table 1.1: Measurements regarding the audibility of inharmonicity in three instruments

Consistently with our statements above, inharmonicity in the Spanish guitar can be assumed to be inaudible (in spite of the 0.2 semitone deviation seen in the 35th partial of the open E2), while the opposite is true for the grand piano. It should be noticed that, in the case of the 35th harmonic in the acoustic guitar's E2, the deviation caused by inharmonicity is equivalent to that of the grand piano in the corresponding tone and harmonic – a large 1.12 semitones. The only difference between them is in the decay time, which is much greater for the piano's partial (5.71 seconds) than for the acoustic guitar's (1.30 seconds). However, the decay time of the latter seems long enough for the frequency of this partial to be well perceived.

Regarding the treble E, in the acoustic guitar the deviation of the 10th partial is of less than a tenth of a semitone. For such a small deviation, we suspect that the inharmonicity will have greater perceptual bearing in contributing to the fluctuations of the wave shape (perfectly harmonic sounds show a static waveform, which comes with an auditory sensation of "inertia") than in making the series sound sharper. Although more tests should be run, it is believed here that a physical model for the higher strings of an acoustic guitar might achieve convincing results with a simple harmonic model such as that of [LEVK01]. However, the inharmonicity of the lower strings was shown here to be comparable to that of the piano on equivalent tones, and hence, is believed to be audible. Where synthesis is concerned, these observations raise the question of a mixed model, including the inharmonicity in the lower strings for realism, and using a perfectly harmonic (or near perfect) model for the upper strings, for computational efficiency.

The subject of this thesis – the sonic extraction of a string from its instrumental environment – offers a means of making experiments regarding the question of inharmonicity audibility. The PDE (1.31), on page 37, shows that the IC is a coefficient of the stiffness-related term, and that it can be modified independently of the rest. In other words, no other aspect of the wave is changed by a modification of β except the frequency of each mode. Upon close examination, we observe that the one physical parameter upon which the IC alone depends is the string's elasticity modulus E, while the tension T, the length L and the radius r – determining the string's mass density, and hence the wave propagation speed – are shared with the fundamental frequency parameter. This is to say that modification of the IC alone is equivalent to modifying the elasticity of the string alone, too.

The string extraction process proposed in this thesis includes the estimate of the IC – in fact, it relies on it to find the partials to extract. Hence, upon extraction of the string's components, it is a simple process to modify the IC and shift the partials in frequency accordingly. In accordance with our physical model, lesser or greater radii and elasticity moduli affect no other aspect of the standing wave, only the frequency of the components – even their decay rate is left untouched. On this basis, we argue that an IC-processed output truly corresponds to the tone of a string identical in all aspects but stiffness. The only nuance that should be brought in this argument is that an instrumental body is not included in this model (the extremities are here assumed to be perfectly rigid). On the one hand, a sound wave emanates from a string instrument mainly from the resonating body, the strings themselves offering too little surface to set a sufficient number of molecules of air in motion [FR91, Ste96]. On the other hand, each instrument has its own frequency-dependent radiation spectrum [FR91], which has the effect of amplifying the string vibrations to a different degree depending on their frequency. Hence, to virtually modify a string's stiffness without modifying the resonant characteristics of the instrument, correction of the components' magnitude based on prior evaluation of the instrument's resonating properties might be an option. However, this might be unnecessary if the frequency shift of the partials caused by the modification of the IC is too little to cause partials to move from one resonant region of the body to another. Indeed, the resonances of an instrument's body are generally a smooth function of frequency, similarly to the *formants* observed in the spectrum of human-produced vowels [Moo04, FR91]. Hence, the modification of the IC in the series prior to resynthesis of the string's partials, followed by the re-mixing of the "processed string" with the body, might be a perceptually valid way of simulating the modification of the stiffness of an instrument's string. In [FBS62], such perceptual tests were made for the piano, comparing synthetic to recorded tones as a means of judging the quality of piano tones, and as part of that, of the audibility of inharmonicity. Because such aspects of the original tones as the sound of the piano hammer hitting the string and the dampening pedal could not be synthesised, the sound files were cropped so that they were removed. The extraction, independent processing and re-mixing of the string such as proposed in this thesis offers a sensible alternative to this problem.

1.2.5 Time-varying Fundamental Frequency and Inharmonicity Coefficient

At rest, the string is a straight line between the two support points. A line being the shortest path between two points, the length of the string is then at its shortest. Any displacement in the string, any motion, and it has to depart from the "line state", and thereby, increase in length – provided that the supports are immobile, at least in the x direction, which we consider them to be.

Up until now, we have been neglecting this increase in length. In the derivation of the Wave Equation at its most basic, we have already approximated $\sin s$ with $\tan s$, which works for a small s, small displacement. This constant-length assumption is synonymous of constant pitch, as the string's fundamental frequency depends, among other things, on its length. However, there are times when the pitch can be *heard* not to be constant: aggressive play on the lower strings of an electric guitar, or even acoustic guitar, can produce this effect. At other times, this pitch glide may not be audible, but visible in fine fundamental frequency measurements. In these situations, what happens is, the transverse displacement imparted by the excitation becomes too large for the increase in string length to be neglected.

String elongation also causes an increase in tension, and the latter is also involved in the physical expression of the fundamental frequency,

$$\omega_0 = \frac{\pi c}{L} = \frac{\pi}{L}\sqrt{\frac{T}{\mu}}. \tag{1.34}$$

Hence, it seems logical to investigate the phenomenon of string length and tension increase in order to explain that of pitch glide. Yet, the coefficient of inharmonicity also depends directly on these physical parameters, as shown in its physical expression, that is, equation (1.30) on page 37. Conclusions similar to those drawn for the fundamental frequency will hence be drawn for the inharmonicity coefficient as well.

In the literature, the phenomenon of pitch glide is referred to as that of *tension modulation*, the term that we may use henceforth. Our task now is to derive an expression for the length of the string in terms of our physical model,

$$s(u,t) = \sum_{k=1}^{\infty} A_k \cos(\omega_k t + \phi_k) e^{\gamma_k t} \sin\frac{k\pi u}{L}. \tag{1.35}$$

The coming derivation is based on the findings of Legge and Fletcher in [LF84], detailed more recently by Bank in [Ban09].

Derivation of string length and tension

The length of the string can be obtained by use of the Pythagoras theorem within calculus,

$$\begin{aligned} L'(t) &= \int_0^L \sqrt{1 + \left(\frac{\partial s}{\partial u}\right)^2} du \\ &\approx L + \frac{1}{2}\int_0^L \left(\frac{\partial s}{\partial u}\right)^2 du, \end{aligned} \tag{1.36}$$

(1.36) being the integration of a Taylor series 4th-order approximation about zero of the square root term.

By substitution of (1.35) into (1.36) and use of the trigonometric identity $\cos^2(\omega_k t + \phi_k) = (1 + \cos(2\omega_k t + 2\phi_k))/2$, the time-dependent length L' of the string can be expressed as

$$L'(t) = L + \frac{\pi^2}{8L}\sum_{k=1}^{\infty} k^2 A_k^2 (1 + \cos(2\omega_k t + 2\phi_k)) e^{2\gamma_k t}. \tag{1.37}$$

This can be split in two expressions: one with the sum of cosine terms that are twice the frequency of the corresponding modes of vibration, and a quasi-static term [Ban09],

$$L_{\text{qs}}(t) = L + \frac{\pi^2}{8L}\sum_{k=1}^{\infty} k^2 A_k^2 e^{2\gamma_k t}. \tag{1.38}$$

(1.38) is the sum of the length of the string at rest, L, and the summation of decaying exponentials. Each k^{th} exponential's decay rate is found to be twice that of the corresponding k^{th} mode of vibration of the string.

In turn, this expression depends on its length according to

$$T'(t) = T + \frac{\pi r^2 E}{L}(L'(t) - L), \tag{1.39}$$

By substitution of (1.37) into (1.39), we get

$$T_{\text{qs}}(t) = T + \frac{\pi^3 r^2 E}{8L^2} \sum_{k=1}^{\infty} k^2 A_k^2 e^{2\gamma_k t}, \tag{1.40}$$

provided that, again, only the quasi-static term is preserved. Similarly to what is stated in (1.39), the quasi-static and time-varying tension and length are simply related:

$$T_{\text{qs}}(t) = T + \frac{\pi r^2 E}{L}\left(L_{\text{qs}}(t) - L\right).$$

Fundamental Frequency and Inharmonicity Coefficient models

All this considered, we can now gather the time-varying expression of the fundamental frequency

$$\omega_0(t) = \frac{\pi}{\sqrt{\mu}} \frac{\sqrt{T + \alpha\Sigma(t)}}{L + \Sigma(t)} \tag{1.41}$$

and inharmonicity coefficient

$$\beta(t) = \frac{\pi^2 r^4 E}{4\left(T + \alpha\Sigma(t)\right)\left(L + \Sigma(t)\right)^2}, \tag{1.42}$$

where

$$\Sigma(t) = \frac{\pi^2}{8L} \sum_{k=1}^{\infty} k^2 A_k^2 e^{2\gamma_k t} \tag{1.43}$$

and

$$\alpha = \frac{\pi r^2 E}{L}. \tag{1.44}$$

Notice that the inverse of the time-varying IC is a cubic polynomial in $\Sigma(t)$,

$$\beta^{-1}(t) = a_3\Sigma^3(t) + a_2\Sigma^2(t) + a_1\Sigma(t) + a_0, \tag{1.45}$$

where

$$a_0 = \frac{4TL^2}{\pi^2 r^4 E} = \frac{1}{\beta},$$

β being the inharmonicity coefficient constant already seen in Equation (1.30), which, consistently with (1.45), had been derived in a string model where the length of the string was approximated as constant.

As will be shown in coming figures, the truncation of the polynomial (1.45) to its zeroth- and first-order terms gives a satisfying approximation. The first-order term a_1 can be found as

$$a_1 = \frac{1}{\beta}\left(2L + \frac{\alpha}{T}\right),$$

and so the inverse of the truncated polynomial can be written as

$$\beta(t) \approx \frac{\beta}{1 + (2L + \alpha/T)\,\Sigma(t)}. \tag{1.46}$$

Knowing that $\Sigma(t)$ is a sum of exponentials which all decay towards zero,

$\beta(t)$ is to be read as a function whose value increases in time, at a lesser and lesser rate, converging towards the constant β. This time-increase trend is clearly visible in the lower plot of Figure 1.8.

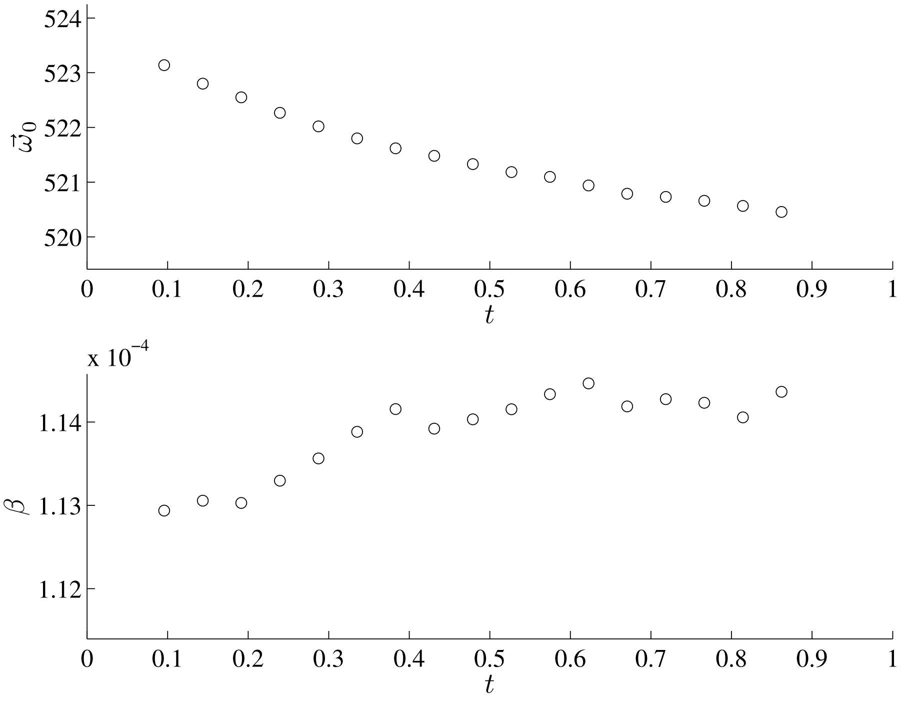

Figure 1.8: Fundamental Frequency & Inharmonicity Coefficient measurements in an Ovation acoustic guitar open bass string.

Where the time-varying FF is concerned, Equation (1.41) is completely non-linear in $\Sigma(t)$. Yet the upper plot of Figure (1.8) is clearly reminiscent of an exponential-plus constant function, which leads us to think that a first-order Taylor series in $\Sigma(t)$ might yield a good approximation [Rab12]. We re-arrange (1.41) and express it in terms of a Taylor series,

$$\begin{aligned}\frac{\sqrt{\mu}}{\pi}\omega_0(t) &= \frac{\sqrt{T+\alpha\Sigma(t)}}{L+\Sigma(t)} \\ &= f(\Sigma(t)) \\ &\approx f'(0)\Sigma(t)+f(0),\end{aligned}$$

which, after mathematical development, yields the approximation

$$\omega_0(t) \approx \frac{\pi^2 r^2 E - 2\pi T}{2L^2\sqrt{\mu T}}\Sigma(t)+\omega_0. \tag{1.47}$$

The constant term, ω_0, yet again proves to be the fundamental frequency constant derived earlier and presented in Equation (1.5), in the context of the assumption of a constant string length.

If the approximation that $\Sigma(t) \approx \xi e^{\gamma t}$ held, for some appropriate constants ξ and γ, we could express equations (1.46) as an exponential-plus-constant function and (1.47) as the *inverse* of an exponential-plus-constant function, respectively. The number of unknowns would then reduce to three for either of these functions, and some non-linear fitting approach could be taken to model these time-varying trajectories. Let us therefore re-formulate (1.47) as

$$\omega_0(t) = \omega_\Delta e^{\gamma_\omega t}+\omega_\infty, \tag{1.48}$$

and (1.46), as

$$\beta(t) = \frac{\beta_\infty\left(\beta_\infty+\beta_\Delta\right)}{e^{\gamma_\beta t}\left(\beta_\infty+\beta_\Delta\right)-\beta_\infty}, \tag{1.49}$$

the idea being that : ω_∞ and β_∞ are the values towards which $\omega_0(t)$ and $\beta(t)$

tend as t tends towards infinity ; ω_Δ and β_Δ are the difference between the respective values of these functions at time $t = 0$ and $t \to \infty$; and γ_ω and γ_β are the decay rate of their respective exponential component. (These are the symbols that were used in the paper where the author first introduced this theory, [HTL10].)

We are now in position to try and fit these parameters in actual measurements, which we are hoping will validate our exponential-plus constant modeling of the time-varying fundamental frequency and inharmonicity coefficient.

1.2.6 Exponential-plus-constant curve fitting

Fitting an exponential-plus-constant curve into data points, especially when the points are so few, is a delicate task in itself. This problem is met in other areas of science as well, and it would seem that most generally, Non-Linear Least Squares are used [NKS04, HWS87, Hag03]. The popularity of this method can be explained by the fact that it is generic to non-linear fitting problems, which spares the need to learn new statistical skills for each new type of problem found. However, one downside to this method is that it requires initial estimates that are sometimes difficult to guess. Also, due to its least-squares nature, it is sensitive to data outliers. Last, but not least, its convergence time is variable, and in some cases may not converge at all – the IC data shown in Figure 1.8, where the points are few and the noise is strong, could well become one of these cases.

Another idea that might cross one's mind is to use differentiation and logarithmic transformation. Indeed, it is straightforward that the logarithm of the derivative of our exponential-plus-constant $\omega_0(t)$ is a function linear in t – formally,

$$\log\left(\frac{d\omega_0}{dt}\right) = \log(\gamma_\omega \omega_\Delta) + \gamma_\omega t.$$

Thereafter, linear least squares can be used to isolate and work out γ_ω. $\omega_0(t)$ then becomes a function linear in $e^{\gamma_\omega t}$, and linear least squares can be used once more to find ω_Δ and ω_0. The issue with this approach is that it is sensitive to noise in places where the derivative $d\omega_0/dt$ is small in relation to the measurement error. Here follows a short formulation of the phenomenon.

To begin with, let us consider a length-U time vector of evenly-spaced sampling instants

$$\mathbf{t} = [t_0, t_0 + T_s, t_0 + 2T_s, ..., t_0 + (U-1)T_s]^T \, ,$$

where t_0 is the time value of the first sample, and T_s , the sampling time interval. We now express the measurements of the fundamental frequency in vector form, as the sum of an uncorrupted exponential-plus-constant vector and a measurement error vector

$$\boldsymbol{\varepsilon} = \left[\varepsilon^0, \varepsilon^1, ..., \varepsilon^{U-1}\right]^T \, ,$$

.i.e.

$$\boldsymbol{\omega}_0 = \omega_0(\mathbf{t}) + \boldsymbol{\varepsilon}.$$

Now we look at the values issued from the Finite Difference of the $\boldsymbol{\omega}_0$ and $\boldsymbol{\varepsilon}$,

$$\dot{\omega}_0^n = \omega_0 \left(t_0 + (n+1)T_s \right) - \omega_0(t_0 + nTs)$$

and

$$\dot{\varepsilon}^n = \varepsilon^{n+1} - \varepsilon^n,$$

for $n = 0, 1, ..., U-2$. It can be derived that a measurement error difference $\dot{\varepsilon}^n$ in the domain of origin becomes, after logarithmic transformation,

$$\dot{\varepsilon}^n_{\log} = \log \left(1 - \frac{\dot{\varepsilon}^n}{\dot{\omega}_0^n} \right).$$

The presence of the finite difference of $\boldsymbol{\omega}_0$ in the denominator is problematic, fundamental frequency measurements are common where little or no decay trend can be seen, hence augmenting dramatically the error after differentiation.

The author presented in [Hod11] a fast and robust method for the fitting of exponential-plus-constant curves to measurements. The method was inspired by the realisation that the exponent's coefficient could easily be found through the Fourier series of the function, when considered over a finite interval of its domain. In practice, the DFT was substituted for the Fourier series to evaluate the coefficient, but before that, so as to reduce the bias introduced by aliasing, the data was modulated by one period of a sinusoid, in the manner of standard windowing. Again, the other two linear coefficients could thereafter be estimated through standard linear least squares.

A Matlab function was encoded, using the syntax

$$(\gamma_\omega, \omega_\Delta, \omega_\infty) = \texttt{fepcf}\,(\mathbf{t}, \boldsymbol{\omega}_0)\,,$$

`fepcf` standing for Fourier-based Exponential Plus Constant Fit. The function takes in two equal-length vectors, $\mathbf{t}$ and $\boldsymbol{\omega}_0$, and outputs the three coefficients looked for, γ_ω, ω_Δ and ω_∞.

Figure 1.9 demonstrates at once the validity of the exponential-plus-constant based models for fundamental frequency and inharmonicity coefficient. In the lower panel, it also shows the overall robustness of the Fourier-based approach. Robustness is here greatly desirable, as inharmonicity coefficient estimates are often very sensitive to noise. Nevertheless, the trend can be seen outlined by our fitted model.

Being in possession of a model and fitting technique for time-varying fundamental frequencies is valuable, for analysis as well as synthesis purposes. However, in the case of the inharmonicity coefficient, this is much less obvious. At reasonable harmonic indices, change in the IC has negligible impact in comparison with the glide in fundamental frequency. However, there may come a harmonic index where the upward trend in inharmonicity tempers the fundamental frequency glide. Figure 1.10 shows the spectrogram of the data already used to extract the information presented in figures 1.8 and 1.9. The partial tracks

$$\omega_k(t) = k\omega_0(t)\sqrt{1 + \beta(t)k^2} \tag{1.50}$$

are represented with the dashed lines. To emphasize the effect of inhar-

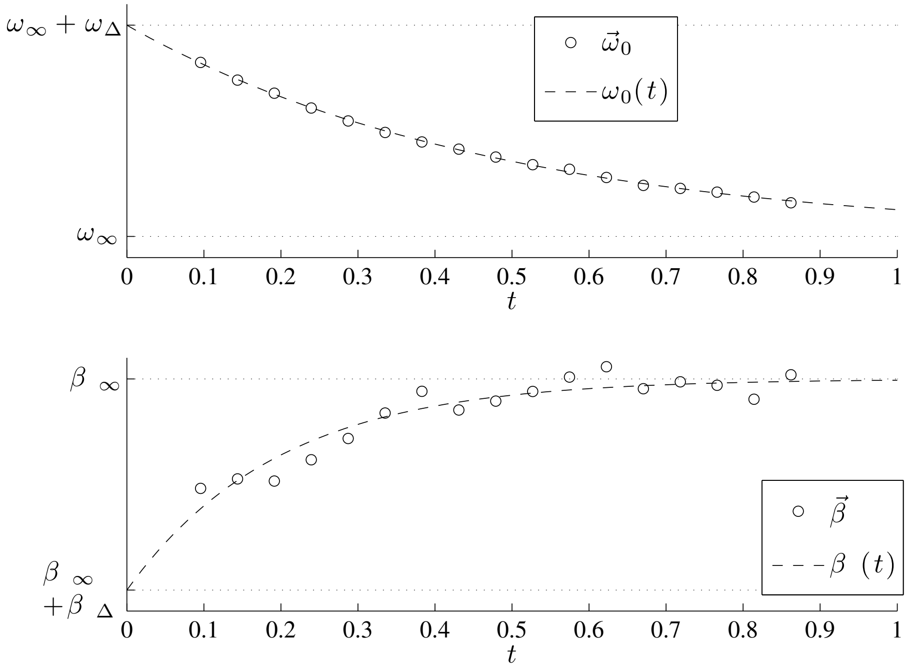

Figure 1.9: FF and IC models (dashed lines) fitted in previously presented measurements (circles), using the Fourier-based method.

monicity time-variance, tracks were also drawn in dash-dotted lines using $\beta(0)$ as constant inharmonicity coefficient. The latter are seen to glide down more rapidly than they should, soon departing from the actual partial tracks, clearly visible in lighter shade. By contrast, the model based on (1.50) faithfully follows the partials, and seem to reveal their path even after their trend is stabilised to a constant.

There is a potential for synthesis and processing in the model derived in this section. Such reliable extrapolation of the partial frequencies avails of the possibility of extending the lifetime of the partials, thereby creating an effect of "sustain increase" – it should be realised that this differs from

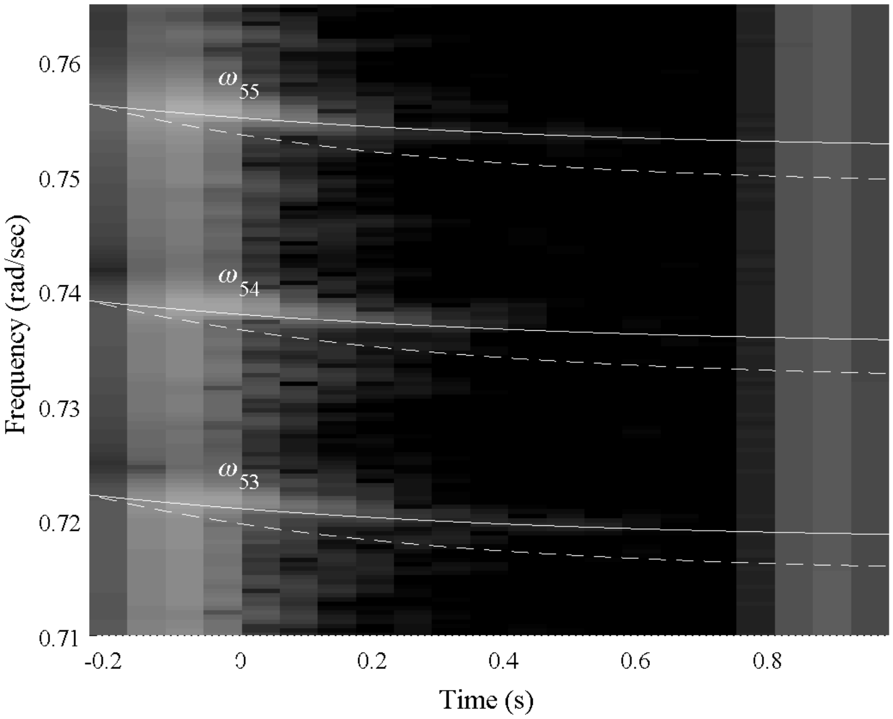

Figure 1.10: Model of partial tracks based on time-varying (solid lines) and fixed (dashed lines) Inharmonicity Coefficient, on top of the Ovation E2 spectrogram. The focus is here on partials 53 to 55.

standard time stretching of the partials tracks, where the decay time of the fundamental frequency and inharmonicity coefficient glides are also stretched. For surrealistic effects, it might also be considered to exploit the fact that the model defines the partial tracks for all values of time, even negative ones, virtually before the excitation actually takes place. However, it must be realised that this also implies a model for the magnitude envelopes of the partials. Where frequency is concerned, however, model (1.50) is in this thesis considered complete.

Before closing, it is important to note that, because of the approximations necessary to obtain the reasonably simple models for fundamental frequency and inharmonicity coefficients, we have departed at the start of section 1.2.5 from a rigorous physical analysis scheme. Neither $\omega_0(t)$ nor $\beta(t)$ as expressed in (1.48) and (1.49) can be substituted into the string displacement expression issued from this chapter's analysis and synthesised in (1.35) and satisfy the Partial Differential Equation (1.31) at the same time. However, these approximations have been proven to be worthwhile by the fitting of the resulting model to actual data, and furthermore, no refinement in the description of the string's vibrations is to be developed on top of these approximations before the chapter on physical analysis is over. These approximations shall thereby have bearings on no other but themselves.

1.2.7 Longitudinal vibrations, phantom partials

For now we have only been considering the transverse vibrations of the string, those perpendicular to the direction of the wave propagation. There exists two more categories of vibrations: torsional, and longitudinal vibrations.

Torsional vibrations are negligible contributors to the sound output of plucked and hit string instruments. Even in bowed instruments, they are not involved directly in the radiation of the sound, only in the interactions between bow and string [TR03, MSW83].

Longitudinal vibrations, on the other hand, are commonly found in string spectra. It has been known for a long time that they contribute to the timbre of piano tones [Kno44], and they were also found in the output of other instruments, such as the acoustic guitar [HAC99], or, more recently, the kantele, a Finnish 5-string instrument [EKHV02]. Figure 1.11 illustrates the reality of longitudinal partials, with examples taken from a Spanish guitar (top) and acoustic guitar (bottom). It can be seen here that transverse partials are still largely predominant, even in these selected regions of the spectra where a "longitudinal series" arises. Longitudinal partials may therefore get *masked* [Moo04] by the transverse partials towering in their vicinity, and become inaudible. While the question of their audibility is fully relevant to synthesis, in the context of this thesis it is much less the case. Our aim is the cancelation of all audible partials issued from the string. Given the levels measured

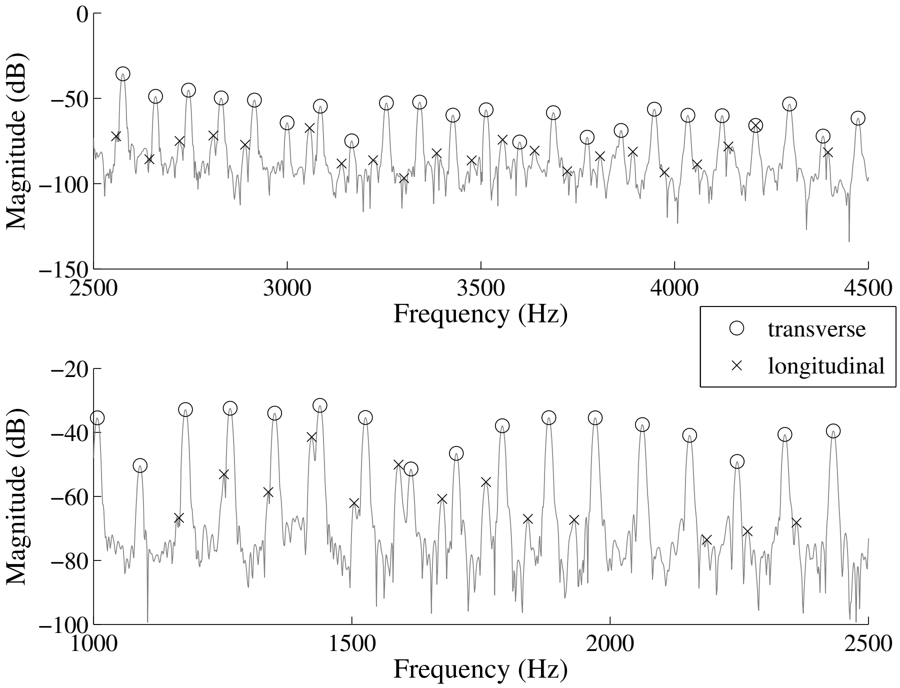

Figure 1.11: Presence of longitudinal partials in Spanish (top) and acoustic (bottom) guitar spectra. Both are open bass E tones.

in Figure 1.11, with longitudinal partials at times reaching -40dB$_{\mathrm{FS}}$ [6] for a normalised input waveform, it is evident that, once the transverse partials canceled, the remaining longitudinal partials must be, on their own, clearly audible. This statement is supported by informal perceptual experiments.

Free and driven longitudinal vibrations

In this thesis we distinguish two types of longitudinal vibrations: free vibrations, and driven vibrations. Free vibrations are found at the longitudinal resonant frequencies of the string, a harmonic series whose fundamental can be derived by physical analysis to

$$\omega_0^{\mathrm{L}} = \frac{\pi r}{L}\sqrt{\frac{E\pi}{\mu}}, \tag{1.51}$$

in radians per second [FR91, Rai00]. This frequency is independent on tension, and hence, on the fundamental frequency of the transverse series. For instance, the first partial of a longitudinal series was found in a 500–3,000Hz interval in piano tones whose (transverse) fundamental frequency lay in a 30–100Hz interval [Kno44].

Having a higher fundamental frequency lessens the density of normal longitudinal partials. Moreover, when they are detectable, they only appear in the "precursive sound", given a very strong decay rate, in the order of 100dB per second [PL87]. On this basis, and to keep complexity down to a reasonable level, normal longitudinal partials will be ignored in the string

[6] dB$_{\mathrm{FS}}$, or *Full-Scale decibels*, are used in the context of digital audio, where the reference amplitude level is 1, or 0dB$_{\mathrm{FS}}$

extraction process. Their perceptual impact is not neglected, but, where present, we shall consider them as part of the excitation.

The longitudinal vibrations measured in Figure 1.11 are in fact of the second type, *driven* longitudinal vibrations. These result from the modulation in length and tension of the string as it vibrates; as such, they are directly related to the phenomenon of pitch glide detailed in the previous section. Because of their audibility, they cannot be excluded from the list of partials to extract in the string extraction process. As a matter of fact, their inclusion in this thesis demands relatively little additional content, as their frequency and decay time parameters are aligned with those of the transverse series, as we are going to see of now.

Derivation of longitudinal vibration parameters

Driven longitudinal vibrations are a nonlinear function of transverse amplitude [GK96]. A PDE for the longitudinal motion based on the standard wave equation and augmented with a term for the influence of the transverse vibrations is developed in [MI86], and simplified in [BS03] as

$$\mu\frac{\partial^2\xi}{\partial t^2} = E\pi r^2\frac{\partial^2\xi}{\partial u^2} + \frac{1}{2}E\pi r^2\frac{\partial\left(\frac{\partial s}{\partial u}\right)^2}{\partial u}. \tag{1.52}$$

The nonlinearity stems from the squaring of the transverse vibrations. The displacement s is the sum of an infinite number of terms, as per our model, expressed in equation (1.35), section 1.2.5. However, as $\partial s/\partial u$ is raised to the power of two, it is sufficient to consider the nonlinear effect on the sum of two components only [BS03], of indices, say, k and l:

$$s_{k,l}(u,t) = s_k(u)e^{\gamma_k t}\cos\left(\omega_k t + \phi_k\right) + s_l(u)e^{\gamma_l t}\cos\left(\omega_l t + \phi_k\right). \tag{1.53}$$

Our purpose allows that we keep the focus on frequency, decay time and initial phase only. Hence, the spatial distribution does not need to be made explicit, and may for each components k and l be concisely expressed as some undefined function of u, $s_k(u)$ and $s_l(u)$.

Differentiation, squaring and differentiation again of (1.53) yields

$$\begin{aligned}\frac{\partial\left(\frac{\partial s_{k,l}}{\partial u}\right)^2}{\partial u} =& \frac{1}{2}e^{2\gamma_k t}\left(1+\cos\left(2\omega_k t + 2\phi_k\right)\right)\frac{\partial\left(\frac{\partial s_k}{\partial u}\right)^2}{\partial u}\\ &+\frac{1}{2}e^{2\gamma_l t}\left(1+\cos\left(2\omega_l t + 2\phi_l\right)\right)\frac{\partial\left(\frac{\partial s_l}{\partial u}\right)^2}{\partial u}\\ &+e^{(\gamma_k+\gamma_l)t}\left(\cos\left[\left(\omega_k+\omega_l\right)t+\phi_k+\phi_l\right]+\cos\left[\left(\omega_k-\omega_l\right)t+\phi_k-\phi_l\right]\right)\\ &\times\frac{\partial\left(\frac{\partial s_k}{\partial u}\frac{\partial s_l}{\partial u}\right)}{\partial u}.\end{aligned} \tag{1.54}$$

Transverse partials are shown in (1.54) to generate two types of longitudinal vibrations, *even*, $e^{2\gamma_k}\cos 2\omega_k t$, and *odd*, $e^{(\gamma_k+\gamma_l)t}\cos\left[\left(\omega_k+\omega_l\right)t\right]$ [HAC97]. Even partials are issued from a single component, with twice its frequency, decay time and phase. An odd partial depends on the combination of two distinct transverse partials, inheriting the sum of their frequencies, decay times and phases as its own. These longitudinal partials driven by transverse motion probably account for most of the *phantom partials* often alluded to in the literature [HAC97, BS03, Smi11]. We may thereby use "phantom partials" as a shorthand for "driven longitudinal vibrations" henceforth.

According to (1.52), K^2 longitudinal partials should be generated due to the presence of K transverse partials. However, Conklin finds in [HAC97] that "odd" partials come predominantly from adjacent transverse partials, a phenomenon that Bank *et al.* attempt an explanation for in [BS03], based on the consideration of space orthogonality of the transverse and corresponding longitudinal partials. Be that as it may, this fact allows for a simple arrangement of the phantom partials into a pseudo-harmonic series,

$$\omega_k^{\mathrm{L}} = \begin{cases} \omega_{(k-1)/2} + \omega_{(k+1)/2} & k \text{ odd,} \\ 2\omega_{k/2} & k \text{ even.} \end{cases} \tag{1.55}$$

It is important to note here that, for perfectly harmonic transverse series, i.e. when $\omega_k = k\omega_0$, the transverse and longitudinal series coincide. Hence, phantom partials driven from transverse vibrations cannot expect to be found in inharmonicity-free string tones. Only when $\omega_k = k\omega_0\sqrt{1+\beta k^2}$ (1.33) may this be the case.

The expression of the longitudinal series can be simplified one step further. On the one hand, substitution of (1.33) into (1.55) simplifies, for the even case, to $\omega_k^{\mathrm{L}} = k\omega_0\sqrt{1+{}^1\!/\!_4\beta k^2}$. On the other hand, it can be verified that, within reasonable inharmonicity coefficient values and partial indices, the odd frequencies differ negligibly from the even frequencies, i.e.

$$k\omega_0\sqrt{1+\frac{1}{4}\beta k^2} \approx \frac{k-1}{2}\omega_0\sqrt{1+\beta\left(\frac{k-1}{2}\right)^2} + \frac{k+1}{2}\omega_0\sqrt{1+\beta\left(\frac{k+1}{2}\right)^2}.$$

Hence we can finalise the phantom frequency series to be

$$\omega_k^{\mathrm{L}} = k\omega_0\sqrt{1 + {}^{1}\!/\!{}_{4}\beta k^2}. \tag{1.56}$$

It is very convenient to have a frequency series for our longitudinal partials that is modeled directly upon the transverse frequency series. For both, the frequencies of the partials are entirely defined by the fundamental frequency ω_0 and inharmonicity coefficient β. Also, the decay rate parameters of the phantom series can be derived from those of the transverse series with equivalent ease. Let us recall that the decay rate γ_k of the k^{th} partial is, according to our model, to be expressed as $\gamma_k = b_1 - b_3k^2$ (equation (1.27)). Similarly to frequency, the decay rate of even phantom partials is twice that of the parent transverse partial, and that of odd partials, the sum of the decay rates of its parents. This yields

$$\gamma_k^{\mathrm{L}} \approx -2b_1 - \frac{1}{2}b_3k^2. \tag{1.57}$$

Equations (1.56) and (1.57) present the longitudinal partials as a series in its own right, a quarter the inharmonicity of the main series, twice its air resistance, and half its internal damping.

Considerations on the longitudinal distribution

Unfortunately, the longitudinal model cannot be completed with a time-zero distribution, as we had in the A_k transverse coefficients, which were specified for plucked and hit strings in section 1.1.3. These could be made explicit in

(1.53), but that would be of little use so long as the longitudinal response of the string is not taken into account. Here, the string in its longitudinal polarisation can be seen as a resonator for the transverse vibrations. At natural longitudinal frequencies of the string only can the transverse-driven longitudinal vibrations be expected to be well amplified. Proximity of the even and odd frequencies seen in (1.54) with longitudinal normal frequencies is, in this regard, pointed in [BS03] as a condition for efficient coupling.

To test this hypothesis, Figure 1.12 highlights the transverse (circles) and longitudinal (crosses) series on the short-time spectrum of a Steinway grand piano E2 (82.4Hz). It was hoped that a trend of longitudinal partial emer-

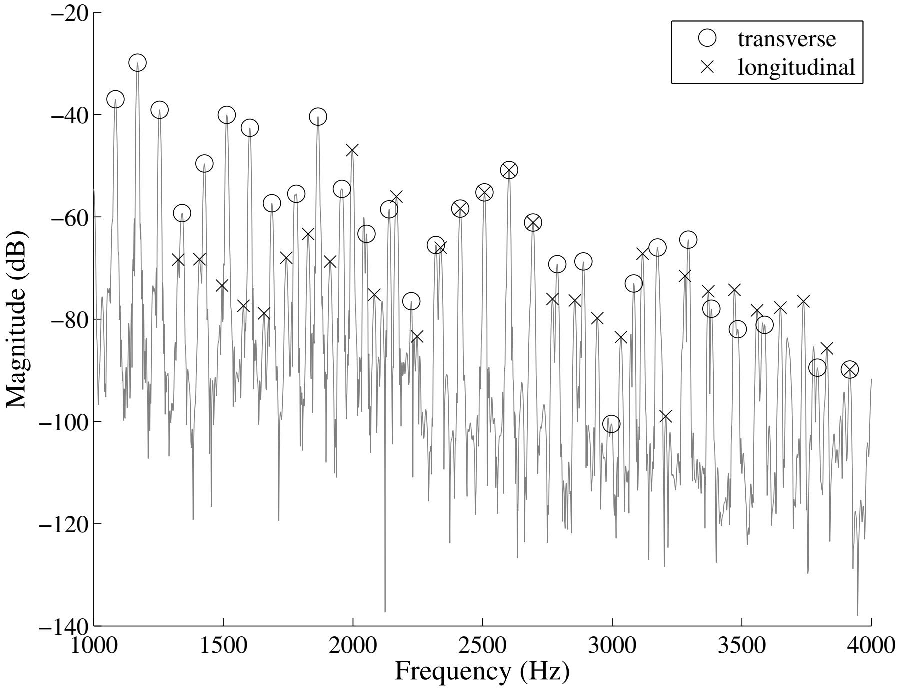

Figure 1.12: Transverse series and transverse-driven longitudinal series in a Steinway E2 piano tone. Transverse-longitudinal conflicting partials are marked with crossed circles.

gence at intervals of the (unknown) longitudinal fundamental frequency could be observed. The reading of the spectrum is in this regard inconclusive, as such trend is invisible. This could be due, in part, to the complex reality of piano tones, far more difficult to fit into templates than other string instruments. More importantly, the phantom partial frequencies regularly coincide with the transverse partial frequencies, in which case it is difficult to decide whether the observed peak is the result of a transverse, a longitudinal, or both partials – in general, preference is given to the transverse choice, due to the overall predominance of the transverse series. On the plot, such conflicting peaks are denoted with crossed circles (except for all partials up to frequency 1,200Hz, for clarity).

Upon appreciation of this complexity, the task of determining a model for the frequency distribution of longitudinal partials is here left aside. It is acknowledged that such a model would be desirable for the synthesis of string tones. Yet, as far as processing is concerned, longitudinal partial magnitudes are obtained from measurement, and it seems sufficient to leave them intact during pitch and inharmonicity coefficient manipulations. This is discarding the frequency-dependent longitudinal responsiveness of the string, but it seems reasonable to think that it compromises little the realism of such a sound effect.

1.3 Recapitulation

In this section, we are going to make the string model developed throughout this thesis more concise and ready for use in subsequent chapters.

First of all, the space variable u can be discarded. During this development it was indispensable to the derivation of the PDEs, and also to the expression of the initial amplitude coefficients of the transverse series. The space variable would also be relevant to synthesis, to construct a model where the plucking position parameter is accessible, but the rest of this thesis is going to focus on the string extraction, which essentially is a processing task. Therefore, the $\sin \frac{k\pi u}{L}$ spatial distribution term, seen as early as section 1.1.2 and throughout the rest of the chapter, will be discarded. Formally, this is simplifying $s(u,t)$ down to $s(t)$.

Our string model $s(t)$ turns out to be the sum of a transverse and a longitudinal polarisation,

$$s(t) = s^{\mathrm{T}}(t) + s^{\mathrm{L}}(t).$$

These sub-models are detailed, level by level, in Table 1.2.

transverse vibrations	longitudinal vibrations	section
$s^{\mathrm{T}}(t) = \sum_k A_k e^{\gamma_k t} \cos\left(\int_0^t \omega_k(t)dt + \phi_k\right)$	$s^{\mathrm{L}}(t) = \sum_k A_k^{\mathrm{L}} e^{\gamma_k^{\mathrm{L}} t} \cos\left(\int_0^t \omega_k^{\mathrm{L}}(t)dt + \phi_k^{\mathrm{L}}\right)$	1
$A_k = \begin{cases} -\frac{2v_{\mathrm{h}}}{\omega_0} \frac{1}{k} \sin \frac{k\pi u_{\mathrm{h}}}{L} & \text{hit} \\ -\frac{A2L^2}{\pi^2} \frac{1}{(u_{\mathrm{p}}-L)u_{\mathrm{p}}} \frac{1}{k^2} \sin \frac{k u_{\mathrm{p}} \pi}{L} & \text{pluck} \end{cases}$	A_k^{L} undefined	1.1.3
$\phi_k = \begin{cases} 0 & \text{pluck} \\ \frac{\pi}{2} & \text{hit} \end{cases}$	$\phi_k^{\mathrm{L}} = \begin{cases} 0 & \text{pluck} \\ \pi & \text{hit} \end{cases}$	1.1.3
$\gamma_k = -b_1 - b_3 k^2$	$\gamma_k^{\mathrm{L}} = -2b_1 - \frac{1}{2} b_3^2$	1.2
$\omega_k(t) = k\omega_0(t)\sqrt{1+\beta(t)k^2}$	$\omega_k^{\mathrm{L}}(t) = k\omega_0(t)\sqrt{1+\frac{1}{4}\beta(t)k^2}$	1.2.5
$\omega_0(t) = \omega_\Delta e^{\gamma_\omega t} + \omega_\infty$ and $\beta(t) = \frac{\beta_\infty(\beta_\infty+\beta_\Delta)}{e^{\gamma_\beta t}(\beta_\infty+\beta_\Delta)-\beta_\infty}$		1.2.6

Table 1.2: String model developed in this chapter.

The end-unknowns, e.g. ω_Δ, γ_β, b_1, or u_h/L, may be evaluated from appropriate fitting techniques for the purpose of physically meaningful sound effects, or for the very purpose of this thesis, string extraction.

Conclusion

Our string model is now in its final form. This conclusion gives the opportunity for us to reflect upon this model.

The model was initiated by the derivation of the Wave Equation. This Partial Differential Equation was obtained by the equation of the forces acting upon infinitesimally small segments of the string, and these forces were deduced from a schematic representation of the string, where the physical parameters of interest – length, tension, mass density – were present. As a result, these very parameters ended up in the basic model. This is also true for the string's stiffness term, whose derivation can be found in the textbooks [FR91, Rai00].

This schematic physical analysis approach is especially attractive for sound synthesis and processing, as the very physicality of the instrument can virtually be manipulated. However, albeit attractive, it also demands very advanced knowledge and skills in acoustics and mechanical physics, and we were unfortunately bound to step down a little as soon as the question of damping arose. There, the appropriate terms were brought in the PDE on the basis of good sense, e.g. the force opposed to the motion of the string due to air resistance being proportional and opposite to its velocity. This resulted in the appearance of anonymous coefficients (such as b_1, to retain the example of air friction), whose physical identity can only be guessed. Yet, it should be pointed out that, for the purpose of this thesis, being able to relate coefficients to physical parameters of the instrument is not necessarily useful, as our string extraction process is blind to the type of string instrument at the origin of the input digital wave. For instance, the constant amplitude term A_k as specified for plucked strings, seen in Table 1.2, depends on both the distance of the plucking point from the string's rest axis and its length. Obviously, initial amplitude estimates in our processing unit will end up with a single coefficient for these two parameters.

The PDE can be seen as some safety guideline, the respect of which is the guarantee of a valid model. It was therefore wise and desirable to follow it faithfully as far as possible into the model. However, on the occasion of the inclusion of the phenomenon of pitch and inharmonicity glide, and of the longitudinal series, more practical approaches had to be taken. Such a breach to our guideline was there tolerated for its practicality, and because it occurs at the end of the development, sitting on top of a physically consistent model. In any case, it permitted the obtention of an Exponential-Plus-Constant (EPC) model for the fundamental frequency and inharmonicity coefficient. Albeit questionable in its physical validity, the EPC model was shown in Figure 1.10 to trace with precision the evolution of entire sets of harmonics in tension-modulated tones.

Chapter 2

Frequency-domain component estimation

Introduction

Chapter 1 was dedicated to the derivation of an analytical model for the vibrations of the string. This was the starting point to the aim of this thesis, namely, the virtual extraction of the string, through the cancelation of all the partials, transverse and phantom.

In this thesis, we propose to cancel these partials in the context of a Phase Vocoder scheme. There, the input is decomposed in short-time segments, or *grains*, and the cancelation is operated at such time scale. The extraction process is common to all grains, and to guarantee that this process is successful at the level of a grain is thereby to guarantee that the string extraction process is successful at the level of the entire input.

All the details on the cancelation of the partials at the time scale of a grain will be given in Chapter 3, but in short, it consists of measuring the relevant parameters of the partials of the string so as to generate synthetic copies, and thereafter to subtract these copies from the grain. To achieve good results at an optimal computational cost, the set of parameters of the string partials that are taken into account in the synthetic model must include neither more, nor less than the parameters that are relevant at such small time scale. Taking into account some unnecessary parameter would make the method more complicated and costly than necessary. Inversely, omitting a parameter that turns out not to be of importance would lead to inefficient cancelation.

In this regard, we will see in Section 2.4 that the parameters of the partials that need, at the time scale of a grain, to be taken into account, are the frequency, magnitude and phase, but also the exponential magnitude envelope. Other methods will be evoked, but it will be explained how the Complex Spectral Phase-Magnitude Evolution (CSPME) method allows the evaluation of frequency and exponential magnitude. The CSPME is an original method, a generalisation of the Complex Spectral Phase Evolution (CSPE) [SG06] to exponential-amplitude signals. It will also be explained how the knowledge of the frequency and magnitude envelope of the partials can be used to, thereafter, obtain the remaining two parameters of constant magnitude and phase.

However, measurement of partial parameters in the frequency domain is inevitably exposed to cross-interference between partials, a consequence of *spectral leakage* [AW03]. Spectral leakage can nevertheless be minimised by the use of some appropriate window. This chapter therefore includes, in Section 2.2, a development on the theory of windowing, which will lead to the choice of the most appropriate window or class of windows, among the many that can be found in the literature [Har78, Nut81]. However, before this development is initiated, a short introduction on the notation used throughout this chapter shall spare on the reader tedious inferences.

2.1 Notation guidelines

The string extraction method introduced in this thesis relies heavily on the frequency-domain properties of windowed sinusoids. The aim of the development that is found in Section 2.2 and following is to bring understanding upon these properties.

2.1.1 Time-to-frequency transforms

Although our application is digital and works in the discrete time domain, time-to-frequency-domain transformations are often more meaningful when stated in the continuous-time domain because integration then replaces summation, and integration is more propitious to the finding of explicit solutions. Helpful approximations will hence follow between spectra issued from continuous and discrete time, but the downside is an increased complexity in the notation, which has to account for these two types of spectra.

Furthermore, it is useful to regard the short-time Fourier analyses of signals as the infinite-time analyses of the product of sinusoids with short-time windows. Short-time and infinite-time spectra are therefore also both going to be involved, whether in the continuous or discrete time domain, and need notational distinction as well. To summarise, time-to-frequency-domain transformations will be achieved with the four types of transformations : the Fourier Transform (FT, infinite-continuous-time) ; the Fourier Series (FS, short-continuous-time) ; the Discrete-Time Fourier Transform (DTFT, infinite-discrete-time) ; and the Discrete Fourier Transform (DFT, short-discrete-time). Our convention to distinguish the four is twofold :

- Continuous-time transforms use rounded capital letters, and discrete-time transforms, standard italic capital letters.

- Short-time transforms without a subscript are infinite-time transforms, and transforms that use a subscript N denote transforms taken over an interval (discrete or continuous) $[0, N)$.

For example, $\mathcal{X}$ would be the infinite-time transform of x in continuous-time, and X, in discrete-time, while $\mathcal{X}_N$ would be the short-time transform of x in continuous time, and X_N, in discrete-time.

2.1.2 Arguments

Also, for reasons that will become evident when the explicit definitions of these transforms are given, the units of the arguments in infinite- and short-time transforms differ. For infinite-time transforms, the argument (usually denoted with ω) is in radians per time unit, while for short-time transforms, the argument (usually denoted with b, for *bins*) denotes the number of periods of the basis function $e^{-jb2\pi/N}$ per interval $[0,N[$. For instance, in $X_N(\xi)$, ξ would be in radians per time unit, while in $X(\xi)$, it would be in periods per N time units. Although it seems complicating, it actually facilitates indexation into DFT and FS spectra.

Finally, another notational distinction that was found useful and that is often used as a convention in articles and textbooks, is the use of round brackets for continuous arguments, and square brackets, for discrete arguments. Hence, from such expression as $x(n)$, the reader should infer that $n \in \mathbb{R}$, while for $x[n]$, that $n \in \mathbb{Z}$.

2.2 Windowing

For the sake of clarity, we assume for now that our input were continuous-time. Spectra inherited from discrete inputs will be considered in due time.

Let us consider some input $x(n)$, of which we want to get the spectrum. Among the Fourier analysis tools at our disposal, we pick the Fourier series, as our recording is continuous and of finite length. The $[0,N]$-interval Fourier series is

$$\mathcal{X}_N[b] = \int_0^N x(n)e^{-jb2\pi n/N}dn. \tag{2.1}$$

Single-component input

We saw in Chapter 1 that a string tone was a sum of sinusoids, whose amplitude and frequency could be considered static over some reasonably short interval of time. In the frequency domain, each time-domain sinusoid transforms to two peaks, as the basis function of the Fourier transform is a complex exponential, $e^{jb2\pi/N}$, and a real sinusoids splits in two complex exponentials, as per Euler's formula,

$$\cos\theta = \frac{1}{2}\left(e^{j\theta} + e^{-j\theta}\right).$$

To begin at the simplest, our signal $x(n)$, to start off, is thus

$$x(n) = Ae^{j\phi}e^{jr2\pi n/N}, \tag{2.2}$$

a single component of specific parameters: a magnitude A, initial phase ϕ,[1] and frequency $r2\pi/N$. Here, the frequency is written in such a way that it depends on r, so as to say that there are r periods of the signal in the Fourier series interval $[0, N]$, or in other words, that the frequency of x is r times the fundamental frequency of the analysis.

The Fourier series of x as as defined in (2.2) is

$$\begin{aligned} \mathcal{X}_N[b] &= Ae^{j\phi}\int_0^N e^{-j(r-b)2\pi n/N}dn \qquad (2.3) \\ &= j\frac{Ae^{j\phi}N}{(r-b)2\pi}\left(e^{-j(r-b)2\pi} - 1\right). \end{aligned}$$

In the simple case where there is an integer number of periods of x in the

[1] A could be negative, but the negative sign might as well be included in the phase information, as a π phase shift is equivalent to a reverse of sign, i.e. $e^{j\pi} = -1$.

analysis interval, or $r \in \mathbb{Z}$, then

$$\mathcal{X}_N[b] = \begin{cases} Ae^{j\phi}N & r = b, \\ 0 & r \neq b. \end{cases} \tag{2.4}$$

This results in a very tidy magnitude spectrum, easy to read, as there is only one peak, whose height is A, scaled by the length of the analysis N. Such a spectrum is presented in the upper plot of Figure 2.1.

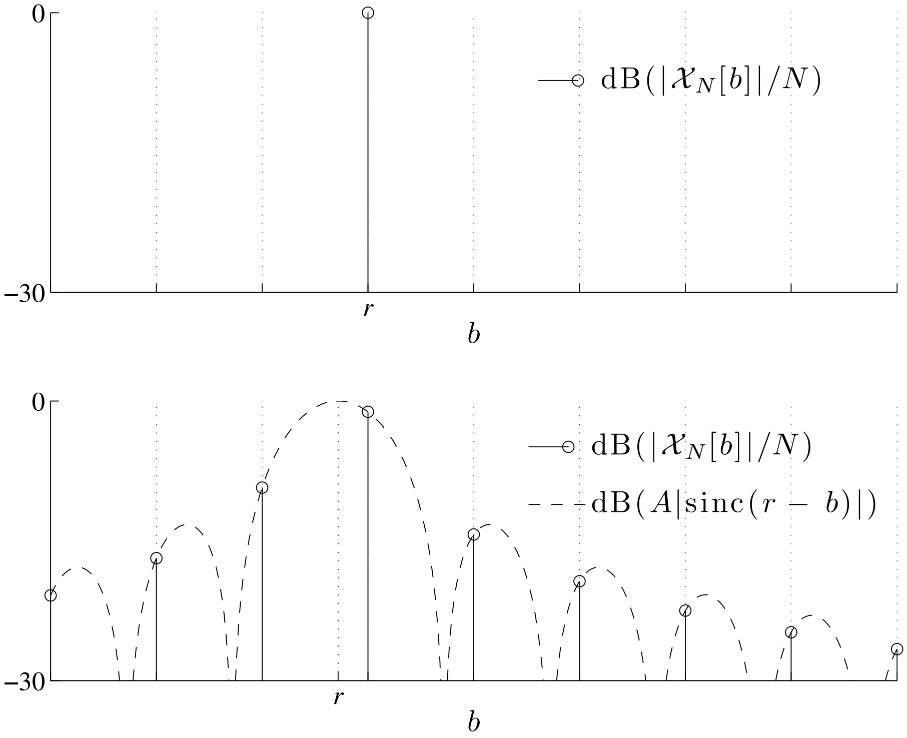

Figure 2.1: Discrete spectra of single complex exponential, synchronised (top) and out-of-sync (bottom) in the Fourier analysis period.

In general, however, the components of the input signal cannot be expected to be integer multiples in frequency of the fundamental frequency of the analysis. We must therefore consider the more general case where $r \in \mathbb{R}$, which yields the spectrum

$$\mathcal{X}_N[b] = NAe^{j\phi}e^{-jb\pi}\text{sinc}(r-b), \tag{2.5}$$

where $\text{sinc}(x) = \sin(\pi x)/\pi x$. An example of a Fourier series magnitude for some non-integer r is shown in the lower plot of Figure 2.1.

It is evident, both from (2.5) and the lower plot in Figure 2.1, that the main lobe of the interpolated spectrum culminates above r, the frequency index of the analysed component. Hence, given some appropriate method of interpolation, the frequency of the component can be found through the finding of r, and thereafter the magnitude, $A = |\tilde{\mathcal{X}}_N(r)|/N$, and phase, $\phi = \angle\left(e^{jr\pi}\tilde{\mathcal{X}}_N(r)\right)$, where $\tilde{\mathcal{X}}_N(.)$ denotes the interpolation of the Fourier series $\mathcal{X}_N[.]$.

Multi-component input

The fact that the analysed signal is multi-component complicates the matter. To see how, let x be not a single complex exponential, but the sum of K of them, each with their own frequency, magnitude and phase,

$$x(n) = \sum_{k=1}^{K} A_k e^{j\phi_k} e^{jr_k 2\pi n/N}. \tag{2.6}$$

The corresponding Fourier series is the sum of the Fourier series of each individual component,

$$\mathcal{X}_N[b] = e^{-jb\pi} \sum_{k=1}^{K} A_k e^{j\phi_k} N \text{sinc}(r_k - b). \tag{2.7}$$

Then, the (interpolated) spectrum above r_k is not the result of the k^{th} component alone anymore, or $\mathcal{X}_N\left(r_k\right) \neq A_k e^{j(\phi_k + r_k\pi)} N$. This tells us that the direct reading of the spectrum of a multi-component signal cannot yield exactly the parameters of its individual components. Either some processing of the spectrum must be done beforehand, or some approximation threshold be set. The latter option is considered now.

The sinc(b) function features an asymptotic trend of a $1/b$ decay as its argument is further from 0. In decibels, this is a $-20\log_{10} b$ decay, or approximately six decibels per octave. On this basis, the threshold below which any spectral value is negligible can be set, and a corresponding interval B defined, such that

$$\mathcal{X}_N(b) \approx 0, \;\; |r - b| > B/2,$$

in the case of a single-component spectrum of frequency index r. Then, for some multi-component spectrum, it can be considered that the parameters of the components are readily available in the spectrum, as

$$\mathcal{X}_N\left(r_k\right) \approx A_k e^{j\phi_k} N \text{sinc}(r_k - b), \tag{2.8}$$

so long as a minimal distance B is respected between every partial frequency, i.e.

$$|r_k - r_l| > B, \qquad\qquad \forall\ k,\ l \in [1, K]. \tag{2.9}$$

The idea is illustrated in Figure 2.2. In the upper plot, condition (2.9) is respected, and it can be seen that the main lobe of each component is faithful to that of a single sinc function. In the lower plot, however, it is not, and the lobes are distorted. Then, the frequency index underneath a given magnitude maximum cannot be trusted to represent correctly that of the corresponding

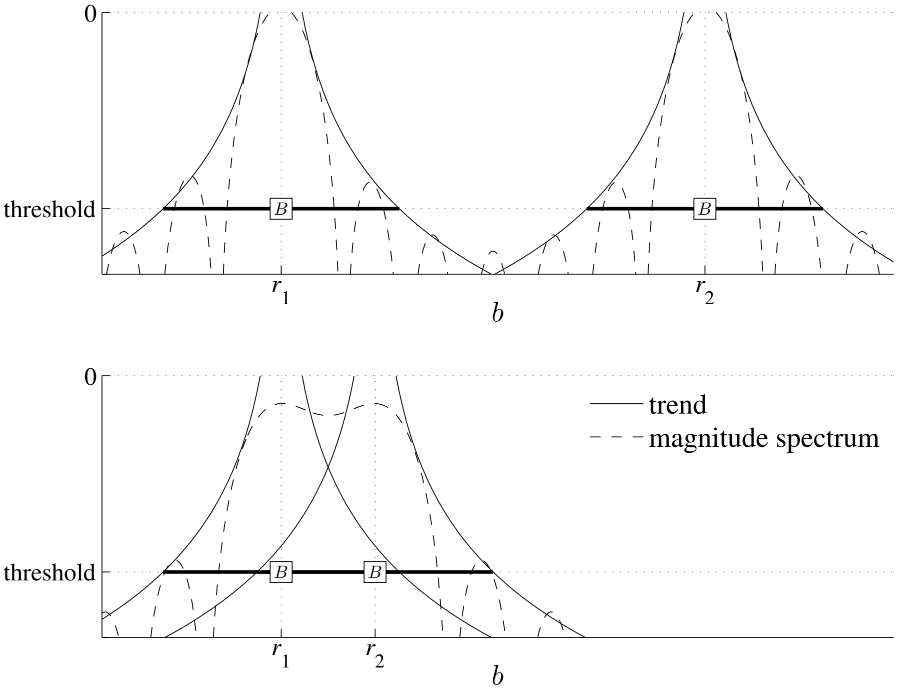

Figure 2.2: Multi-component spectra, with negligible (top) and prejudicial (bottom) overlap.

component anymore. Also, the figure shows that the magnitude value at these maxima is affected, here, dramatically. The same goes for the phase, which is not represented on the figure.

We can now say that as long as some distance between the center frequencies is respected, the parameters of the partials can be faithfully obtained. The decay trend of a component's spectral representation does not have to be that of a sinc function, though. The latter is indeed rather slow. For example, the threshold under which some spectral energy can be considered negligible may be set to -60 decibels. For the decay trend of the sinc function, this gives the equation $-20\log_{10}(B) = -60$, which implies a distance B of one thousand frequency index units between adjacent partials. Considering that in this thesis, we are dealing with harmonic signals, with components evenly spaced in frequency, it would imply that at least a thousand periods of the recorded wave is taken into one analysis frame. Yet our string output, with its partials decaying exponentially and some potential frequency glide, can only be considered to be static over a much shorter period of time. It is therefore desirable to find a means of accelerating the decay of the sinc's trend. This is the object of the next section.

2.2.1 Rectangular window

Some great insight into the spectral curves seen in the previous section can be achieved when, for a start, it is realised that the Fourier series of the complex exponential (2.3) is equal to the Fourier transform of the product of this same signal with a *rectangular window*, i.e.

$$\begin{aligned}\mathcal{X}_N[b] &= \int_0^N x(n)e^{-jb2\pi n/N}dn \\ &= \int_{-\infty}^{\infty} x(n)\ {}_\varsigma\overset{\mathrm{r}}{w}\ (n)e^{-jb2\pi n/N}dn, \qquad (2.10)\end{aligned}$$

where

$${}_\varsigma\overset{\mathrm{r}}{w}\ (n) = \begin{cases} 1 & n \in [0, N) \\ 0 & n \notin [0, N). \end{cases}$$

The notation of our rectangular window ${}_\varsigma\overset{\mathrm{r}}{w}$ may seem overly complicated, but it will be seen further down this development how it fits in a broader picture. The stacked r, on top of the w, stands for *rectangular* – other types of windows with different properties will be seen later, and it is useful to make the difference in their notation. As for the prescript ς, we ask the reader to ignore it for now : its meaning and use will be shown in due time during the course of this development.

As demonstrated in Appendix C.2, the Fourier transformation of the product of two signals equals the convolution of their spectra, divided by the period interval of the basis function, here, N.[2] So we write

$$\mathrm{FT}\left\{x(n) \cdot\ {}_\varsigma\overset{\mathrm{r}}{w}\ (n)\right\} = \frac{1}{N}\left(\mathcal{X} *\ {}_\varsigma\overset{\mathrm{r}}{\mathcal{W}}\ _N\right)(\omega), \qquad (2.11)$$

[2] In the appendix, this is 2π, as the independent frequency variable is then in radians. See Equation (C.11), page 256

where we have substituted ω for $b2\pi/N$, and used the definition,

$$\mathrm{FT}\{x(n)\} \triangleq \mathcal{X},$$

$\mathcal{X}$ being the Fourier transform of x.

Let us therefore examine the Fourier Transform of the single-component x of Equation (2.2) on its own:

$$\begin{aligned}\mathcal{X}(b2\pi/N) &= Ae^{j\phi}\int_{-\infty}^{\infty} e^{-j(b-r)2\pi n/N}dn \\ &= Ae^{j\phi}N\delta(b-r). \qquad (2.12)\end{aligned}$$

By substitution of (2.12) into (2.11), it can be found that

$$\begin{aligned}\mathrm{FT}\left\{x(n)\cdot\ {}_{\varsigma}\overset{\mathrm{r}}{\dot{w}}\ (n)\right\} &= Ae^{j\phi}\ {}_{\varsigma}\overset{\mathrm{r}}{\dot{\mathcal{W}}}\ {}_{N}\left((b-r)\frac{2\pi}{N}\right) \\ &= Ae^{j\phi}\ {}_{\varsigma}\overset{\mathrm{r}}{\dot{\mathcal{W}}}\ {}_{N}(b-r),\end{aligned}$$

where, we recall the user, ${}_{\varsigma}\overset{\mathrm{r}}{\dot{\mathcal{W}}}\ {}_{N}(b-r)$ is, as indicated by the subscript N, the Fourier series of ${}_{\varsigma}\overset{\mathrm{r}}{\dot{w}}$, but with a continuous-frequency argument $b-r$, as indicated by the round brackets.

This sheds light on the Fourier series of the single-component complex exponential (2.5): we now see that the Fourier series of $Ae^{j\phi}e^{jr2\pi n/N}$ over the interval $[0, N]$ really is the spectrum of the rectangular window ${}_{\varsigma}\overset{\mathrm{r}}{\dot{\mathcal{W}}}$, only scaled by A, phase-shifted by ϕ radians, and frequency-shifted by r frequency bins.

The rectangular window, albeit the simplest of all, is nevertheless a window. To make them symmetric about 0, and hence ensure that their spectrum is real – which is convenient – windows are conventionally centered around $n = 0$ [Har78, Nut81]. For clarity, we have been using ${}_{\varsigma}\overset{\mathrm{r}}{w}$, which in fact is the zero-centered rectangular window $\overset{\mathrm{r}}{w}$, delayed by half its length, i.e.

$${}_{\varsigma}\overset{\mathrm{r}}{w}\,(n) = \overset{\mathrm{r}}{w}\left(n - \frac{N}{2}\right). \tag{2.13}$$

By the time-shift property of the Fourier transform (see Appendix C.1), the spectrum of ${}_{\varsigma}\overset{\mathrm{r}}{w}$ results in being that of $\overset{\mathrm{r}}{w}$, frequency-modulated by $e^{-jb\pi}$, i.e.

$${}_{\varsigma}\overset{\mathrm{r}}{\mathcal{W}}_{N}(b) = e^{-jb\pi}\,\overset{\mathrm{r}}{\mathcal{W}}_{N}\,(b). \tag{2.14}$$

We can now conclude this section with the derivation of the rectangular window spectrum proper, free of frequency modulation:

$$\begin{aligned}\overset{\mathrm{r}}{\mathcal{W}}_{N}\,(b) &= \int_{-N/2}^{N/2} e^{-jb2\pi n/N} dn \\ &= N\mathrm{sinc}(b),\end{aligned}$$

a simple N-scaled sinc function.

2.2.2 The Hann window

Let us recall that the motivation of this development is to accelerate the decay trend of the sinc spectrum, which we now know to be that of the rectangular window. The question is therefore to find a window with faster-decaying side lobes.

We can think in terms of “brightness”, or richness in high-frequency components. The fact that the side lobes of the spectrum of the rectangular window decay slowly with frequency is a one-way relationship with the brightness of the rectangular window. In its original domain, the rectangular window is a sudden jump from 0 to 1, and a subsequent drop back to 0, An illustration is provided in the upper plot of Figure 2.3, in solid line. These infinitely sharp corners are responsible for strong high-frequency components. To reduce these components, and thereby accelerate the decay rate of our window's trend, one can consider picking a smoother window.

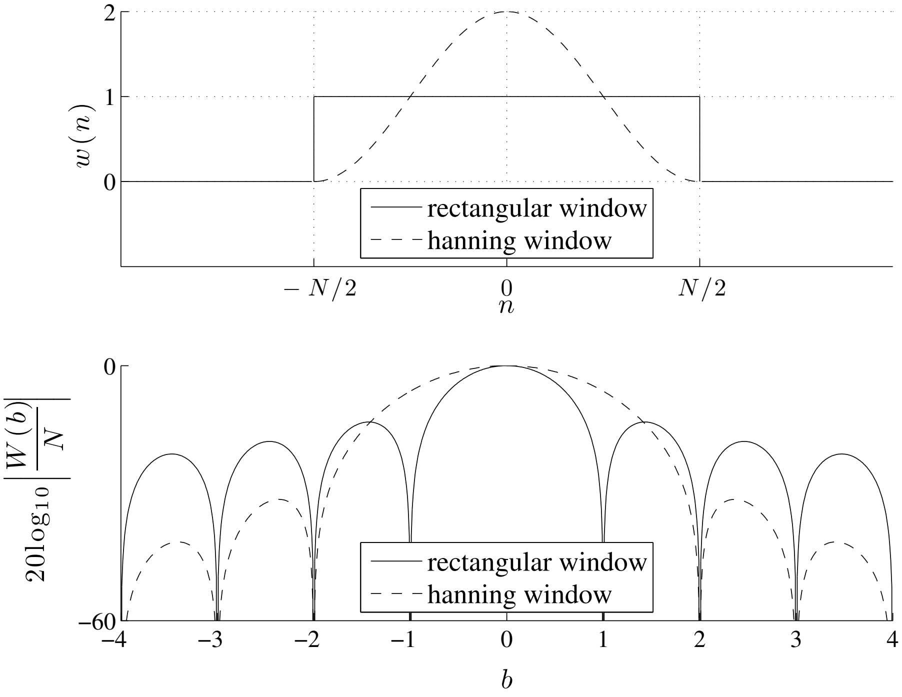

Figure 2.3: Rectangular and Hann windows, in time (top) and frequency (bottom) domains.

A good window to start with is the Hann window,

$$\overset{\mathrm{h}}{w}(n) = \begin{cases} 1 + \cos\frac{2\pi n}{N} & -\frac{N}{2} \le n \le \frac{N}{2} \\ 0 & |n| \ge \frac{N}{2}, \end{cases} \tag{2.15}$$

seen alongside the rectangular window in the upper plot of Figure 2.3. Not only is it visibly smooth, but also, its spectrum has a relatively simple analytical expression. Indeed, with the help of Euler's formula, we see that the window is the sum of three complex exponentials, of frequencies $-2\pi/N$, 0, and $2\pi/N$, and respective magnitude 1/2, 1, and 1/2. Hence, by the convolution theorem, we can foresee that the Fourier transform of the Hann window is the sum of three sinc functions, scaled and shifted accordingly. This sum simplifies to

$$\overset{\mathrm{h}}{\mathcal{W}}_N(b) = \frac{N}{\pi}\frac{\sin b\pi}{b^3 - b},$$

which, in accordance with our speculation, is a spectrum with a faster decay rate, of $-60\log_{10}(b)$, or 18 decibels per octave, thrice as much as the rectangular window's. Now, in the presence of a harmonic signal to analyse, and to achieve a -60 decibel attenuation between the center frequency of each partials, a minimum of ten periods is required. This is already much more practical than the thousand mentioned before, in the case of the rectangular window.

In fact, the Hann window, and even the rectangular window, belong to an entire family of windows, made up of cosine waves. The next section is dedicated to the description of these windows, which we may call *cosine* *windows*.

2.2.3 Cosine windows

In [Nut81], Nuttall explains that the $1/b$ decay trend of the rectangular window spectrum is due to the discontinuity of the window at the boundaries of its non-zero interval $[-N/2, N/2]$. Consistently, the Hann window at these points is zero,

$$\overset{\text{h}}{w}\left(\pm\frac{N}{2}\right) = 0,$$

and its decay is faster than $1/b$.

More generally, it is said in [Nut81] that the discontinuities of a window at its boundaries in its derivatives dictate the asymptotic behaviour of its spectrum – the higher the derivative order of continuity, the faster the decay. Regarding the Hann window, we observe that it is also continuous in its first derivative,

$$\begin{aligned} \left.\frac{d\overset{\text{h}}{w}}{dn}\right|_{\pm\frac{N}{2}} &= -\frac{2\pi}{N}\sin\pi \\ &= 0 \end{aligned}$$

but not in its second. However, it is possible to satisfy

$$\left.\frac{d^q w^P}{dn^q}\right|_{\pm\frac{N}{2}} = 0 \tag{2.16}$$

for some differentiation order $q \leq 2P$, given the presence of the first $P +$

1 cosine harmonics (including zero-frequency) in the non-zero part of the window. The expression for *cosine windows* thus generalises to

$$w^P(n) = \sum_{p=0}^{P} a_p \cos\frac{p2\pi n}{N}, \quad |n| \leq \frac{N}{2}, \tag{2.17}$$

where the $P+1$-length coefficient vector $\mathbf{a} = [a_0, a_1, ..., a_P]^T$ remains to be found.

The problem of finding $\mathbf{a}$ can be reduced to solving a system of linear equations. First, it should be realised that every odd-order derivative of a cosine window is made of $\sin(p2\pi n/N)$ components, which are all zero at $n = \pm N/2$, making the window's derivative continuous at its boundaries. Odd-order derivatives therefore need not to be worried about. Hence, the P first even-order derivatives can be equated to zero by adjustment of the $P+1$ coefficients of the window. For example, for a third-order cosine window, there are four coefficients, and these can be used in a linear system to make the 0^{th}, 2^{nd} and 4^{th} derivatives of the window continuous for all n.

But this is $P+1$ coefficients for P equations – an underdetermined system, with an infinity of solutions. The missing equation shows up with the fact that a complex exponential of magnitude 1 should come out in the magnitude spectrum as a peak of height N, for some N-length analysis. This is equivalent to saying that the transform of the window should be N at frequency 0, or

$$\int_{-\infty}^{\infty} w^P(n)dn = N. \tag{2.18}$$

The cosine components of w^P are periodic in N, and for this reason, their integration is going to yield 0. The constant part of the window is thus the only contributor at frequency 0, and so we are left with $a_0 \int_{-N/2}^{N/2} dn = N$, or $a_0 = 1$. With the first coefficient of the vector forced to be 1, P coefficients remain to be determined. In matrix notation, this can be written as

$$\begin{bmatrix} d_1^0 & d_2^0 & \cdots & d_P^0 \\ d_1^2 & d_2^2 & \cdots & d_P^2 \\ d_1^4 & d_2^4 & \cdots & d_P^4 \\ \vdots & \vdots & \ddots & \vdots \\ d_1^{2(P-1)} & d_2^{2(P-1)} & \cdots & d_P^{2(P-1)} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad (2.19)$$

where

$$d_p^q = \frac{d^q}{dn^q} \cos \frac{p2\pi n}{N}\bigg|_{\frac{N}{2}}.$$

Continuous windows and Minimal windows

Cosine windows whose coefficients are derived as per (2.19) may be called *continuous windows*. Figure 2.4 shows the spectra of the first four such windows. The decay trend for continuous windows can seemingly be generalised as $(2P+1)6$ decibels per octave, for a large frequency index (see minimal window trends in Figure 2.5). This is partly explained by the generalised spectrum $\mathcal{W}_N^P$ of a cosine window or order P,

$$\mathcal{W}_N^P(b) = N \frac{\sin b\pi}{\pi} \left(\frac{1}{b} + b \sum_{p=1}^{P} \frac{a_p(-1)^p}{b^2 - p^2} \right),$$

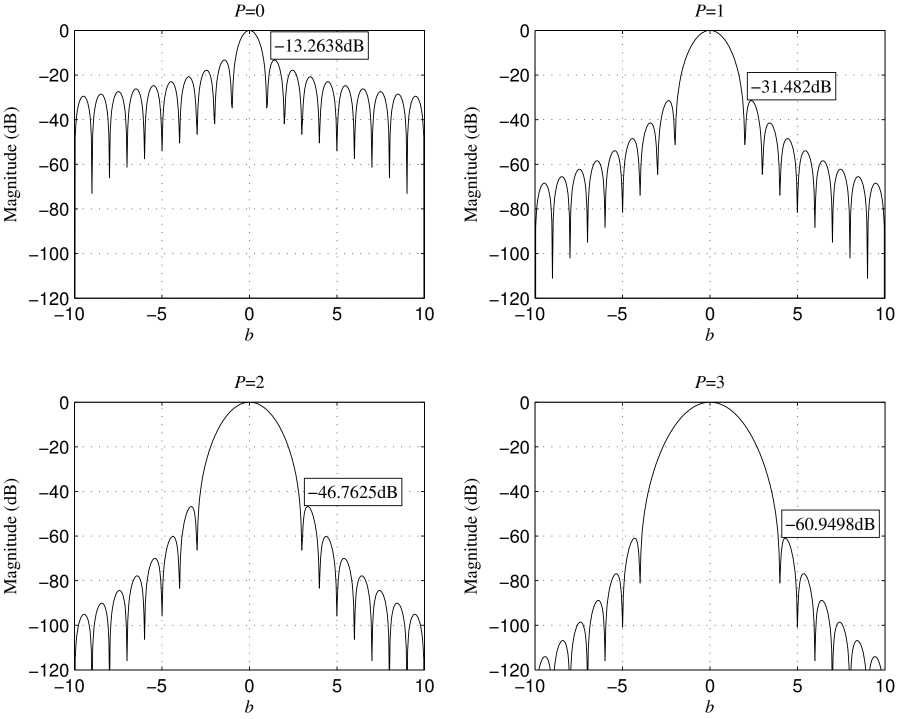

Figure 2.4: Magnitude spectra of first four continuous windows.

for which the addition of the numerators requires the multiplication of the denominators, a process during which the powers of b add up. However, other cosine windows, with discontinuous derivatives, share the same spectral expression, and yet have slower decays.

Instead of finding the coefficients with the sole purpose of the acceleration of the decay trend, the reduction of the highest *sidelobes* can be set as a priority. We may call these *minimal sidelobe windows*, or *minimal windows* for shorthand. The simplest example of such a window is the Hamming window, a first-order cosine window, discontinuous at its boundaries, but whose largest sidelobe is -43.19 decibels, as opposed to -31.47 for the Hann window. The coefficients for this minimal sidelobe window are given in [Nut81], but no explanation is given there as to how they were obtained, neither for the Hamming window, nor for higher-order windows with minimal sidelobes. Figure 2.5 shows the spectral trend of the first- (upper plot), second- (middle plot) and third-order (lower plot) continuous and minimal sidelobe windows, and Table 2.1 lists the properties and coefficients of these windows.

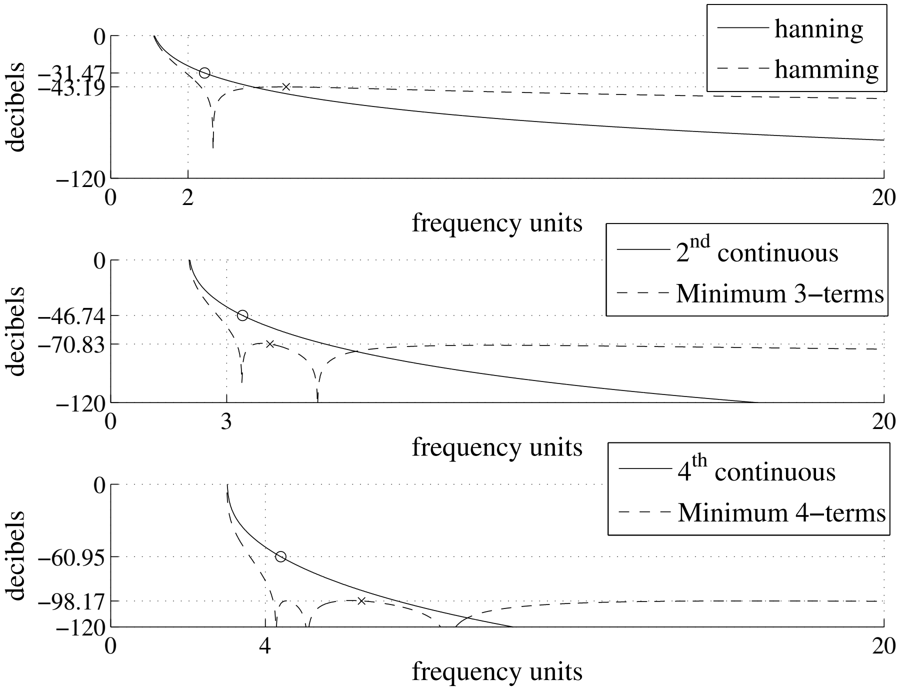

Figure 2.5: Trend in spectra of 1st-, 2nd- and 3rd-order continuous (solid lines) and minimal (dashed lines) windows.

Here, the coefficients of the minimal sidelobe windows were read from [Nut81][3]. This article shows a wider variety of windows. For instance, win-

[3] Some names found in the table were adapted from the original article to the context of this thesis. For example, the ”Continuous third derivative of weighting” window became here the ”2nd-order continuous window”, as all odd-order derivatives were shown to be zero at the boundaries anyway.

order	name	coefficients	decay	sidelobe
0	rectangular	1	-6	-13.26
1	Hann	$[1, 1]$	-18	-31.47
	Hamming	$\left[1, \frac{349}{407}\right]$	-6	-43.19
2	2^{nd} continuous	$\left[1, \frac{4}{3}, \frac{1}{3}\right]$	-30	-46.74
	”Minimum” 3-term	$\left[1, \frac{1152}{983}, \frac{515}{2792}\right]$	-6	-71.48
3	4^{th} continuous	$\left[1, \frac{15}{10}, \frac{6}{10}, \frac{1}{10}\right]$	-42	-60.95
	”Minimum” 4-term	$\left[1, \frac{1667}{1239}, \frac{866}{2305}, \frac{161}{5501}\right]$	-6	-98.17

Table 2.1: 0^{th}, 1^{st}-, 2^{nd}- and 3^{rd}-order continuous and minimal windows.

dows need not be continuous *or* minimal sidelobe – they can be a bit of both. Some high-order window can be designed in such a way that its first derivative (or higher) is continuous, thereby guaranteeing some 18dB (or more) per octave decay rate, but the remaining degrees of freedom are used to minimise the side lobes, instead of further accelerating the decay trend. We dwelled here on the extreme cases to emphasize the contrast in properties there can be between cosine windows, even of the same order. Also, Table 2.1 offers already plenty to choose from and to discuss.

First of all, it is pertinent to state that the width of the main lobe depends on the order of the window, i.e.

$$B = 2(P + 1). \tag{2.20}$$

This is well illustrated in Figure 2.4. A rule of thumb to follow in spectral analysis is to avoid the overlap of adjacent main lobes [IS87], which requires the inclusion in the analysis of at least $2(P+1)$ fundamental periods of the harmonic sound. Hence, higher-order windows require longer analysis periods. This would not be a problem if the analysed sound were truly static, but it is not the case: string components feature an exponential decay, and their frequency, sometimes a downward glide. Projection of some time-varying frequency components onto some static complex exponentials during the Fourier transformation may introduce unpredictable bias. Care must therefore be taken not to stretch the length of the analysis too much.

If avoiding overlapping of adjacent main lobes is the only requisite, then minimal sidelobe windows must be favoured. However, the maximum sidelobe level of a window may not provide sufficient attenuation – recall that, in the previous section, we were considering some -60 decibels attenuation. For such constraint, the Hamming window would prove less convenient than the Hann window, 1st-order as well. In fact, the upper plot in Figure 2.5 shows that only over a spectral region of about one frequency unit, immediately outside the main lobe, is the outline of the Hamming window lesser than that of the Hann window. Some similar comment could be given on the next two subplots as well, comparing the trends of 2nd- and 3rd-order continuous and "minimal" window spectra. Yet, the -70.83dB minimal attenuation of the Minimum 3-term may be found sufficient, in which case the second-order minimal window is preferable to the second-order continuous window, due to the steeper decay near the main lobe[4].

[4] Terminology here may be confusing. A P^{th}-order window has $P+1$ terms, or coeffi-

One final point regarding the choice between a continuous window or a minimal sidelobe window may regards the appearance of the spectrum. Minimal windows have the visual advantage over continuous windows of a very flat floor of side lobes (which may be considered a noise floor). Figure 2.6 gives an illustration of this tendency, which is due to the greater drop in energy immediately on either side of the main lobe, and to the more flat decay trend. Spectra obtained with minimal windows may therefore look more like the ideal spectrum, featuring peaks where there are frequency components, and a flat floor elsewhere. This might also be useful for discerning main lobes

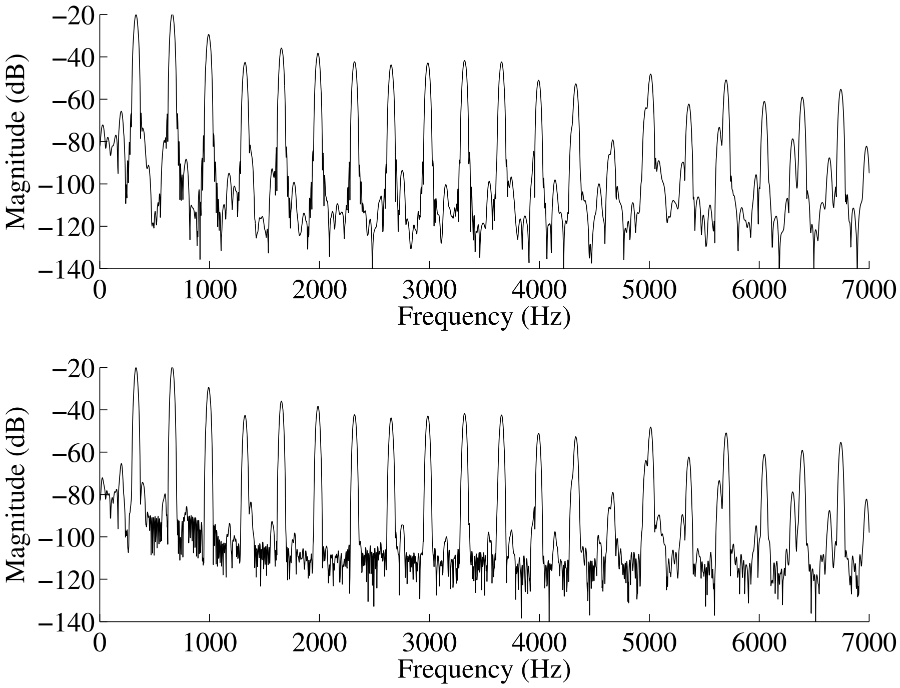

Figure 2.6: Acoustic guitar spectra derived with 2^{nd}-order continuous (top) and minimal (bottom) windows. Notice the steady noise floor in the lower plot.

cients. Hence, the Minimum 3-term is a fourth-order window.

from sidelobes, which, in automated peak detection, can at times be found difficult.

Constant-sum property of cosine windows

Before we have a look at windows other than cosine windows, an important property of these should be evoked. Cosine windows are such that, when they are duplicated, spaced by some regular time interval, and optionally raised to some power, they can add up to some constant. This property is formalised in the equation below,

$$\sum_{u=-\infty}^{\infty} \left(w^P \left(n - \frac{uN}{PQ+i} \right) \right)^Q = \alpha, \tag{2.21}$$

where Q and i are positive integers, i greater than 0, and α is a constant real number. In (2.21), the ratio $N/(PQ+i)$ is the number of time units separating each window, and is known as the *hop size* in the literature around the Short-Time Fourier Transform (STFT) [Por81, Cro80, Ser89, Zö02].[5] Alternatively, we can evoke the *overlap* of the windows, which is sometimes expressed as a percentage [IS87], or a factor

$$O = PQ + i. \tag{2.22}$$

This property is illustrated in Figure 2.7, where only a finite number of window overlap, to outline the phenomenon. There, the window order

[5] In discrete time, these time units are samples, and then the window length has to be an integer multiple of $PQ+i$.

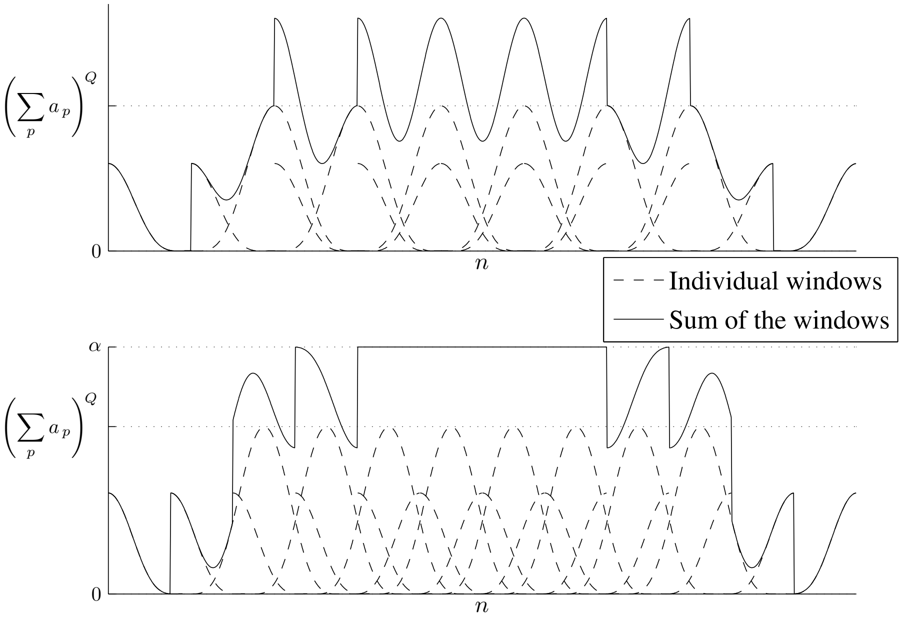

Figure 2.7: Constant-power-sum property of Cosine windows. In the upper plot, the minimal overlap required is short of one, the condition of equation (2.21) that $i > 0$ not being respected. The lower plot respects this condition: see how the sum suddenly stabilises over the interval where $PQ + i$ windows overlap. The coefficients of the cosine window were chosen randomly, to emphasize the phenomenon.

P is set to 2 (i.e. three-terms cosine), the power Q, to 2. In the upper plot, the overlap O is one short of the minimum for the windows to add up to a constant. In the lower plot, the minimal overlap required is satisfied. The sum of the windows then clearly stabilises to α. The coefficients of the window were chosen randomly, so as to emphasize the phenomenon.

We now set to give an explanation for this property. To simplify the rest of the development, we begin by noting that when the time shift index in (2.21) reaches the number of overlapping windows, i.e. $u = PQ + i$, the time shift becomes equal to one window interval. Instead of an infinity of windows, we can thereby do as if the windows were non-zero over infinity (as opposed to $[-N/2, N/2]$ only) and replace the left-hand side term in (2.21) with a sum of $PQ+i$ of them, i.e.

$$\sum_{u=0}^{PQ+i-1}\left(\sum_{p=0}^{P} a_p \cos\left[p\frac{2\pi}{N}\left(n-\frac{uN}{PQ+i}\right)\right]\right)^{Q}=\alpha.$$

We know that a cosine window of order P contains $P+1$ cosine terms, of respective frequencies $p2\pi/N$ for $p=0,1,...,P$. Raised to the power Q, the window now features all frequencies $p2\pi/N$ for $p=0,1,...,PQ$, and each p^{th} frequency term is then scaled by a new factor, let us call it c_p. On this basis, we re-express the Q^{th} power of our cosine window as

$$\left(w^P(n)\right)^Q=\sum_{p=0}^{PQ} c_p \cos p\frac{2\pi n}{N}, \quad |n|\leq N/2. \tag{2.23}$$

Now it is possible to express (2.21) as a linear equation,

$$\sum_{p=0}^{PQ} c_p \sum_{u=0}^{PQ+i-1} \cos\left(2\pi p\frac{u}{PQ+i}-2\pi p\frac{n}{N}\right)=\alpha. \tag{2.24}$$

For the left-hand side term in 2.24 to be constant, the components of non-zero frequency must cancel out. It turns out that the sum of samples of a cosine wave, taken over p periods at a rate of $(p+i)/p$ samples per periods, is invariably nil. For us, this translates as

$$\sum_{u=0}^{PQ+i-1} \cos\left[p\left(2\pi\frac{u}{PQ+i}+\phi\right)\right] = \begin{cases} 0 & p \neq 0, \\ PQ+i & p = 0, \end{cases}$$

meaning that only the zero-frequency components are to survive the summation. Hence we get

$$\alpha = c_0(PQ+i) = c_0 O,$$

c_0 being the factor of the cosine terms of frequency zero after the window is raised to the power Q.

Property (2.21) is useful for transparent analysis/processing/resynthesis using the Short-Time Fourier Transform. The FFT-based Phase Vocoder, for instance, processes a signal in the frequency domain. In general, windowing is then necessary both going from time domain to frequency domain *and* from frequency domain to time domain. The interest of the time-to-frequency windowing really is to reduce the cross-interference of partials in the frequency domain, while frequency-to-time windowing is to ensure that the grain smoothly tends to zero at its edges to avoid audible clicks. Technically, windowing a signal segment Q times in succession with the same window is equivalent to windowing the signal once with the window raised to the power of Q. Going to and coming from the frequency domain therefore requires a window and overlap that sums up to a constant when raised to the power of two. In this regard, it can be found that

$$\alpha_2 = \frac{O}{2}\left(a_0^2 + \sum_{p=0}^{P} a_p^2\right), \tag{2.25}$$

denoting by α_Q the constant to which cosine windows raised to the power of Q and overlap-added by a factor of O stabilise. However, it was realised during the course of our research that in the context of frequency-domain partial cancelation, frequency-to-time domain windowing is superfluous, for reasons that will be pointed out in due time. If windowing takes place only once, the summing constant is then simply

$$\alpha_1 = Oa_0. \tag{2.26}$$

2.2.4 Other windows

Windows need not be made of cosine components. A window, after all, is only a composite mathematical function that is non-zero over a given interval, and zero elsewhere, i.e.

$$w(n)\begin{cases} \neq 0 & |n| \leq \frac{N}{2}, \\ = 0 & |n| > \frac{N}{2}. \end{cases} \tag{2.27}$$

A wide variety of windows exists, an extensive list of which is given and described in [Har78, Nut81]. The most popular seem to be the cosine windows, but some other windows are regularly found in the literature as well, notably the Kaiser-Bessel window, and the Gaussian window.

Kaiser-Bessel window

The Kaiser-Bessel window (or simply *Kaiser window*) is defined as

$$\overset{\mathrm{k}}{w}\,(n) = \begin{cases} I_0\left(\alpha\pi\sqrt{1-\left(\frac{2n}{N}\right)^2}\right) & |n| \le \frac{N}{2} \\ 0 & |n| \ge \frac{N}{2} \end{cases},$$

where $I_0(x)$ is the 0^{th}-order modified Bessel function of the first kind, i.e.

$$I_0(x) = \sum_{k=0}^{\infty} \left[\frac{\left(\frac{x}{2}\right)^k}{k!} \right]^2 .$$

The spectrum of this window can be expressed as [Nut81]

$$\overset{\mathrm{k}}{\mathcal{W}}_N\,(b) = N \frac{\sin\left(\pi\sqrt{b^2-\alpha^2}\right)}{\pi\sqrt{b^2-\alpha^2}}, \tag{2.28}$$

from which it is very visible that, for $\alpha = 0$, the Kaiser-Bessel window reduces to a rectangular window.

It is visible in (2.28) that the parameter α can be used to set the first zero-crossing index b_0 of the spectrum, and consequently the width of the main lobe, satisfying the equality

$$\alpha = \sqrt{b_0^2 - 1}, \quad |b_0| > 1. \tag{2.29}$$

Figure 2.8 shows the Kaiser window and window spectrum for two different settings of b_0. It is evident from both the figure and (2.29) that the zero first crossing does not need to be at an integer frequency index, which is one at-

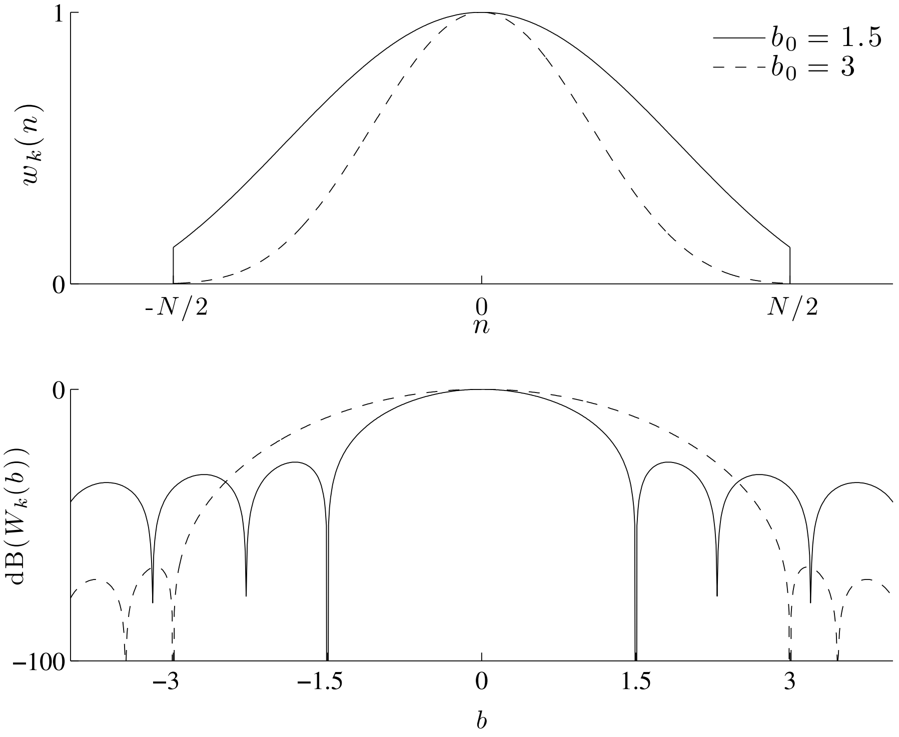

Figure 2.8: Kaiser-Bessel window and window spectrum (normalised), for two different zero-crossing settings.

tractive feature for a window. On the other hand, the roll-off of the sidelobes is bound to 6 decibels per octave. The level of the sidelobes diminishes as the main lobe widens. More specifically, it is found in [Nut81] that, when the main lobe of the Kaiser window is adjusted to be the same width as that of the 1^{st}-, 2^{nd}- or 3^{rd}-order cosine windows, the Kaiser spectrum obtained is very similar to that of the corresponding-order minimal-sidelobe window, with the highest sidelobe greater only by a few decibels.

Gaussian window

The gaussian function has a relatively simple time-domain expression,

$$\overset{\mathrm{g}}{w}(n) = \frac{1}{\gamma_{\mathrm{g}}\sqrt{2\pi}}\exp\left(-\frac{n^2}{2\gamma_{\mathrm{g}}^2}\right), \tag{2.30}$$

which is normalised by the constant $\gamma_{\mathrm{g}}\sqrt{2\pi}$ so that its total area is one, i.e.

$$\int_{-\infty}^{\infty} \overset{\mathrm{g}}{w}(n)dn = 1.$$

One of the many specificities of this function is that its Fourier transform is also a gaussian,

$$\overset{\mathrm{g}}{\mathcal{W}}_N(b) = \exp\left(-b^2 2\pi^2\gamma_{\mathrm{g}}^2/N^2\right).$$

An example of a gaussian is given in solid line in Figure 2.9, with the time-domain waveform in the upper panel, and the decibel spectrum in the lower panel.[6] It should be noted that, as opposed to the sinc-like spectra of the other windows, the gaussian spectrum has no zero-crossings, and consequently no sidelobes. Some other criterion should therefore be set for the width of the lobe. We choose a decibel attenuation criterion, equating the decibel spectrum at some frequency index b_0 to some attenuation α, i.e.

$$20\log_{10}\overset{\mathrm{g}}{\mathcal{W}}_N(b_0) = -\frac{40b^2\pi^2\gamma_{\mathrm{g}}^2}{N^2\log 10} = \alpha \tag{2.31}$$

[6] Notice that, after logarithmic transformation, the gaussian function becomes a quadratic function. Hence, quadratic interpolation in discrete points of the log of a gaussian is *exact*. This may become of interest when the problem of partial frequency estimation is raised, in a later section of this chapter.

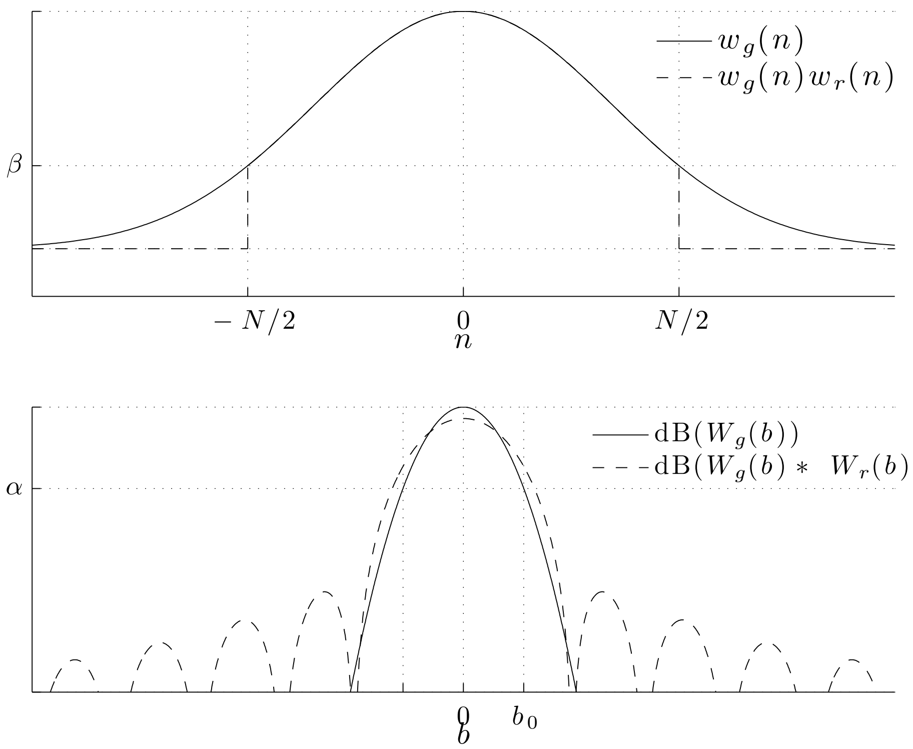

Figure 2.9: Gaussian infinite (solid lines) and truncated (dashed lines) window (linear scale) and spectrum (decibel scale). The frequency-domain approximation by the dashed line is only satisfying when the time-domain lobe is narrow in relation to the analysis interval.

The width of the lobe of the gaussian function is controlled with γ_{g}, which we can now define using (2.31) in terms of the desired attenuation α at the desired frequency index b_0,

$$\gamma_{\mathrm{g}} = \frac{N}{2\pi b_0}\sqrt{-\frac{\alpha \log 10}{10}}. \tag{2.32}$$

The idea is illustrated in the lower plot of Figure 2.9, with the marking of some b_0 and α coordinates.

It has been assumed so far that the time-domain interval of analysis was infinite, but this is of course not the case in practice. For practical use, the time-domain gaussian must be truncated, or, equivalently, be multiplied by a rectangular window, here of length N. The value β_{g} of the truncated gaussian at its boundaries can be obtained by substitution of γ_{g} as defined in (2.32) into the time domain expression of the gaussian (2.30), and evaluation for $n = \pm N/2$, yielding

$$\beta_{\mathrm{g}} = \exp\left(\frac{5\pi^2 b_0^2}{\alpha \log 10}\right).$$

The truncated gaussian window and its spectrum are shown in dashed lines in Figure 2.9. Here, the truncation is largely exaggerated, so as to make the discrepancy between the finite and infinite gaussians visible. This effect was obtained by the setting of a large attenuation α at a small frequency index b_0. It is clear in Figure 2.9 that lowering α and/or getting b_0 closer to 0 would make the lobe narrower. But the gaussian window is subject to the uncertainty principle [Har98] like all other windows, and the width of its lobe in the one domain is inversely proportional to that in the other domain.

The time-domain multiplication of the gaussian window with a rectangular window results in a spectrum that is the convolution of a gaussian with a sinc function. A rectangular window width N that is wide in comparison to the lobe width factor γ_{g} of the gaussian will make the frequency-domain sinc function narrow, impulse-like, with little effect on the gaussian lobe. In the opposite case, i.e. a large sinc lobe for an impulse-like gaussian lobe, the spectrum would be closer to that of a rectangular window. In the lower plot of Figure 2.9, the spectrum of the truncated gaussian features the side lobes reminiscent of the spectrum of the rectangular window. To reduce these, either the attenuation factor α should be made lesser in magnitude (higher on the plot), or the attenuation frequency index b_0 greater, or both.

Now knowing the limits of the gaussian window, we compare its capabilities with those of some minimal sidelobe windows. In Figure 2.10, the solid lines show the spectra of truncated gaussians, and the dashed lines, the spectra of the Hamming (upper plot) and Minimum 3-term [Nut81] (lower plot) windows. The attenuation α of each gaussian window was set to the level of the highest sidelobe of the corresponding cosine window, and the

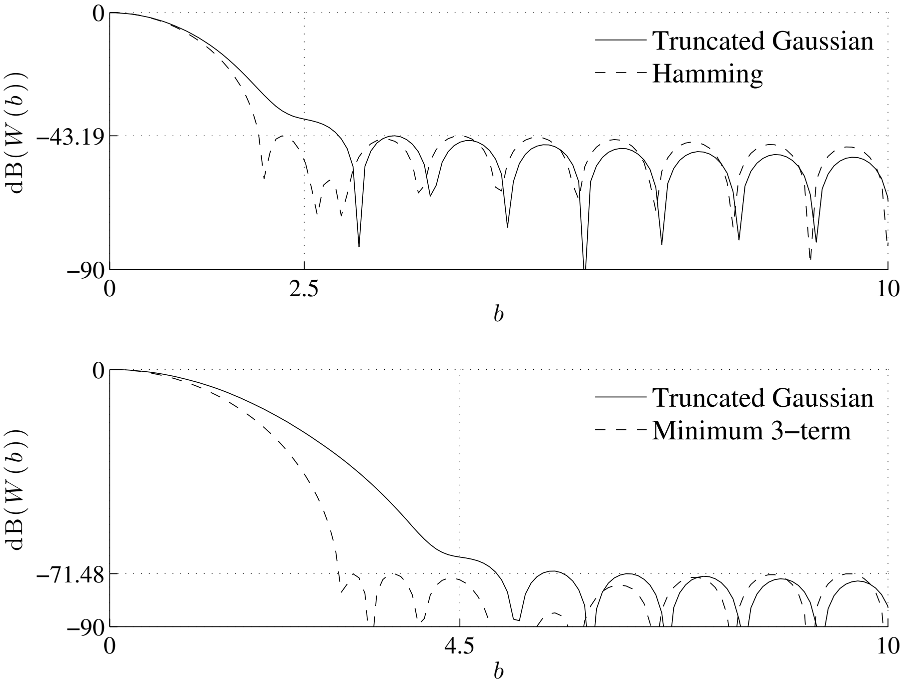

Figure 2.10: Confrontation of truncated gaussian window with the minimum two- (upper plot) and three-terms (lower plot) cosine windows. Cosine windows here show narrower main lobes.

attenuation index, adjusted manually so as to get sidelobes of similar levels. With such a sidelobe constraint, the main lobe of the truncated gaussian window is slightly wider than that of the Hamming window, and wider by almost 50% than the Minimum 3-term.

2.2.5 Discussion

A Window is characterised by a main lobe width and a sidelobe level. In turn, a sidelobe level is characterised by a highest-sidelobe level, and a frequency-dependent decay.

In a way, the main lobe of a window may be seen as a shrine to the parameters of a frequency component, and as such is to remain uncorrupted. Such corruption can be due to the overlapping of another main lobe, or the intrusion of sidelobes from some other component. Regarding main lobe overlapping, it must be avoided by ensuring some minimal frequency index spacing between adjacent components. This spacing, although proportional to the length of the analysis N, is restricted by the available length of the analysed signal on the one hand, and by the assumption that the analysed signal is made of static frequency components on a short interval of time on the other hand. The spacing between the components can therefore not be arbitrarily large. The spacing being restricted, the alternative way of avoiding main-lobe overlap is to use some window with narrow main lobe.

However, a narrow main lobe is synonymous with high sidelobes. Sidelobes are susceptible of intruding into the main lobes of other components, and hence are an important factor to the choice of a window. As sidelobes extend to infinity and are common to all types of windows, some partial is bound to corrupt and be corrupted by other partials in the analysis of multi-component signals. Yet below a certain level, sidelobes can be considered negligible. The analysis can thus be made practical via the acceptance of some sidelobe threshold.

In all three types of windows seen, the width of the main lobe is adjustable, one way or another. For the Kaiser window, it can be adjusted smoothly and continuously, the sidelobe level going down as the lobe goes wide, and *vice-versa*. The width of the truncated gaussian window main lobe is also controllable continuously, but the behaviour of the sidelobes near the main lobe has been seen to be erratic. Finally, the main lobe width for cosine windows is necessarily an integer multiple of the analysis' fundamental frequency, in other words, an integer frequency index multiple.

For all windows, the sidelobe level depends on the width of the main lobe. For cosine windows however, its trend is also adjustable, in terms of decay rate and highest sidelobe level, with high decay rates for high sidelobe levels near the main lobe, or low decay rate but minimal highest sidelobe. Some compromise between these two extrema can also be achieved, by satisfying, in the design of the window, of the derivative constraint up to some lower order than the order of the window would allow, and using the remaining degrees of freedom to minimise the highest sidelobe [Nut81]. Depending on the situation, command over sidelobe decay rate may be useful. Also, measurements in Section 2.2 have shown that the minimal-sidelobe cosine windows performed better than the Kaiser-Bessel and gaussian windows.

Convenience of use can also be a criterion in the choice of a window. Cosine windows are very accessible due to the simplicity of their expressions, and their spectrum is easily derived. In this regard, gaussian windows are easy and even more convenient, inasmuch as their time- and frequency-domain expressions are essentially identical. Yet care must be taken that the approximation due to time-domain truncation does not bias the frequency-domain representation significantly. As for the Kaiser-Bessel window, the frequency-domain expression is very simple, but the infinite sum in its time-domain expression is computationally costly and can only be approximated.

Finally, the constant-sum property of cosine windows may come as an major argument for their use in the context of FFT-based analysis / processing / resynthesis systems. In discrete time, a hop size of one sample is in general required for overlapping windows to sum up to a constant [IS87]. In this regard, cosine windows stand out, as the minimal hop size for a length-N window of order P raised to the power of Q is $N/(PQ+1)$ samples. In terms of *data sample density* [All77], i.e. the number of frequency-domain samples for each time-domain sample, this is a minimum of $PQ+1$ only, as opposed to N in general.

Based on this discussion, an attempt at summarising the advantages and disadvantages of the windows seen in this thesis is given in Table 2.2.

	Pros	Cons
Cosine	· Simplicity of expressions · Unequaled performances · Constant power sum · Control over sidelobe decay	· Each cosine window is a different set of coefficients that need be stored.
Kaiser	· Main lobe first zero adjustable to non-integer index · Results close to minimal windows	· 6dB/octave decay · Costly time-domain synthesis
Truncated gaussian	· Log is quadratic	· 6dB/octave decay · Poor lobe width / sidelobe level trade off

Table 2.2: Recapitulated advantages and disadvantages of windows seen in this thesis.

2.3 Discrete signals

To facilitate the development in Section 2.2, the time independent variable n was considered continuous. Mainly, this allowed the use of the Fourier series as opposed to the Discrete Fourier Transform (DFT), and thereby to deal with integrals as opposed to summations, which is generally easier. However, the application proposed in this thesis is computerised, and works in discrete time. Discrete time signals and signal transforms have a few particularities ; those which matter to our purpose are being examined in this section.

2.3.1 Band-limitedness

First and foremost, it should be said that, inherently, a signal sampled at a rate of f_s samples per second cannot contain frequency components higher than $f_s/2\text{Hz}$. This is another wording of the Sampling Theorem, which, in its paper of origin, states that [Sha49]

> If a function $f(t)$ contains no frequencies higher that $f_s/2\text{Hz}$, it is completely determined by giving its ordinates at a series of points spaced $1/f_s$ seconds apart.

To prove this, we begin with establishing the distinction between a continuous signal and its discrete counterpart,

$$x[n] = x\left(n/f_s\right), \quad n \in \mathbb{Z}.$$

A continuous-time signal can be constructed from the discrete signal with the following convolution [Sha49]:

$$x'(t) = \sum_{n=-\infty}^{\infty} x[n]\text{sinc}\left(tf_s - n\right). \tag{2.33}$$

Taking the Fourier transform of (2.33),

$$\mathcal{X}'(2\pi f) = \int_{-\infty}^{\infty} x'(t)e^{-j2\pi ft}dt,$$

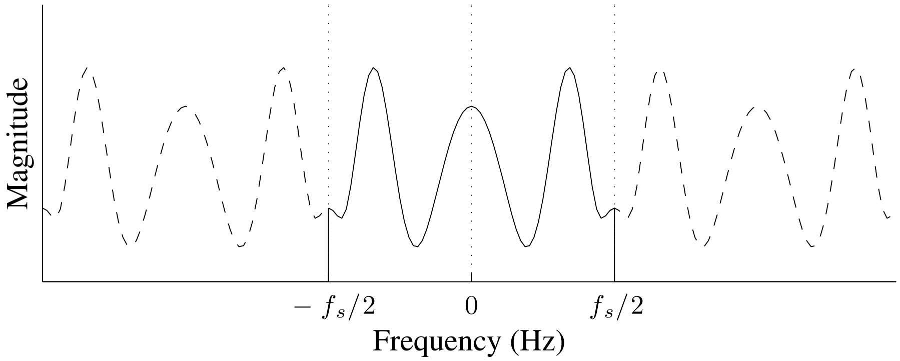

Figure 2.11: Fourier transform (solid line) and DTFT (dashed line) of a discrete-time signal. In continuous time, the discrete signal is expressed as a sum of sinc functions, which induces the rectangular windowing seen in the frequency domain, responsible for the absence of components beyond $|f_s/2|$Hz.

yields the spectrum

$$\begin{aligned} \mathcal{X}'(2\pi f) &= \frac{1}{f_s} \overset{\text{r}}{w}_{f_s}(f) \sum_{n=-\infty}^{\infty} x[n] e^{-jn2\pi f/f_s} \\ &= \frac{1}{f_s} \overset{\text{r}}{w}_{f_s}(f) X\left(\frac{2\pi f}{f_s}\right), \end{aligned} \tag{2.34}$$

where $\overset{\text{r}}{w}_{f_s}$ is a rectangular window of value 1 in the interval $[-f_s/2, f_s/2]$, 0 elsewhere, and X is the Discrete-Time Fourier Transform (DTFT) of $x[n]$,

$$X(\omega) \triangleq \sum_{n=-\infty}^{\infty} x[n] e^{-j\omega n}. \tag{2.35}$$

Equation (2.34) and Figure 2.11 show that the Fourier transform of $x'(t)$ is, due to the frequency-domain rectangular windowing, zero for $|f| > f_s/2$, meaning that the frequency components of $x'(t)$ can only be in the interval

$[-f_s/2, f_s/2]$, and nowhere else. As per the sampling theorem, if the original signal $x(t)$ features no component outside this interval either, then its reconstruction from discrete time is identical, i.e. $x'(t) = x(t)$.

So what happens when a continuous-time signal having components of frequencies beyond $\pm f_s/2$ is sampled? The answer lies in the phenomenon of *aliasing*.

2.3.2 Aliasing

Aliasing is well covered in textbooks, where it is said to occur upon the sampling of a signal which contains frequencies higher than half the sampling rate, called the *Nyquist frequency* [Zö2, Har98, Bou00, Ste96]. The key to understanding aliasing is to realise that discrete-time signals are periodic in frequency. First, to relate the sampling frequency to the system of notation used through Section 2.2, let us state openly the relation

$$\frac{b}{N} = \frac{f}{f_s},$$

saying that the sampling frequency corresponds to frequency index N, and the frequency unit $b = 1$ is f_s/N Hz wide.

Now let us consider the single-component complex exponential already seen in (2.2), except now with the independent variable being an integer,

$$x[n] = Ae^{j\phi}e^{-jr2\pi n/N}. \tag{2.36}$$

This is a signal of angular frequency $r2\pi/N$. Yet, now that $n \in \mathbb{Z}$, it is indistinguishable from a signal of identical magnitude A and phase ϕ, but of frequency $(r + lN)2\pi/N$, for any integer l. Formally,

$$\begin{aligned} x[n] &= Ae^{j\phi}e^{-jr2\pi n/N} \\ &= Ae^{j\phi}e^{-j(r+lN)2\pi n/N}, \quad l \in \mathbb{Z}. \end{aligned} \tag{2.37}$$

In the frequency domain, the situation is consistent with this, the spectrum of a discrete signal being periodic in N, i.e.

$$\begin{aligned} X\left(2\pi\frac{b}{N}\right) &= \sum_{n=-\infty}^{\infty} x[n]e^{-jb2\pi n/N} \\ &= X\left(2\pi\frac{b+lN}{N}\right), \quad l \in \mathbb{Z}. \end{aligned}$$

Figure 2.12 gives an illustration of these two phenomena: on its upper panel, a magnitude spectrum of $x[n]$ as defined in (2.36), is presented over a few frequency periods N. There, we have delineated a frequency band spanning N frequency units, an interval by which the frequency shift of the spectrum is futile, as it yields an identical spectrum. Consistently with the Sampling theorem, this band is chosen to be centered on frequency 0, and hence spans the interval $[-N/2, N/2]$. This band might be called the “useful band”.

Yet, recorded signals are real signals, which can be seen, as per Euler’s formula, as the sum of two complex exponentials, complex conjugates of one

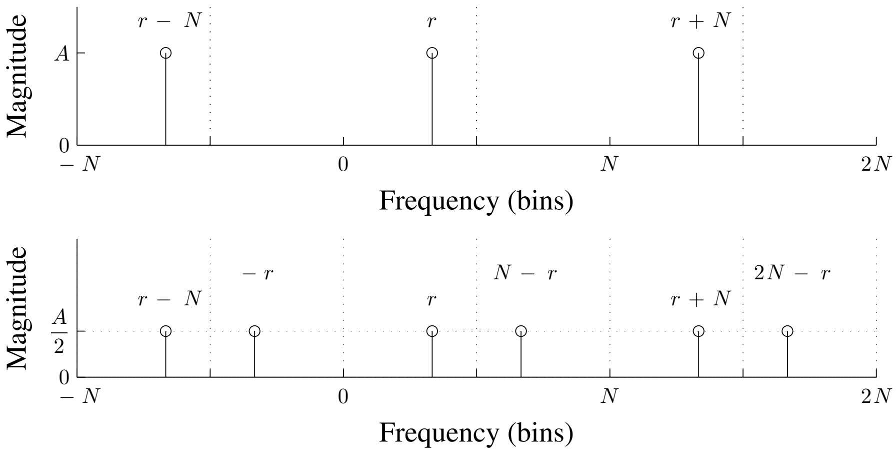

Figure 2.12: Periodicity of the DFT for complex (upper plot) and real (lower plot) signals.

another. For example,

$$\begin{aligned} x[n] &= A\cos\left(r2\pi n/N + \phi\right) \\ &= A\cos\left((r + lN)2\pi n/N + \phi\right) \\ &= \frac{A}{2}\left(e^{j((r+lN)2\pi n/N+\phi)} + e^{-j((r+lN)2\pi n/N+\phi)}\right). \end{aligned} \tag{2.38}$$

In the spectrum of such signal, a peak will be found every $r + lN$ and $lN - r$ bins. Such a spectrum is given in the lower plot in Figure 2.12. There, it is visible that, for real signals, the useful frequency band is $f_s/2$ frequency units wide. One might consider that the frequencies found in a real signal are only positive (in reality, they are both positive and negative, as per Euler's formula), and then set the boundaries of the useful frequency band to $[0, f_s/2]$.

In discrete time, a frequency component of frequency fHz is going to be the cause for the presence an alias every $f + lf_s$Hz. If some continuous component of frequency f outside the $[-f_s/2, f_s/2]$ interval is sampled, this component will be folded back in the aforementioned interval at frequency $f_s/2 + \mathrm{mod}\,(f - f_s/2, f_s/2)$. Upon continuous-time reconstruction, the folded-back component will survive, but due to the rectangular windowing of the spectrum, the original component is going to vanish. It is such frequency-folding artifact that, due to its nature, is commonly called *aliasing*.

2.3.3 Dealing with aliasing

In the design of Analogue-to-Digital Converters (ADC), aliasing is seen as an artifact, and is avoided by the elimination of any continuous-time component of frequency greater than half the sampling rate, ideally without any perceptual reduction the analogue signal's quality. This is achieved through low-pass filtering in the continuous domain, with the alignment of the Nyquist frequency and the filter's cut-off frequency and stop-band on some frequency threshold, often the threshold of audibility of humans. Aliasing in ADC is not of concern here, however, as we are dealing with signals exclusively in discrete time, whether originally issued from some ADC, or obtained directly from synthesis. The only care to give so as to avoid foldback is not to synthesise frequencies or to shift existing components outside the useful band.

In digital signal analysis and processing, aliasing is of a different concern. Mainly, it augments the sidelobe density in a signal's spectrum, as sidelobes extend to infinity, and there is a copy of each peak every $\pm N, \pm 2N, \pm 3N...$ frequency units, *ad infinitum.* The "sidelobe floor" is therefore the combined effect of an infinity of virtual components, even if one component only originally figures in the signal. However, we have seen that the spectral decay trend of windows asymptotically converges towards zero, which allows the setting of a window-dependent width B, beyond which the level of sidelobes can be considered negligible. While remote aliases can therefore be considered not to interfere, the nearest aliases, situated beyond the limits of the useful frequency band $[-N/2, N/2]$, are clearly susceptible of interfering with original partials, just as original partials may interfere between themselves. Not only original components, but original components *and* aliases, must thereby all be spaced by B frequency units. Aliases and negative-frequency halves of real components altogether considered, this spacing is guaranteed if condition (2.9) is respected (Section 2.2), and, for all component of frequency $r_k \in [0, N/2]$,

$$r_k > \frac{B}{2},$$

and

$$N/2 - r_k > \frac{B}{2}.$$

This second condition is intended to keep both “insiders” – original components – and “outsiders” – aliases – at a distance of $B/2$ bins away from the limits of the useful band, thereby ensuring the required B-spacing.

2.3.4 The Fast Fourier Transform

Due to the sampled nature of digital signals, we are, in discrete time, to use summation-based (as opposed to integration-based) Fourier analysis tools.

These are the Discrete-Time Fourier Transform (DTFT), already introduced in (2.35), and the Discrete Fourier Transform (DFT),

$$X_N[b] \triangleq \sum_{n=0}^{N-1} x[n]e^{-jb2\pi n/N}. \tag{2.39}$$

The DTFT cannot be computed due to the infinite-index summation, although it can be useful at an analytical level. Of the four essential Fourier transforms, the only left for us to analyse input discrete signals is therefore the DFT. This section is going to discuss the computational load inherent to the DFT, introduce an efficient algorithm to its calculation, and explore applications of this algorithm beyond the mere computation of DFTs.

As seen in (2.39), the DFT requires, for each spectral sample, the calculation of $x[n]e^{-jb2\pi n/N}$ for N distinct time indices. The inverse DFT,

$$\mathrm{DFT}^{-1}\left\{X_N\right\} \triangleq \sum_{b=0}^{N-1} X_N[b]e^{jb2\pi n/N}, \tag{2.40}$$

in turn requires N spectral samples for the reconstruction of the original sequence. Hence, the computational complexity for the complete frequency-domain description of some N-length discrete-time segment is, in *big-oh* notation, $O\left(N^2\right)$ [7]. In spite of the constant improvement in processing power of computers, real-time applications can suffer from such computational load. Yet there exists an algorithm, or *class* of algorithms [OSB99], essentially based on a divide-and-conquer approach [Ste96], which can compute the N

[7] Big-oh notation is covered in Computational Complexity in standard third-level Computer Science curriculums. The textbook [Sip06] gives a comprehensive introduction on the subject.

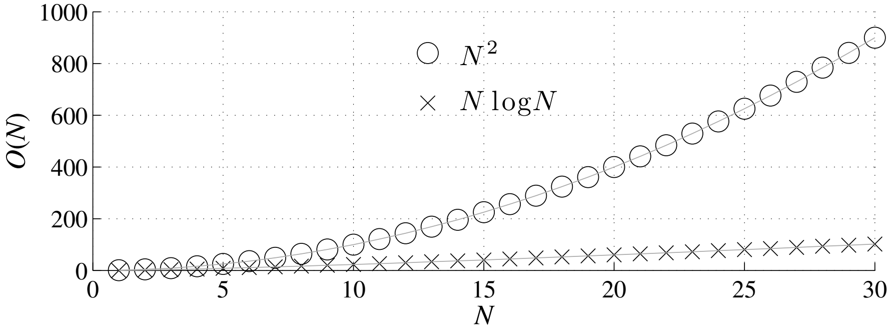

Figure 2.13: N-length DFT (circles) and FFT (crosses) big-oh computational complexity.

spectral samples of the DFT in $O(N \log N)$ time, provided that N is a power of two. This algorithm is the Fast Fourier Transform (FFT). Figure 2.13 gives an impression of the computational savings obtained through the use of the FFT. There, the computational cost of the DFT computed straightforwardly is seen to grow in the manner of a parabola, besides which the $N \log N$ growth seems almost linear. And yet, for the figure to be visually meaningful, N was limited to 30, while analyses of several thousand of samples are commonplace.

Using the *convolution theorem* (Appendix C.2), the FFT can also be used to compute discrete-domain circular convolutions and cross-correlations at lesser cost. Convolution is especially useful in the theory of Linear Time-Invariant (LTI) digital systems, where the output of a filter can be obtained from the convolution of the system's impulse response with the input [8]. Cross-correlation, in turn, is of interest for us as auto-correlation (i.e. the cross-correlation of a signal with itself) can be used to estimate the pitch of

[8] The theory surrounding this statement is well covered in Digital Signal Processing textbooks, e.g. [Ste96, OSB99].

harmonic signals.

Let us first establish the definitions required to make our point. We define the N-length FFT operation onto some signal x as

$$\mathrm{FFT}_N\left\{x\right\} \triangleq X_N,$$

The inverse transform FFT_N^{-1} is such that

$$\mathrm{FFT}_N^{-1}\left\{\mathrm{FFT}_N\left\{x\right\}\right\} = x,$$

and is of equal computational load to the FFT, i.e. $O\left(\mathrm{FFT}_N^{-1}\right) = O(N\log N)$.

The N-length circular convolution and cross-correlation of signals x and w span the interval $[0, N-1]$ and are defined as

$$\left(x \circledast w\right)_N[n] \triangleq \sum_{m=0}^{N-1} x[m]w[\mathrm{mod}(n-m, N)] \tag{2.41}$$

and

$$\left(x \star w\right)_N[n] \triangleq \sum_{m=0}^{N-1} x^*[m]w[\mathrm{mod}(n-m, N)], \tag{2.42}$$

respectively.[9]

The convolution theorem states that the Fourier transform of the convolution/product of two signals equals the (scaled) product/convolution of the Fourier transform of each signal. This theorem is transposable to all four

[9] The asterisk seen in (2.42) is the *complex conjugation* operator, i.e. $\left(e^{j\phi}\right)^* = e^{-j\phi}$.

Fourier transforms, and in the present case implies that

$$(x \circledast w)_N = \frac{1}{N}\mathrm{FFT}_N^{-1}\left\{\mathrm{FFT}_N\{x\} \cdot \mathrm{FFT}_N\{w\}\right\}, \tag{2.43}$$

and, applied to cross-correlation, that

$$(x \star w)_N = \frac{1}{N}\mathrm{FFT}_N^{-1}\left\{\mathrm{FFT}_N^{*}\{x\} \cdot \mathrm{FFT}_N\{w\}\right\}^{10}. \tag{2.44}$$

In terms of cost, the direct computation of some N-length convolution (2.41) or cross-correlation (2.42) takes, similarly to the DFT case, $O\left(N^2\right)$ time. However, each of the FFT-based equivalents, (2.43) and (2.44), takes

$$\begin{aligned} O\left(\frac{1}{N}\mathrm{FFT}_N^{-1}\left\{\mathrm{FFT}_N \cdot \mathrm{FFT}_N\right\}\right) &= O(N + 3N\log N) \\ &= O(N\log N) \end{aligned}$$

time only, following the conventions of big-oh notation [Sip06]. The gain in time is shown here to be of the same order of magnitude as that seen in Figure 2.13.

Using the FFT is, in a variety of contexts, a substantial gain of time. Yet at first glance, the constraint of setting the analysis length N to some power of two may seem inconvenient. Indeed, to keep the quality of near-stationary and/or finite-length signal analysis at an optimum, it is essential to have complete command over analysis length. To combine both arbitrary analysis length and FFT efficiency, the solution is to use zero-padding.

[10] Formal statement and proof of the convolution theorem and its extensions can be found in Appendix C.

2.3.5 Zero-padding

The purpose of this section is to examine the effect of zero-padding a signal on the signal's DFT. In DSP terminology, zero-padding a discrete signal means concatenating a finite-length sequence of zeros to the tail of the signal. Zero-padding can be emulated with the use in the DFT of a length-N rectangular window, and setting the DFT length to some power-of-two greater than N, $M = 2^i$, $i \in \mathbb{Z}$, $M \geq N$. The M-length FFT of x_N then becomes

$$\begin{aligned} X_N^M[b] &= \sum_{n=0}^{M-1} x[n] \, {}_{\varsigma}\overset{\mathrm{r}}{w}\,(n) e^{-jb2\pi n/M} \\ &= \sum_{n=0}^{N-1} x[n] e^{-jb2\pi n/M}. \end{aligned} \tag{2.45}$$

These equalities show that the M-length DFT (or FFT) of a signal of length N padded with $M-N$ zeros is similar to a straightforward N-length DFT, to the difference that it is evaluated at frequency multiples of $2\pi/M$ instead of $2\pi/N$, the "natural frequency" of the analysis. A important consequence is that, while the number of discrete frequency indices increases from N to M, the width of "frequency events" increases proportionally. Concretely, what this means is that the width of the spectral lobe of a window increases by a factor of M/N, as does the number of samples per frequency period. For instance, consider ${}_{\varsigma}W_N^M$, the length-M DFT spectrum of a window ${}_{\varsigma}w$ of regular length N. It follows from Equation (2.45) that

$${}_{\varsigma}W_N^M\left[b\frac{M}{N}\right] = {}_{\varsigma}W_N[b].$$

For instance, say ${}_{\varsigma}W_N$ has its first zero-crossing at bin b_0, and hence has a main lobe $2b_0$ bins wide. In the spectrum ${}_{\varsigma}W_N^M$ of the zero-padded window, the first zero is going to be found at index b_0M/N, as ${}_{\varsigma}W_N\,[b_0] = {}_{\varsigma}W_N^M\,[b_0M/N]$. Whether in the original spectrum or the "zero-padded spectrum", the ratio *useful band/lobe width* thereby remains the same, i.e. $N/2b_0$.

In the light of this discussion, we see that the setting of the length of an FFT to M does not modify the spectrum's attributes stemming from the length N of the window. In our spectral notation, e.g. X_N^M, we can say that the subscript denotes the periodicity of the frequency-domain signal, while the superscript is a direct indicator of the "analysis resolution". Considering the meaning we give to this term, it is therefore more appropriate to see the effect of zero-padding as a reading of the spectrum between the multiples of the analysis frequency $2\pi/N$, in other words, as a means of interpolation [IS87], rather than an increase in resolution.

This type of interpolation is nevertheless particularly good, as it is faithful to the actual continuous-frequency contour yielded by the corresponding DTFT. To increase the interpolation even more, M may be derived as $2^{\lceil \log_2 N \rceil + i}$, for some non-zero integer i, instead of merely satisfying the power-of-two constraint by setting the FFT length to $M = 2^{\lceil \log_2 N \rceil}$. In Section 2.4, where the problem of "extracting" the parameters of a partial from its spectral lobe is approached, this might be found to be quite useful.

2.4 Determination of a sinusoidal model suitable for short-time partial cancelation

A wide variety of signals consist of sinusoids, with slow-changing parameters, mixed up with some level of noise. Estimating the frequency, phase and amplitude parameters of the sinusoids making up such signals is the basis to a wide variety of applications, among which, the very purpose of this thesis: string extraction.

Because of its special status, the literature on the subject of *sinusoidal analysis* is voluminous. Among the varied approaches therein proposed, we are to pick that which is best suited to our purpose. Let us recall that we are favouring a phase-vocoder-based approach to string extraction. This implies a Fourier-based approach to sinusoidal analysis. Surveys of such methods for signals that can be assumed stationary – constant-amplitude and constant-frequency components – are given in [KM02, HM03]. Of the solutions proposed, the most popular is probably the quadratic-fit method, because it is very intuitive, simple to implement, and has the potential of yielding readily accurate frequency, magnitude and phase estimates [AI04]. Another is the phase difference method, a little bit less straightforward to implement, but yet a simple way of surpassing the time/frequency resolution trade-off inherent to Fourier analysis [ZÖ2, p. 337]. There, the phase evolution between successive spectra is kept record of and used to refine the rounded frequency corresponding to a local magnitude peak in the spectrum. In this sense, it is not so far off from the Complex Spectral Phase Evolution method, which, although quoted less often in the literature, is arguably a simpler, more accurate and secure way of exploiting the evolution of phase over time [SG06]. Another method worthwhile mentioning is the method that uses the spectrum of the time derivative of the signal along with the traditional spectrum to refine the frequency estimate [Mar98]. Although the improvement in frequency estimate there is real, the method is, in practice, relatively complex and computationally costly to implement.

All these techniques, as designed originally, assume that the frequency and amplitude of the sinusoids making up the signal is approximately constant over the analysis period. Much research has been put in recent years for the generalisation of such techniques to the evaluation of non-stationary signals. In our choice for a technique, it is an important question to consider whether string signals, over the period of the analysis, can indeed be considered static, and if not, in which aspects – frequency, amplitude, or both? The purpose of this section is to answer this question. To do so, we are going, in Section 2.4.1, to simulate the use of a constant-amplitude model for the cancelation of an exponential-amplitude partial – which is indeed the nature of our string partials, as seen in Chapter 1. Likewise, a constant-frequency model is going to be used against a linear-frequency model in Section 2.4.2. (We saw in Section 1.2.6 that, in the presence of tension modulation, the frequency of the partials followed the non-linear model of Equation (1.50). Yet it is deemed that a first-order polynomial approximation, at the time scale of an analysis window, is in any case largely exact.) The conclusion to these simulations will tell us what parameters it matters to estimate, and thereby help us choose an appropriate analysis method.

2.4.1 Canceling exponential-amplitude partials with sinusoids of constant amplitude

The standard partial cancelation process consists of synthesising a sinusoid identical to the target sinusoid, and subtracting it thereafter. In this section, the focus is on amplitude, and the complexity of the mathematical notation can be reduced if we assume that the target sinusoid is perfectly measured in phase, frequency and amplitude at the sample index of the analysis, only the synthesised tone has constant amplitude, while the actual measured tone has exponential amplitude. We therefore omit the phase and frequency terms in our expressions, to concentrate on *levels*.

We have seen in Chapter 1 that the decay rate of the harmonic series, γ_k, was a second-order polynomial in k (1.27). At higher harmonic indices, the decay rate is therefore much greater, and the cancelation of partials with some constant-amplitude wave, worse. Yet, it must be realised that higher partials are initially lesser in amplitude, and thereby require to be attenuated less before becoming inaudible. We have seen in our physical analysis that, for hit strings, the level of a harmonic was inversely proportional to its number (1.18), and for plucked strings, it was *at best* inversely proportional to its number (1.13). In a normalised digital tone, the maximum amplitude of a partial is 1, corresponding to a Full-Scale decibel ($\mathrm{dB}_{\mathrm{FS}}$) level of 0. On the basis of all this, we can say that the optimal level of a string's k^{th} harmonic is

$$\kappa(t) = \frac{1}{k} e^{-\gamma_k t}.$$

To simplify the expressions to come, we consider that κ is well approximated with a first-order polynomial for a short period of time. Approximated around $t = t_u$, the level thus becomes

$$\kappa(t) \approx -\frac{\gamma_k}{k} e^{-\gamma_k t_u}(t - t_u) + \frac{1}{k} e^{-\gamma_k t_u}, \quad \text{for small } |t - t_u|.$$

Say we evaluate the amplitude of the partial at time t_u exactly, and cancel around this time index the component with a constant-amplitude wave. The level of the canceled wave, $\lambda(t)$, comes out as

$$\lambda(t) = -\frac{\gamma_k}{k} e^{-\gamma_k t_u}(t - t_u).$$

Next: set λ to some fixed value corresponding to a minimal tolerated attenuation in decibels, $\lambda = 10^{\lambda_{\text{dB}}/20}$; substitute t_{dB}, the time it takes from the analysis time t_u to reach λ, in place of $t - t_u$; and solve for t_{dB}. This yields

$$t_{\text{dB}} = \frac{k}{\gamma_k} 10^{\lambda_{\text{dB}}/20 + \gamma_k t_u / \log 10}. \tag{2.46}$$

Three comments can be made on Equation (2.46):

- The time it takes for the cancelation to degrade beyond what can be tolerated is an exponential function in the decibel attenuation threshold λ_{dB}. Roughly speaking, this means that, if we raised the level of tolerance, the cancelation process will be satisfying over a longer period of time. Lower levels of tolerance, consistently, induce shorter "full cancelation times".

- t_{dB} is also exponential in t_u, the place where the cancelation occurs. Indeed, the gradient of the exponential curve is visibly smaller later into the sound. Short-term stationary-amplitude cancelations might therefore not be satisfying near the excitation time, but improve as time passes by, later into the tone.

- Finally, the relation between t_{dB} and the partial index k is more complex. At time 0, it would be proportional to $k/(b_1 + b_3k^2)$, and so we could say that, for large k, it is inversely proportional to the harmonic index. However, as glimpsed in figures 1.6 and 1.5, the 2^{nd}-order coefficient b_3 is generally around a hundred times smaller than b_1, which implies that this trend will not be visible until some elevated harmonic number is reached.

The time it takes for the cancelation to become less than acceptable is not meaningful on its own. It must be put in relation to some indicator of the period over which the cancelation is going to take place. In our phase vocoder setup, this period is going to be the analysis window length N, itself adjusted in accordance with the fundamental period of the analysed tone. Using some P^{th}-order cosine window, this length is going to be at least $2(P+1)T_0$, for the minimal requirement that spectral main lobes from adjacent harmonics do not leak one into another. As the analysis time index is at the center of the window, t_{dB} only needs to be equal to, or greater than, half this length for the cancelation to be considered successful. The order of the window

P being chosen independently of the note being played, we can express the condition for successful cancelation as

$$t_{\mathrm{dB}}/T_0 \geq P + 1. \tag{2.47}$$

Before we proceed on to the evaluation of constant-amplitude cancelation effectiveness in string tones, some sensible maximal level of inaudibility λ_{dB} and cancelation time index t_u must be chosen. A reasonable decibel level of tolerance for the canceled partials is -60dB$_{\mathrm{FS}}$, a referential attenuation level in other audio engineering contexts, that corresponds to a linear level of one thousandth. It may be argued that $-60\mathrm{dB}_{\mathrm{FS}}$ sinusoidal components are still slightly audible when they occur in some frequency band where the human hearing is the most sensitive. However, we have been rather harsh in other places, which leaves us a bit of headroom here. For instance, we have neglected, for the sake of simple expressions, the attenuation due to the window, important near the edges of the analysis. Also, an initial optimal level for the harmonics of $1/k$ at time 0 is *really* optimal: the harmonics add up together, and for the final waveform not to clip the harmonics must generally be scaled down. Finally, we choose for simplicity to set the analysis time t_u to zero, which is the time when the gradient of the amplitude envelopes is the greatest, making it harder for constant-amplitude cancelation. In addition, it is also the time when the partials are the greatest in amplitude, and hence when the attenuation constraint is the most severe. All things considered, we can confidently say that the respect of condition (2.47) at the onset of the sound and for a required post-cancelation wave level λ_{dB} of $-60\mathrm{dB}_{\mathrm{FS}}$ guarantees a successful cancelation process. Now whether string tones satisfy this condition or not remains to be seen.

In spite of setting t_u and λ_{dB} to some fixed value, the level time t_{dB} still depends on the decay rate of each harmonic, in each note, for each instrument. It was tried to resort to the γ_k model of equation (1.27), but the fitting of this model does not yield satisfying b_1 and b_3 coefficients across all tones of all instruments. It was therefore chosen to use the measured decay time of each harmonic individually. Then t_{dB} could be evaluated as $\frac{k}{\gamma_k}10^{-3}$, which is Equation (2.46), after substitution of 0 for t_u, and of -60 for λ_{dB}. Finally, t_{dB} had to be divided by the fundamental period T_0 of the corresponding tone to evaluate if the condition for successful cancelation (2.47) is met.

The experiment was run over four string instruments of contrasting character: an American Fender Stratocaster electric guitar, a grand piano, a double bass plucked in jazz style, and a harpsichord.[11] For each instrument, several notes were chosen, evenly spaced across the instrument's range. For each harmonic, the decay rate was measured, and t_{dB} thereby derived. The measurements are shown in Figure 2.14. These values, except for some outsiders and local trends, are seen to be relatively homogenous across each instrument's plane. The median value, highlighted with a black stem, can thereby be considered as a meaningful indicator of whether a stationary-amplitude model is sufficient for the string extraction on a given instrument. The median values are listed in Table 2.3.

These results can be clearly put in relation to the "brightness" of the

[11] The samples were issue from Yamaha sample CDs, A5000 Professional Studio Library

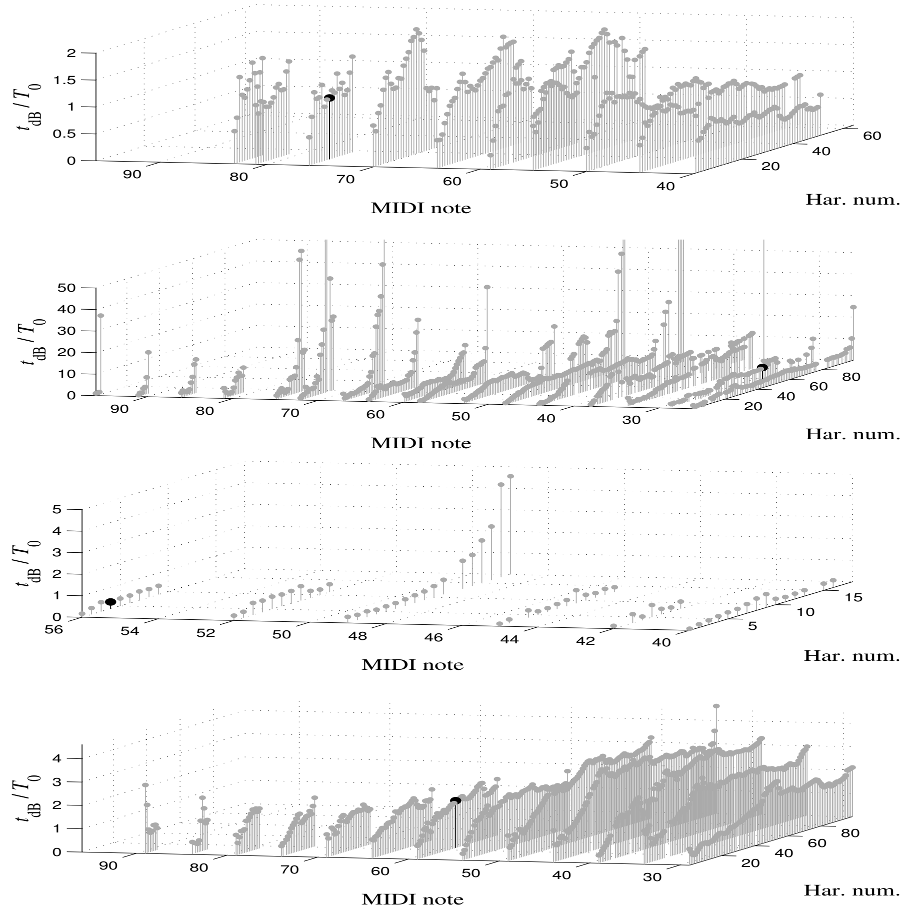

Figure 2.14: "Full-cancelation time", t_{dB}, over the note's fundamental period, T_0, measured for each harmonic ("Har. num." axis) of a number of notes ("MIDI note" axis) and four instruments, from top to bottom: electric guitar, grand piano, double bass (plucked) and harpsichord. The rule for successful cancelation of a partial with exponential envelope approximated with a constant-amplitude sinusoid is that $t_{\text{dB}}/T_0 > P+1$, P being the order of the window used. The median values t_{dB}/T_0 for each instrument, in black, are shown in Table 2.3.

instruments. Bright instruments are instruments where the partials decay slowly, and hence where cancelation based on static-amplitude synthetic tones is most effective. The grand piano is indeed the instrument whose

	Stratocaster	Steinway	Double Bass	Harpsichord
median(t_{dB}/T_0)	1.1393	5.4936	0.3132	2.0219

Table 2.3: Median values of the ratio t_{dB}/T_0 across all harmonics of all notes, for four instruments of contrasting character.

tones, over most of its range, exhibit the greatest sustain. The harpsichord is also a relatively bright instrument, as well as the electric guitar. The double bass, on the other hand, is known for its heavily dampened, “round” and muted character, with very short-lived tones. The table is consistent with this. It indicates that canceling the partials of a grand piano can be done with a constant-amplitude sinusoid using an analysis window of large order, up to 4 if we put the median value in relation to the condition of successful cancelation, (2.47). It also says that constant-amplitude cancelation can be used for a harpsichord if the window is of order 1, i.e. a Hann or Hamming window. However, only a rectangular window seems to be admissible for the Stratocaster, and we know that this is not an option due to the high sidelobe level of this window. As for the double bass, the decay rate of partials is such that the cancelation must be done with synthetic sinusoids of linear, or, better, exponential amplitude.

According to these readings, a constant-amplitude sinusoidal analysis technique is sufficient for the string extraction of the brightest string instruments. However, our ambition is to devise a string extraction technique applicable to all plucked- and hit-string instruments, ambition for which taking the exponential decay of the partials into account is necessary. Also, the minimal length N to resolve adjacent harmonics can be found to be insufficient, especially for tones which feature distinct phantom partials, the likes of which we saw previously in figures 1.11 and 1.12. The conclusions drawn from this section therefore orientate us toward a sinusoidal analysis that accounts for amplitude changes in the components.

2.4.2 Canceling linear-frequency partials with constant-frequency sinusoids

The process of evaluating the need or not for a linear-frequency model in the cancelation of the partials is similar to that in the constant-amplitude case. First we express the k^{th} harmonic x_k with an optimal amplitude level of $1/k$, disregarding, for simplicity, the dependence of amplitude on time:

$$x_k(t) = \frac{1}{k}\exp\left[j\left(\int_0^t \omega_k(u)du + \phi\right)\right]. \tag{2.48}$$

The frequency $\omega_k(t)$ is, according to our string model, $k\left(\omega_\Delta e^{-\gamma_\omega t} + \omega_\infty\right)$ (Equation (1.48), Section 1.2.5), if we neglect the contribution of the time-varying inharmonicity coefficient. This is still a relatively complex expression, and to allow for a simple, readable outcome to this development, we approximate the frequency with a first-order polynomial, as we did for amplitude. To be short, we readily choose this approximation to be made about time 0, where the glide is the steepest, to get

$$\omega_k(t) \approx -k\gamma_\omega\omega_\Delta t + k\omega_\infty \quad \text{for small } t. \tag{2.49}$$

Substitution of (2.49) into (2.48) yields

$$x_k(t) = \frac{1}{k}\exp\left[j\left(-\frac{1}{2}k\gamma_\omega\omega_\Delta t^2 + k\omega_\infty t + \phi\right)\right]. \tag{2.50}$$

Again, we assume that, at the time of the analysis, the constant-frequency term $k\omega_\infty$ and the phase ϕ are evaluated perfectly, and are used synthesize a signal

$$x'_k(t) = \frac{1}{k}\exp\left(j\omega_k t + j\phi_k\right)$$

for the cancelation of x_k. We denote again the level of the wave after cancelation with $\lambda(t)$, which is the magnitude of the difference of the two waves, i.e. $\lambda(t) \triangleq |x_k(t) - x'_k(t)|$, reducing to

$$\begin{aligned}\lambda(t) &= \frac{2}{k}\sin\left(\frac{k}{4}\gamma_\omega\omega_\Delta t^2\right)\\ &\approx 2\sin\left(\frac{1}{4}\gamma_\omega\omega_\Delta t^2\right),\end{aligned}$$

where we used the approximation $\sin(kt)/k \approx \sin(t)$, valid for small t.

Setting $\lambda(t)$ to some fixed value λ corresponding, again, to the maximum level acceptable for the canceled wave to remain inaudible, and solving for the corresponding time t_{dB}, we get

$$t_{\mathrm{dB}} = \sqrt{\frac{4}{\gamma_\omega\omega_\Delta}\sin^{-1}\frac{\lambda}{2}}.$$

We see here that the time lapse t_{dB} before the level of the output wave becomes greater than λ is inversely proportional to the square root of both the decay rate of the non-stationary part γ_ω of the frequency ω_Δ, and ω_Δ itself. For most string tones, these are negligible, and there t_{dB} can be considered to be infinitely large. The tones where the frequency glide was found to be the most conspicuous were the Ovation acoustic guitar *fortissimo* tones, whose measurements were shown in the upper plot of Figure 1.9. In Table 2.4, we show the decay rate and time-varying part of the fundamental frequency over a range of notes of our Ovation guitar samples, alongside the t_{dB}/T_0 ratio.

MIDI note	γ_ω	ω_Δ	t_{dB}/T_0
40	2.1272	0.5208	3.5015
44	3.9312	0.4857	3.3604
48	4.3385	0.4313	4.2768
52	1.8438	0.6241	6.8710
64	4.3999	1.6048	5.5476
68	2.3838	1.6135	9.4703

Table 2.4: The decay rate γ_ω and magnitude ω_Δ of the non-constant part of the fundamental frequency of string tones can be used to estimate the ratio t_{dB}/T_0. The measurements shown here are from *fortissimo* tones of an Ovation acoustic guitar, spanning two octaves and a major third. The evaluation of t_{dB}/T_0 indicates here that, in the worst case, a second-order cosine window can be used in a cancelation process using a constant-frequency model, and yet ensure an output wave of less than -60dBFS.

It can readily be concluded from the reading of this table that string tones can be canceled with frequency-stationary signals over a period of time that is at least over six times the fundamental period of the corresponding tones. It is acknowledged that some electric guitar tones can exhibit stronger pitch glides than those seen and heard in acoustic guitar tones. However, notwithstanding that the attenuation of the window itself is omitted here, this concerns a very small minority of tones, for which the added complication and computational load of using sinusoidal analysis techniques for signals of linear frequency is probably not worthwhile.

Section 2.4 has shown that it is necessary to account for the exponential envelope of the plucked- and hit-string partials, even though the time interval over which cancelation takes place is, in our Phase Vocoder approach, no longer than a few times the string's fundamental period. However, a generalisation of the model to first-order frequency was shown unnecessary. It was said that the various options in terms of choice of sinusoidal analysis methods that met this requirement should, at this stage, be examined. Generalisations of the quadratic fit method and the derivative method were respectively introduced in [AS05] and [MD08], but it was found that the Complex Spectral Phase Evolution (CSPE) method introduced in [SG06], simpler than the previous two, could be generalised to returning the decay rate of sinusoidal signals as easily as their frequency. In acknowledgment to the original method, we have named this method the Complex Exponential Phase and Magnitude Evolution (CSPME) method. Because of its simplicity and its accuracy, it can be said in anticipation that this is the method we are going to choose. It is introduced in Section 2.5.1, and how the frequency and exponential amplitude estimates that it returns can be used to estimate the phase and amplitude constants will be shown in Section 2.5.2.

2.5 Parametric estimation of our short-time sinusoidal model

The short-time sinusoidal model we have argued for in Section 2.4 is recapitulated in Figure 2.15. These parameters are ω_r, γ, ϕ and A.

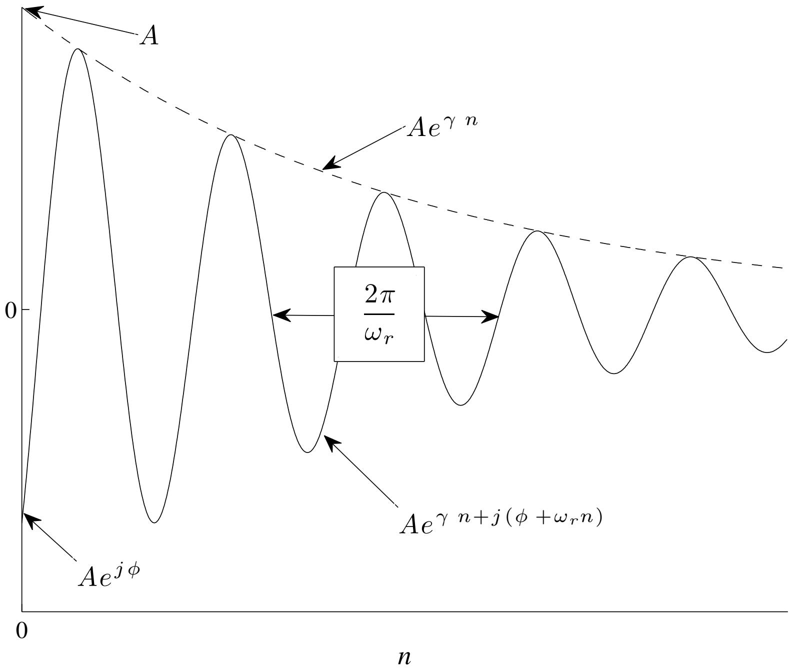

Figure 2.15: The four parameters of our short-time sinusoidal model: frequency (ω_r), growth rate (γ), initial phase (ϕ) and amplitude (A).

As said previously, the CSPME method will yield ω_r and γ, the frequency and exponential amplitude constants. It will be shown thereafter how these are key to the evaluation of the phase and amplitude constants.

2.5.1 The Complex Exponential Phase Magnitude Evolution method.

Throughout Section 2.2, we have been referring to the signal $x(n)$ as a constant-amplitude and frequency complex exponential (c.f. Equation (2.2)). To keep the mathematical symbols simple, we now denote such signal with x', as in

$$x'[n] = Ae^{j\phi}e^{jr2\pi n/N}, \tag{2.51}$$

and x is generalised to an exponential-amplitude signal,

$$x[n] = Ae^{j\phi}e^{jr2\pi n/N}e^{\gamma n}.$$

The two signals are related as follows,

$$x[n] = x'[n]e^{\gamma n}.$$

Now let the signal y be the signal x forward one sample, i.e. $y(n) = x(n+1)$. We see that it can also be related to x by

$$y[n] = e^{\gamma}e^{jr2\pi/N}x[n]. \tag{2.52}$$

Similarly, if we take the DFT of y, we realise that we get that of x, except multiplied by $\exp(\gamma + jr2\pi/N)$, i.e.

$$Y_N[b] = e^{\gamma}e^{jr2\pi/N}X_N[b].$$

Notice that the quotient of X and Y can be used to get to γ and r. We define

$$Z_N[b] \triangleq \log \frac{Y_N[b]}{X_N[b]}, \tag{2.53}$$

get γ as

$$\gamma = \text{real}\{Z_N[b]\}, \tag{2.54}$$

and r,

$$r = \frac{N}{2\pi}\text{imag}\{Z_N[b]\}. \tag{2.55}$$

An equation similar to (2.55) can be found in the paper that introduced the CSPE method [SG06]. The finding of the exponential amplitude coefficient through (2.54) is, on the other hand, the result of our generalisation of the method. As such, this augmented method shall be called Complex Spectral Phase-Magnitude Evolution (CSPME).

It can be seen that, with the CSPME, it is very straightforward to obtain the first-order coefficients of phase and exponential amplitude. Evaluating the corresponding zeroth-order coefficients, however, infers a lengthier development and more elaborate theory.

2.5.2 Amplitude and phase constants

In [SG06], it was shown that the exact frequency of a component could be used to evaluate its exact amplitude as well, with a method equivalent to that already seen in the phase difference method [Z0̈2] and the derivative method [Mar98]. In short, the method consists of "raising", or normalising, the magnitude values of the spectrum to the sought amplitude constant, by dividing the signal's spectrum at the bin nearest to the component's frequency with the spectrum of the analysis window, shifted in frequency by the frequency of the component. Here we propose a generalisation of this approach to the exponential-amplitude case.

It was found that the simplest way to visualise the DFT of x was as the DTFT of the product of three signals: x', which is the same as x but without the exponential amplitude envelope; v, any length-N window (cf. (2.27)), but delayed to the interval $[0, N)$, and periodicised in N, repeating itself *ad infinitum*; and g, an "exponential decay window" of length N. Written formally, this is

$$X_N[b] = \sum_n (x' \cdot g \cdot v)[n] e^{-jb2\pi n/N} \tag{2.56}$$

where

$$g[n] = \begin{cases} e^{\gamma n} & n \in [0, N) \\ 0 & n \notin [0, N-1] \end{cases} \tag{2.57}$$

and

$$v[n] = {}_{\varsigma}w[\operatorname{mod}(n, N)].$$

Now that we have broken down x into such entities, the convolution theorem enables us to express the transform of this product as the circular convolution of the transform of each of these entities, i.e.

$$X_N[b] = \frac{1}{N^2} \left(X' \circledast G \circledast V \right) (\omega_b). \tag{2.58}$$

The sinusoidal parameters A and ϕ are easy to retrieve from Equation (2.58).

First, examine the DTFT of x',

$$\begin{aligned} X'(\omega_b) &= Ae^{j\phi}\sum_n e^{j(r-b)2\pi n/N} \\ &= NAe^{j\phi}\delta(b-r), \end{aligned}$$

then make the definition

$$GV(\omega_b) \triangleq \frac{1}{N}(G \circledast V)(\omega_b), \tag{2.59}$$

derive the convolution of X' and GV,

$$(X' \circledast GV)(\omega_b) = NAe^{j\phi}GV(\omega_b - \omega_r), \tag{2.60}$$

substitute (2.59) and (2.60) into (2.58) and get $Ae^{j\phi}$:

$$\log A + j\phi = \log \frac{X_N[b]}{GV(\omega_b - \omega_r)}. \tag{2.61}$$

In the right-hand side of equation (2.61), we have, in order of appearance, a Discrete Fourier Transform term and a Discrete-Time Fourier transform term. Because $v[n]$ is only non-zero over the interval $[0, N)$, it turns out that the DTFT of the product of s and v is equal to its DFT, i.e.

$$GV(\omega_b) = GV_N(b).$$

In the DFT, however, the argument is an integer, which is not the case for $b - r$. The way to "delay" the DFT by a fractional number of bins r is to introduce in the transform a modulation signal

$$\xi(n) \triangleq e^{jr2\pi n/N}. \tag{2.62}$$

The CSPME formulae (2.55) and (2.54) provide us with r and the exponential amplitude coefficient γ, respectively. For an arbitrary window v, we can now synthesize $(\xi \cdot g \cdot v)[n]$, and hence get the desired spectrum

$$GV_N(b-r) = \mathrm{DFT}\left\{(\xi \cdot g \cdot v)[n]\right\}. \tag{2.63}$$

Without the presence of noise or other partials, and ignoring the digital round-off errors, this procedure yields the exact values of A and ϕ. However, it requires the synthesis of N samples as well as their DFT – or FFT. A computationally cheaper alternative is to approximate the DTFT of $(g \cdot v)[n]$, GV, with the continuous-time Fourier transform of $(g \cdot v)(n)$, $\mathcal{GV}$, establishing

$$GV(\omega_b) \approx \mathcal{GV}(\omega_b),$$

which is true if N, the length of the window, is sufficiently large, and for small b. Similarly to the discrete-time case (2.59), the Fourier transform spectrum $\mathcal{GV}$ is here the convolution of the transform of g, $\mathcal{G}$, with that of v, $\mathcal{V}$. The former is relatively simple,

$$\mathcal{G}(\omega_b) = \frac{N}{N\gamma - jb2\pi}\left(e^{N\gamma - jb2\pi} - 1\right). \tag{2.64}$$

An analytic expression for $\mathcal{V}$ is easily obtained if we use in our analysis a cosine window. Then, v becomes

$$v[n] = \sum_{p=0}^{P} (-1)^p a_p \cos(p2\pi n/N), \tag{2.65}$$

its spectrum,

$$\frac{1}{N}\mathcal{V}(\omega_b) = a_0 \delta(b) + \frac{1}{2}\sum_{p=1}^{P} a_p (-1)^p \left[\delta(b-p) + \delta(b+p)\right]. \tag{2.66}$$

and the scaled convolution of $\mathcal{G}$ with $\mathcal{V}$,

$$\mathcal{GV}(\omega_b) = a_0 \mathcal{G}(\omega_b) + \frac{1}{2}\sum_{p=1}^{P} a_p (-1)^p \left[\mathcal{G}(\omega_b - \omega_p) + \mathcal{G}(\omega_b + \omega_p)\right], \tag{2.67}$$

where $\omega_p = p2\pi/N$. Now the amplitude and phase can be obtained with replacement in (2.61) of GV with $\mathcal{GV}$, with the condition that $|\omega_b - \omega_r|$ is small, and so that b, the integer frequency index used for the estimate, is close to r. To satisfy this condition, let b_0 be the bin closest to r, i.e. $b_0 = \lfloor r \rceil$. Then

$$Ae^{j\phi} \approx \frac{X_N[b_0]}{\mathcal{GV}(2\pi b_0/N - \omega_r)}. \tag{2.68}$$

Equation (2.68) shows that the estimation of the amplitude and phase constants of our target sinusoid requires $\mathcal{GV}$ to be evaluated for one value of its argument only. This is obviously more efficient than having to compute an FFT and store its entire set of M values, even temporarily. As will be seen in Section 3.2.1, $\mathcal{GV}$ can also be used for the frequency-domain cancelation of partials, but then it has to be evaluated for as many bins as there are in the main lobe of the partial. The analytical approach yet remains computationally cheaper than the FFT approach, so long as care is taken to evaluate all these values of $\mathcal{GV}$ in an efficient manner. A computationally efficient approach to the computation of the entire main lobe of $\mathcal{GV}$ was therefore elaborated, and so as to keep these mathematics in one same place, it was decided to present this approach now rather than later.

2.5.3 Computationally optimal calculation of the exponential window spectrum

In our analytical approach to frequency-domain cancelation, $\mathcal{GV}$ has to be evaluated for all bins of the main lobe. To do so efficiently, a matrix $\mathbf{\Omega}$ must be created at initialisation time,

$$\mathbf{\Omega} = -j2\pi \left(\frac{1}{M} \mathbf{b} \mathbf{J}_{1,|\mathbf{p}|} + \frac{1}{N} \mathbf{J}_{|\mathbf{b}|,1} \mathbf{p} \right),$$

where $\mathbf{b}$ is the vector of the zero-centered frequency-domain indices that cover the main lobe of the analytical cosine window of order P,

$$\mathbf{b} = \left[- \left\lceil (P+1)\frac{M}{N} \right\rceil, \left\lceil (P+1)\frac{M}{N} \right\rceil \right]^T,$$

$\mathbf{p}$ is a row-vector of the indices of the spectral components of the cosine window,

$$\mathbf{p} = [-P, P],$$

$|.|$ denotes the cardinality (i.e. the length) of its argument vector, and $\mathbf{J}_{m,n}$ is a matrix of ones, with m rows and n columns.

Another matrix should be created at initialisation time, whose role is going to scale the values of $\mathbf{\Omega}$ with the coefficients of the cosine window. This matrix is

$$\mathbf{A} = \mathbf{J}_{|\mathbf{b}|,1}\mathbf{a},$$

where $\mathbf{a}$ is the following arrangement of the weighted window coefficients found in (2.59),

$$\mathbf{a} = \frac{1}{2}\left[\begin{array}{ccccccccc} (-1)^P a_P & (-1)^{P-1}a_{P-1} & \ldots & -a_1 & 2a_0 & -a_1 & \ldots & (-1)^{P-1}a_{P-1} & (-1)^P a_P \end{array}\right].$$

The run-time operations follow the initialisation-time preliminary steps. For each peak during the cancelation process, upon the detection of the peak bin b_0 and the obtention of the frequency and amplitude modulation CSPME estimates ω_r and γ, the entries of the column vector $\mathbf{GV}$ can be obtained as

$$\mathbf{GV} = \left[\mathbf{A} \cdot \left(e^{N(\mathbf{\Omega}+\zeta)} - 1\right) \div (\mathbf{\Omega} + \zeta)\right] \mathbf{J}_{|\mathbf{p}|,1},$$

where $\zeta = \gamma + j(\omega_r - b_0 2\pi/M)$, and $\cdot$ and $\div$ are used to denote *pointwise* multiplication and division operations.

Recapitulation and additional comments

A summary of the method for the evaluation of all four parameters of our sinusoidal model is given in Figure 2.16. There, the abbreviation "p.d." stands for "peak-detection", and $\hat{r} = \hat{\omega}_r \frac{M}{2\pi}$, where M is the FFT length,

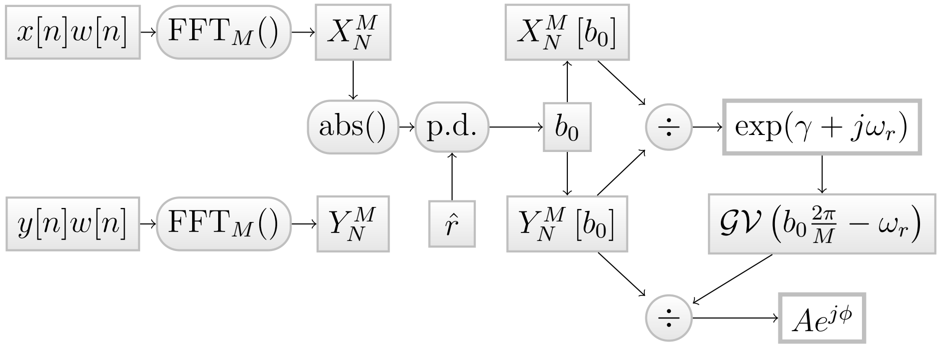

Figure 2.16: Summary of our parametric estimation of first-order amplitude and phase complex exponential. The notation X_N^M denotes an FFT of length M using a cosine window of length N. “p.d” stands for “peak detection”.

and $\hat{\omega}_r$ is an *a priori*, rough estimate of the frequency of the target partial. The use of this estimate is to find the bin b_0, corresponding to the closest magnitude maximum. As for the cosine window, $w[n]$, its coefficients can be picked from Table 2.1, and should be kept for the evaluation of the expression $\mathcal{GV}\left(b_0\frac{2\pi}{M} - \omega_r\right)$, in reference to equations (2.66) and (2.67). The possibility of using the FFT spectrum GV_N^M is, of course, also a possibility, with the inconvenience of greater computational cost, and the advantage of enlarging to the window choice to any type of window, not only the cosine windows whose exponential-modulated spectra we have derived.

The core of the CSPME method resides in the obtention of $\exp(\gamma + j\omega_r)$ through the quotient of the DFTs of a signal and a delayed copy of that signal. The resulting equations (2.61) and(2.68) of the subsequent development are similar to equation (9) in [Mar98] and equation (40) in [SG06], except that now, the complex coefficient(s) of the denominator accounts for the exponential amplitude modulation.

In comparison to the quadratic fit method, the CSPE, and by extension, the CSPME, also has other advantages. Where computation is concerned, zero-padding is unnecessary to the CSPE, while the accuracy of the quadratic fit estimate depends directly on this means of “natural interpolation”. The CSPME requires the FFT of two segments, but the zero-padded segment used for the quadratic fit method can be 2, 4 or 8 times longer than the segment used in the CSPME case. Given the $O(N \log N)$ computational complexity of the FFT, the CSPME remains, in this regard, computationally cheaper.

In terms of convenience of use, also, the CSPME, as opposed to the quadratic fit, *does not* require the constraints of zero-phase analysis for the unbiased estimate of ϕ. These constraints include the need for a window that is symmetric about its central sample, and the need for circular shifting subsequent to zero-padding [IS87]. This freedom comes from the fact that all the side effects of windowing and exponential amplitude modulation on phase, inherent to X_N, are annihilated by the spectrum of the amplitude- and frequency-modulated window in the denominator in (2.61) and (2.68). Although the zero-phase spectra constraints are computationally negligible if approached correctly, the CSPME here allows thereby for a more efficient software encoding.

Most importantly, the CSPME responds to our needs, inasmuch as we saw in sections 2.4.1 and 2.4.2 that zeroth- and first-order exponential amplitude and frequency coefficients, no more, no less, were necessary to the successful frame-by-frame cancelation of string partials in a Phase Vocoder setup.

Conclusion

There cannot be successful string extraction without successful sinusoidal analysis, whose study was the topic of this chapter. We began, in Section 2.2, on the topic of windowing, as the good choice of a window and adjustment of the length of the analysis is an important preliminary step, especially for signals that are not truly stationary, and where several frequency components can be found. We put a focus on *cosine windows*, which in the next chapters are going to be our windows of choice, due to their competitive frequency-domain properties, and also their contant-sum property, very desirable in a Phase Vocoder system.

In Section 2.2, the independent variable of time was considered continuous. Section 2.3 extended the discussion to discrete-time signals, explaining the frequency-domain periodicity of DFT spectra and aliasing, introducing the Fast Fourier Transform, formulating the effects of zero-padding. These basics established, the discussion proceeded to Section 2.4 and to the choice of a method for sinusoidal analysis. This choice was to be made according to the aspects of the string partials that needed to be accounted for, which were shown to be, as in “classical” sinusoidal analysis, constant amplitude, constant and linear phase, but also, decay rate. It was deemed that the most suitable method was therefore the the Complex Spectral Phase-Magnitude Evolution method, generalisation of the Complex Spectral Phase Evolution method.

Chapter 3

A Phase-Vocoder approach to string extraction

In Chapter 1, we determined a string model that enlightened us as to what should be looked for and canceled for a successful string extraction. In Chapter 2, we saw how the use of analytical windows in Fourier analysis could help reduce spectral leakage, we studied the resulting frequency-domain representation of the partials, and, on the basis of our knowledge of the string model, we picked an existing measurement method, the Complex Spectral Phase Evolution method, that we generalised to exponential-amplitude signals to suit the string model exactly.

In this new chapter, we are ready to approach the higher-level structure of the string extraction process. In Section 3.1, we are going to study the temporal structure of the process, which is to conciliate the constraints of analysis length, window type, and process transparency, a concept explained in Section 3.1.2. The essential part of this first section will therefore be the presentation of the Phase Vocoder (PV) scheme, simplified from the general formulation [Por81] to constant analysis-synthesis rate, and with a syntax in accordance with the rest of this thesis. This presentation will nevertheless be preceded by the description of a simple, common pitch detection algorithm based on autocorrelation. This preliminary step is necessary to adjust the window length of the Phase Vocoder. Also, this section on the time-domain structure of the string extraction process will conclude on a "user guide" subsection, where the "granulation" step of the Phase Vocoder (also introduced in Section 3.1.2) is described for finite-length inputs in a pragmatic way.

Thereafter, we will leave the time-scale of the whole signal to examine how the string partials can be canceled at the level of a single analysis frame. We will first see how to cancel string partials, and second, how to detect all the partials of the string present in the spectrum. Once all partials are detected and canceled, the inverse Fourier transform of the spectrum can be carried out, and the following frame can be dealt with. Once all frames are treated as such, they can be summed back together to produce the final output.

3.1 Temporal structure of the String extraction process

The adjustment of the length of the Fourier analysis window is subject to the well-known time-frequency trade-off [Har98]. As usual in the analysis of harmonic tones, it is necessary, as we saw in Chapter 2, that the window length is at least twice the fundamental period for the partials to be resolved, in the case of rectangular windows, and longer for higher-order cosine windows.

The minimal length of the analysis thereby depends on the fundamental frequency of the tone under analysis. Windows that are longer than the minimal length required have the advantage of reducing the sidelobe interference between partials, and also, in the case of inharmonic string tones, of increasing the potential of the analysis to resolve the near-parallel transverse and phantom partial series. However, it is desirable to keep the window length to a minimum for the following reasons:

First, the model assumed in our Fourier analyses is that of sinusoids with exponential amplitude and constant frequency, which should work well on a long-time scale for the majority of string tones, but which may be inaccurate for tones that undergo tension modulation, e.g. electric guitar tones with strong dynamics. There, the constant-frequency approximation was estimated, in Section 2.4.2, Table 2.4, to be valid over a period spanning between 5 and 10 times the fundamental period.

The second reason regards the onset of the tone. As will be seen in Section 3.2.2, spectral leakage in the analysis frames that overlap with the attack of the tone can worsen dramatically, making the estimation of the string partial parameters less accurate and the subsequent cancelation process less effective. This can be perceptually acceptable if the duration of the time segment where there are windows overlapping with the onset is short enough, so that a minimum of fundamental periods “escape” the cancelation process.

Given these circumstances, it is important to have some preliminary knowledge of the tone's fundamental frequency so as to set the window length optimally.

3.1.1 Preliminary estimate of the Fundamental Frequency

A variety of techniques have been developed over the years, for the estimation of the pitch of a signal, an overview of which is given in [Ger03]. There exists both time- and frequency-domain methods. At first one might think that, given the necessity that we have to go to the frequency domain for the measurement and cancelation of the string partials anyway, some time could be saved by opting for a frequency-domain technique, and running this technique directly on one of the many spectra of our Phase Vocoder sliding analysis. However, because we do not know the fundamental frequency beforehand, there is no guarantee that the arbitrary DFT length used for the pitch estimation will be the optimal window length. Hence, the pitch estimate method should be chosen according to its robustness and computational efficiency only, notwithstanding whether it is a time- or frequency-domain technique.

In the time domain, the autocorrelation of a harmonic signal is such that the fundamental period of the signal can easily be estimated. This is especially true for string tones, as they are relatively well-behaved. A famous algorithm, which uses both autocorrelation and the difference method, is the YIN algorithm [dCK02]. In the frequency domain, an intuitive approach consists of using the ratios of the measured frequency components: for each pair of partials, the lowest integer ratio is found, and this is carried for a number of combinations of two partials. The fundamental frequency can thereafter be inferred [PG79]. We would probably opt for this approach, carried directly onto the PV frames, if the window length we unimportant, but it is. Instead, a simple autocorrelation-based design is possible, even simpler than the complete YIN algorithm, given that string tones are more steady than the majority of otherwise pitched sounds.

Autocorrelation is a special case of cross-correlation. In (2.42), Section 2.3.4, we already defined circular cross-correlation. We re-express this definition here, except cross-correlating the signal x with itself, i.e.

$$xx_N[n] \triangleq \sum_{m=0}^{N-1} x^*[m]x[\mathrm{mod}(n-m,N)], \tag{3.1}$$

x^* being the complex conjugate of x. Similarly to the convolution theorem, xx can be obtained from the inverse Fourier transform of the product of the spectrum of x with its complex conjugate – in other words,

$$xx = \frac{1}{N}\mathrm{FFT}^{-1}\left\{|X_N|^2\right\}. \tag{3.2}$$

Now let us see how this can be used for the estimation of the FF. It turns out that the cross-correlation signal xx shows a trend of decay towards the center of the analysis, as outlined in Figure 3.1. For well-conditioned cases, there is a correlation index N_{FF}, between a minimal period index n_{min} and the centre of the analysis, where the correlation is positive and maximal.

This index coincides with the fundamental period of the input. For this phenomenon to emerge, the length of the input must be at least twice its fundamental period. This can be explained by the fact that $|X_N|^2$ is real, and its inverse DFT is symmetric about $N/2$. N_{FF} must thereby be lesser than $N/2$, or $N > 2N_{\mathrm{FF}}$. On top of that, it is possible to zero-pad the input, so as to attain a power-of-two length, and benefit from the computational efficiency of the FFT. Figure 3.1 shows a well-conditioned case of autocorrelation, with zero-padding.

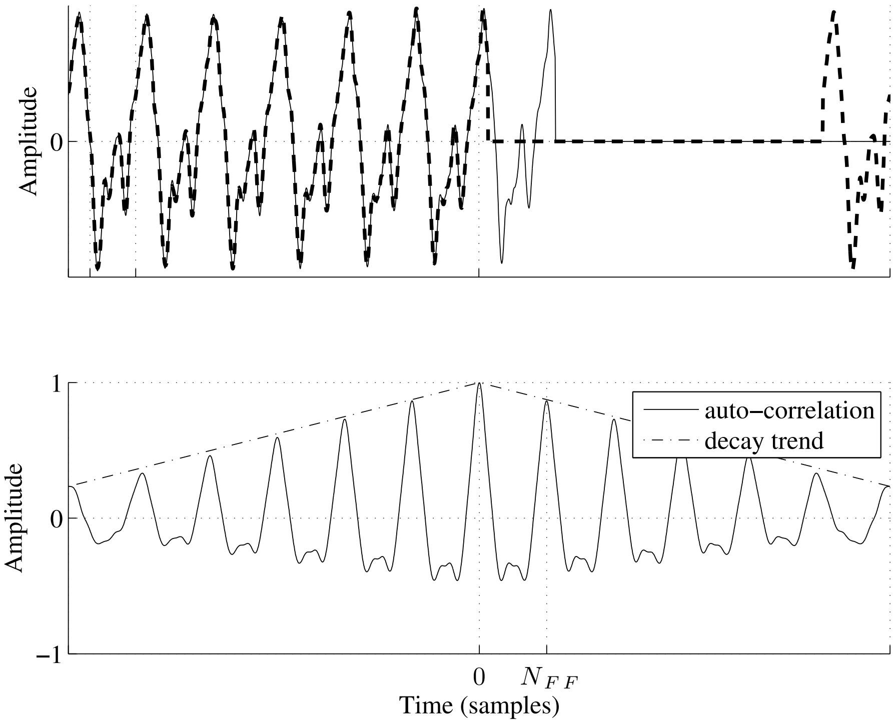

Figure 3.1: Zero-padded signal (upper plot) and its autocorrelation (lower plot). In the upper plot, the periodicity of the waveform is highlighted with the time-shift, in dashed line, of the original waveform by one period. This time shift corresponds to the autocorrelation peak at index N_{FF}.

The power-of-two length can also be achieved by truncating the signal. Complete string tones normally feature many more than two fundamental periods. To keep the computational cost to a minimum, and more importantly, to minimise the time lag of the analysis, it is important that the rough FF estimate is as reactive as possible. Although this thesis strives to facilitate a real-time application, more work is needed here to find a way of adjusting the autocorrelation length to a minimum without preliminary knowledge of the FF. In the meantime, we can set the autocorrelation in accordance with the maximal period length that is susceptible of being found, that of the grand piano's lowest note, an A0 (27.5Hz). Also, the minimal period length $n_{\min}$ is used to prevent the algorithm from returning some unreasonably high pitch, and could be aligned with the period of the highest tone on the grand piano, a C8 (4,186Hz). The grand piano's extreme keys are lower and higher than any other note of any other string instrument ; it is therefore convenient to set our pitch detection interval to its range.

Overcoming the frequency quantisation

In general, the index N_{FF} is only a rounding of the "exact" period, which may be expressed as $N_{\mathrm{FF}} + dN$, were $dN \in \mathbb{R}$. Estimating dN is usually achieved by fitting a quadratic polynomial in the data points $xx_N[N_{\mathrm{FF}} - 1]$, $xx_N[N_{\mathrm{FF}}]$ and $xx_N[N_{\mathrm{FF}}+1]$ [dCK02].[1] A second-order polynomial $q(n) = q_2 n^2 + q_1 n + q_0$ is thereby obtained, whose derivative is equated to 0 so as to obtain the fractional index $N_{\mathrm{FF}}+dN$ of the quadratic function's maximum. This process

[1] This is nothing more than the polynomial fit technique, mentioned repeatedly in Section 2.4, here used in another context.

is illustrated in Figure 3.2.

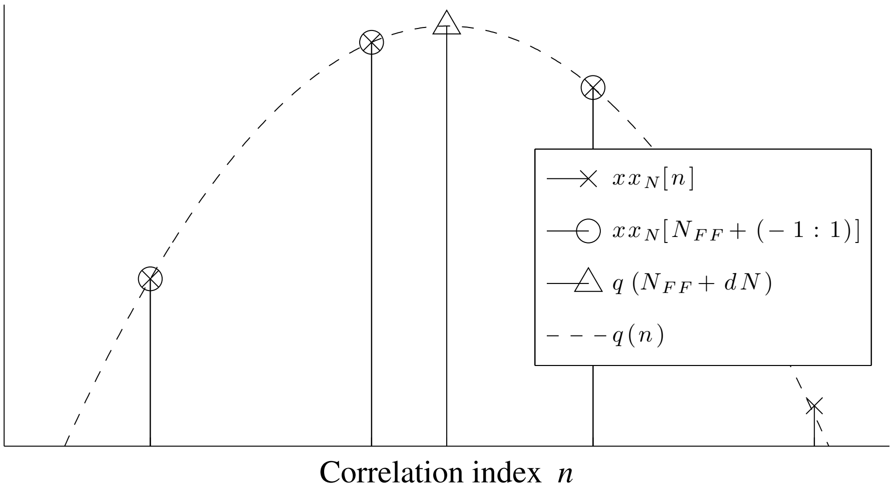

Figure 3.2: Peak index refinement with quadratic fit

Here, the substitution of a quadratic function as a simulation of the continuous shape of the peak seems reasonable. This fractional index refinement improves the period estimate, and hence the fundamental frequency estimate, for a modest computational cost.

From the aproximation of the fundamental period as $N_{\mathrm{FF}} + dN$, a fundamental frequency estimate ω_{AC}, in radians per second, can be obtained as

$$\omega_{\mathrm{AC}} = \frac{2\pi f_s}{N_{\mathrm{FF}} + dN},$$

where f_s is the sampling rate of the time-domain waveform. Such an estimate is deemed reliable enough to be used in the derivation of the parameters of the Phase Vocoder setup.

3.1.2 Constant-rate Phase Vocoder: a formulation

The aim of this section is to describe the Phase Vocoder scheme in an intuitive manner, and at the same time to give a proof of the potential for the "transparency" of the process. The Phase Vocoder is a dynamic audio processing unit, which takes short windowed segments, or *grains* of sound, transforms them one after the other from the time to the frequency domain, to process them there, and bring them back to the time domain afterwards. A transparent Phase Vocoder setup is a setup which, if we omit the processing, should return an output identical to the input, even after the granulation, the frequency-domain forward and inverse transformations and the de-granulation processes. This proof shall be the conclusion of this section.

In Figure 3.3, we illustrate this process with a flowchart. It is easy here to see the symmetry of the process, and it gives an idea of the type of syntax that we are going to use to describe the signal at the various stages of the scheme. It can be seen that the processing, at the U-turning point, is the only place where some changes in the output might come from. If X_N^M and Y_N^M were equal, then so would x and y. At this processing step of the procedure lies the detection and cancelation of the string partials, which results, at the level of the whole tone, in the virtual extraction of the string. Following the flowchart of Figure 3.3, we arrange the formal description of the process on the input $x[n]$ in seven steps, the fourth being the processing step.[2]

[2] Please bear in mind that, in the de-granulation step, it is generally desirable to apply windowing a second time, to avoid discontinuities at the boundaries of the window caused by frequency-domain processing. In the context of string extraction, this additional windowing is superfluous, for a reason that will be made clear soon, and so this additional windowing is left out of the demonstration. The same demonstration of transparency

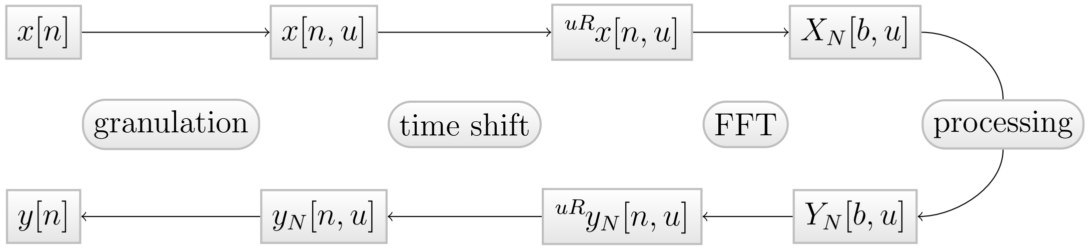

Figure 3.3: Phase-Vocoder process: from the signal at its original time indices, to the frequency domain, and back.

1. Granulation: This term, borrowed from *Granular Synthesis* [Zö2, Roa96], is an metaphor for the process of splitting the input signal in a number of short, windowed "grains" of sound, each of which is, in this context, going to be transformed to the frequency domain and processed there. Let us denote the window index with the integer u, and the hop size in samples, between one window and the next, with R. Then we can express the u^{th} grain as

$$x[n,u] = x[n]{}_{\varsigma}w[n - uR], \tag{3.3}$$

where ${}_{\varsigma}w[n]$ is an arbitrary window, non-zero over the interval $[0, N-1]$.

2. Time-shifting: We are going to make use of the Fast Fourier Transform, whose summation is conventionally operated on the sample interval $[0, M-1]$, for an M-length transform. Our grain $x[n,u]$ must therefore be time-shifted so as to make the first non-zero sample of the window, ${}_{\varsigma}w[0]$, coincide with the start of the analysis. We thus bring the u^{th}

could nevertheless be given with this second round of windowing, too.

grain forward by uR samples, i.e.

$$^{uR}x[n,u] = x[n + uR, u]. \tag{3.4}$$

3. FFT: At this stage, the grain is ready to be transformed to the frequency domain, through fast Fourier transformation

$$X_N^M[b,u] = \sum_{n=0}^{M-1} {}^{uR}x[n,u]e^{-jb2\pi n/M}, \quad b = 0, 1, ..., M-1. \tag{3.5}$$

4. Processing: The processing is left as a black box here, taking in X_N^M, and returning Y_N^M:

$$Y_N^M[b,u] = \text{processing}\left\{X_N^M[b,u]\right\} \tag{3.6}$$

5. IFFT: The time-domain output grains are obtained from the inverse-FFT of the processed spectrum,

$$^{uR}y_N^M[n,u] = \frac{1}{M}\sum_{b=0}^{M-1} Y_N^M[b,u]e^{jb2\pi n/M}. \tag{3.7}$$

6. Time-shifting: The output grain is shifted back to where the input grain originally was:

$$y_N^M[n,u] = {}^{uR}y_N^M[n - uR, u] \tag{3.8}$$

7. De-granulation: Finally, the grains are scaled and summed together, to

yield the final output,

$$y[n] = \frac{1}{\alpha} \sum_{u} y_N^M[n, u], \tag{3.9}$$

where α is a scaling constant.

Now we set to prove that, if no processing takes place in step 4, then the output y equals the input x, or, formally, that $Y_N^M = X_N^M \iff y = x$. To do so, we begin by replacing Y_N^M, in (3.7), by the unprocessed input, X_N^M:

$$\begin{aligned} {}^{uR}y_N^M[n, u] &= \frac{1}{M} \sum_{b=0}^{M-1} X_N^M[b, u] e^{jb2\pi n/M} \\ &= \frac{1}{M} \sum_{b=0}^{M-1} \sum_{m=0}^{M-1} {}^{uR}x[m, u] e^{jb2\pi(n-m)/M} \\ &= \frac{1}{M} \sum_{m=0}^{M-1} {}^{uR}x[m, u] \sum_{b=0}^{M-1} e^{jb2\pi(n-m)/M}. \end{aligned}$$

The summation $\sum_{b=0}^{M-1} e^{jb2\pi(n-m)/N}$ is zero except when $n = m$, where it equals M. So it is easy to get

$${}^{uR}y_N[n, u] = {}^{uR}x[n, u],$$

and, without time shift,

$$y_N[n, u] = x[n, u].$$

We can therefore write that

$$\begin{aligned} y[n] &= \frac{1}{\alpha}\sum_u x[n,u] \\ &= x[n]\frac{1}{\alpha}\sum_u {}_{\varsigma}w[n-uR]. \end{aligned} \tag{3.10}$$

The equality (3.10) only holds if the sum of the square of all the windows equals the constant α, i.e. $y[n] = x[n] \iff \sum_u {}_{\varsigma}w[n-uR] = \alpha$. Cosine windows, as stated in (2.21), Section 2.2.3, have such a potential of summing up to a constant.[3] In that same section, we saw that this coefficient, α, could be found as per (2.26), if R were set to $P+i$, P being the cosine window order, and i, some positive non-zero integer.

As a complement to these explanations and the above formulation, Figure 3.4 pictures the granulation and time-alignment processes. Here, our input signal is only non-zero over a finite interval, as occurs in practice. The point of this illustration is also to realise that, in normal circumstances, only a finite number U of windows is necessary, as any window that does not overlap at all with the non-zero interval of the input shall contribute nothing to the output. However, it also is visible that zero-padding before and after the signal is necessary. To facilitate the implementation of the scheme, we dedicate the next section to the condensed formulation and derivation of the various parameters of the scheme: window length, hop size, window scalar, but also the amount of zero-padding to the left and right.

[3] Even after being raised to some power, had a second round of windowing taken place in the de-granulation process.

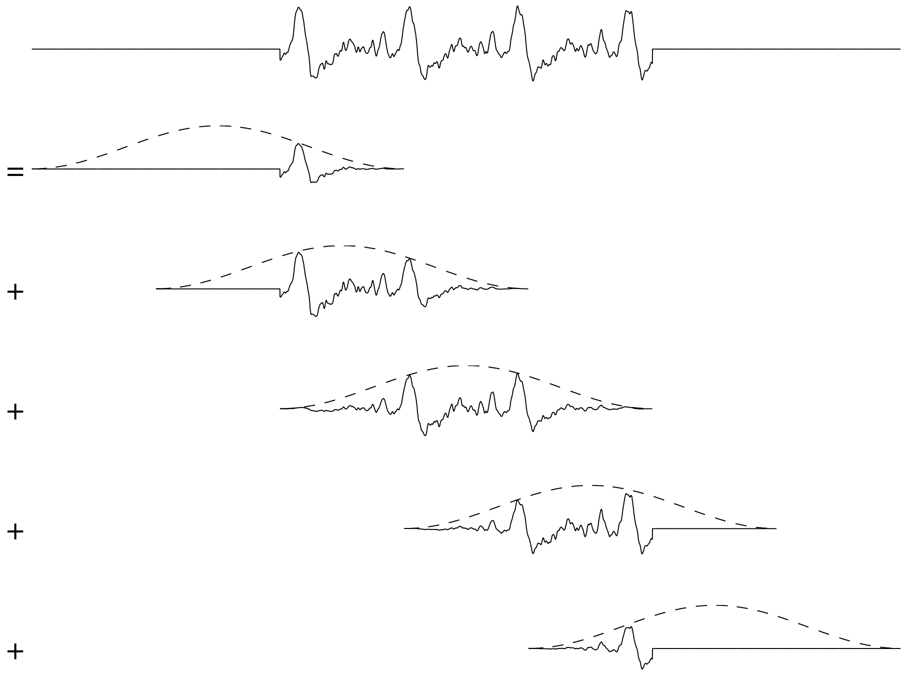

Figure 3.4: Granulation process: the original signal is first zero-padded (top plot) before being granulated (lower plots, c.f. Equation (3.3)), time-aligned to zero (Equation (3.4)) and DFT-analysed (Equation (3.5)).

3.1.3 Preparation of time-domain data

The overlap factor O can be defined in terms of the window order P alone. However, adjusting the window length N and hop size R such that N is O times R, and yet keeping reasonable control over the window length to optimise the time-frequency resolution, is a little bit more delicate. Here we propose a step-by-step description of these derivations.

1. Choose P, the order of the window. Usually, this is 1, 2, or maybe, 3.

2. The minimal overlap to satisfy the constant-sum condition is $O = 2P + 1$. Any integer greater than this works as well.

3. We call the width of the main lobe of the window B, and this width can be derived from P as $B = 2(P + 1)$.

4. Now let us call i the number of main lobes that hold in the frequency band separating each harmonic in the spectrum. For example, if i is one, then the series of lobes of the harmonics will just touch one another, with no spacing between. It was found that a spacing of two or three main lobes between each harmonic generally made the cancelation process easier. The drawback of i larger than 1 is, however, greater time lag, and poorer time resolution to deal with the attack of the tone.

5. The window length could not be set to iBN_{FF}, because it has to be an integer multiple of the overlap. We therefore begin by defining the hop size as $R = \lfloor iBN_{\mathrm{FF}}/O \rceil$.

6. Now, we can give the window length its final value, $N = OR$.

7. Finally, calculate the window scalar as $\alpha = \frac{1}{2}O(a_0^2 + \sum_{p=0}^{P} a_p^2)$, where a_p is the p^{th} coefficient of the chosen cosine window (see Table 2.1).

Now there remains to derive the time index of the start of each grain, as well as the amount of zero-padding needed on either side of the input. The rule of thumb to follow is, for the transparency condition to be respected, that for all sample indices of the input, there are O windows overlapping. In relation to lower plot of Figure 2.7, seen in Section 2.2.3, the totality of the input should lie within the interval where the sum of the windows steadies to the constant α. First of all, we should specify the number U of windows needed for the O-overlap interval to cover the entire input. This is found as

$$U = \left\lceil \frac{N_{\text{in}}}{R} \right\rceil + O - 1,$$

where N_{in} is the length, in samples, of the input. The time indices m_u of the start of each window are therefore obtained as

$$m_u = (u - O + 1)R, \quad u = 0, 1, ..., U - 1.$$

In computer applications, the input comes in as a vector of length N_{in}, undefined for any sample index n outside the interval $[0, N_{\text{in}} - 1]$. To avoid out-of-bound referencing when windowing the input with windows reaching outside this interval, the practical approach is to zero-pad the input to the left and right, with Z_{l} and Z_{r} zeros, respectively. These two values can be computed as

$$\begin{aligned} Z_{\text{l}} &= (O - 1)R, \text{ and} \\ Z_{\text{r}} &= \left(O + \left\lfloor \frac{N_{\text{in}}}{R} \right\rfloor\right) R - N_{\text{in}}. \end{aligned}$$

An illustration of the overall process is given in Figure 3.5.

We have, in Section 3.1.1, obtained a rough estimate of the fundamental frequency of the sound with an autocorrelation method, and in this section, defined all the parameters of our Phase Vocoder scheme, in such a way that we can now go to the frequency domain, find well-conditioned spectral data,

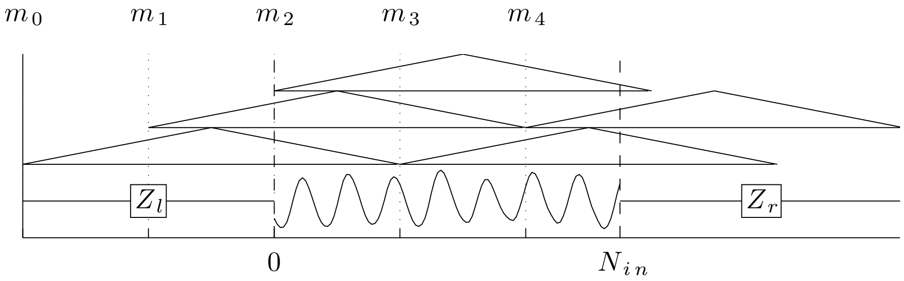

Figure 3.5: Phase Vocoder setup for a finite-length input. A finite number of windows is necessary (here, 5) depending on the window length and overlap. Zero-padding the signal on either end to accommodate windows outside the original interval is a practical approach to the granulation process in computer applications.

and go back to the time domain, transparently if we wish to do so. But to virtually extract the string from the input, some processing must take place: the cancelation of all audible string partials. This is the subject of the rest of this chapter.

3.2 Frame-level partial cancelation

The virtual extraction of the string is, in this thesis, a subtractive process. The idea is to measure the string partials, synthesize them, and subtract this synthesis from the original tone. In a Phase Vocoder setup, this is done at a frame's level, grain after grain.

We look upon the overall string tone, the input x, as the sum of the string, s, and the rest of the instrument, ϵ, (which includes the response of the body to the attack, possibly short-lived resonances, faint sympathetic vibrations of unmuted strings, and environmental noise,)

$$x = s + \epsilon. \tag{3.11}$$

Let us momentarily disregard the phantom partials, for convenience, and express the string s as the sum of K partials s_k,

$$s = \sum_{k=1}^{K} s_k. \tag{3.12}$$

We saw, in Chapter 2, that the string partials could very well, at the time scale of an analysis frame, be approximated as complex exponentials of constant frequency, with an exponential magnitude envelope,

$$s_k[n] = \exp\left[\log A_k + \phi_k + j\left(2\pi r_k/N + \gamma_k\right) n\right]. \tag{3.13}$$

In Section 2.5, we gave the mathematical basis to the measurement of the four parameters A_k, ϕ_k, r_k and γ_k of such a partial, allowing its resynthesis, and by extension, the resynthesis of the whole string s. Subtracting this synthetic version of the string from the original tone isolates ϵ, as per Equation (3.11). The synthesis of the string and its subtraction can be done in the time domain, but we are going to see that, equivalently, it can be done in the frequency domain. This approach turns out to be more straightforward and computationally economical.

3.2.1 Frequency-domain cancelation and its advantages

The Discrete Fourier Transformation of the input x is, as per Equation (3.11), equal to the DFT of the sum of the string and the instrument,

$$\mathrm{DFT}\left\{x\right\} = \mathrm{DFT}\left\{s+\epsilon\right\}. \tag{3.14}$$

On the basis of the linearity of the Fourier transform, and by substitution of (3.12) into (3.14), we see that it actually is equal to the DFT of the instrument, plus the DFTs of each string partial,

$$\mathrm{DFT}\left\{x\right\} = \sum_{k=1}^{K} \mathrm{DFT}\left\{s_k\right\} + \mathrm{DFT}\left\{\epsilon\right\},$$

or

$$X = \sum_{k=1}^{K} S_k + E.$$

Logically, and as seen in Section 2.5, the frequency-domain expression S_k depends on the of the same parameters as the time-domain expression s_k. These parameters can be obtained from the CSPME estimates, and it is therefore just as easy to synthesize S_k, and subtract it from X. Hence we can get ϵ from the inverse transform of the input's spectrum, X, less the string's spectrum,

$$\epsilon = \mathrm{DFT}^{-1}\left\{X - S\right\}.$$

This is the principle of frequency-domain string extraction. Figure 3.6 shows a snapshot of the frequency-domain cancelation process: a peak detection progresses upwards in frequency, and each string partial is canceled, one after the other. We took advantage of this opportunity here to compare the cancelation processes when the exponential amplitude envelope is accounted for (top) and when it is not (bottom). The generalisation of the CSPE method to exponential-amplitude components is shown here to improve the cancelation of the harmonics significantly, especially at high indices, where the exponential decay is strong.

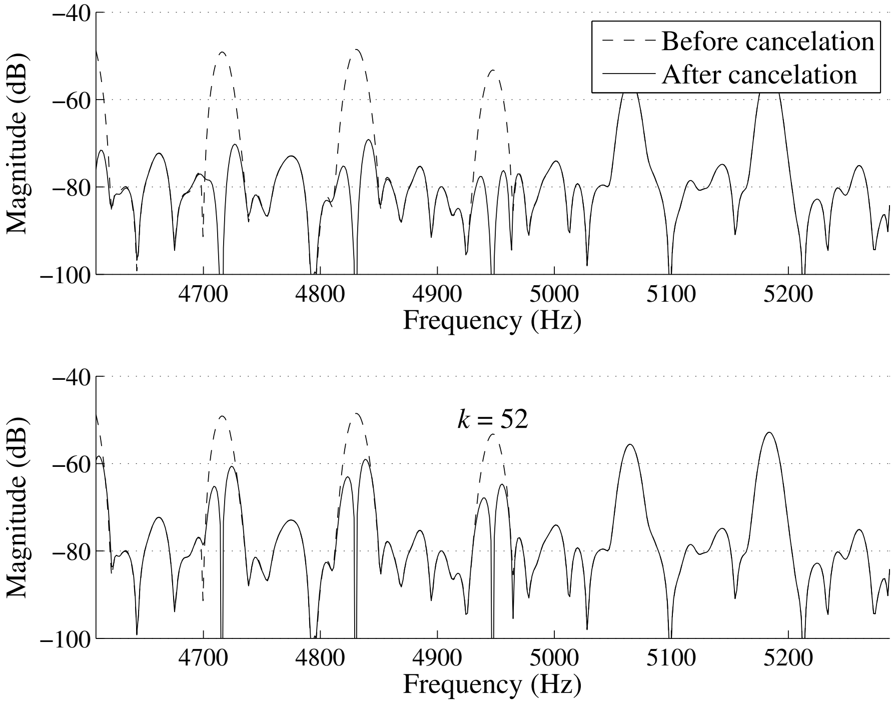

Figure 3.6: Snapshot of the frequency-domain string cancelation process, with the CSPME method introduced in Section 2.5.1 (top), and with the CSPE method as originally introduced in [SG06] (bottom). The partial index k of the peak standing at the centre of the figure is 52.

Time- and frequency-domain cancelation as described here are strictly equivalent. Doing it in the frequency domain, however, is more straightforward and a computationally cheaper approach, especially when phantom partials must be accounted for.

Before we study the process in more detail, we would like to justify here why, in the context of frequency-domain partial cancelation, windowing is only necessary in the granulation step of the Phase Vocoder scheme, and not in the de-granulation step. The synthetic lobes that are subtracted from the original spectrum are the spectra of synthetic partials that are *windowed*. Upon inverse Fourier transformation, the waveforms of these partials will therefore, like the window used for their modeling, tend smoothly towards zero near their left and right ends. The processed grain being the original, smoothened grain minus a sum of these inherently smooth synthetic signals, the usual risk of discontinuities is absent. Hence the frequency-to-time domain additional windowing only results in bringing unnecessary complications and computational load. Indeed, the minimal overlap required when windowing happens once only is $P+1$, against $2P+1$ when it happens twice, P being the order of the cosine window.

Main-lobe-only subtraction

An important saving can take place thanks to the frequency domain energy distribution of windows, which, as we saw in Section 2.2, is concentrated in the main lobe – this is especially true for high-order cosine windows, which have a larger main lobe and lower sidelobes. In mathematical terms, we can write that

$$S_k[b] \approx 0, \quad b \notin \left(r_k - \frac{B}{2}\frac{M}{N}, r_k + \frac{B}{2}\frac{M}{N} \right),$$

where B is the width, in frequency bins, of the main lobe, as seen already in (2.20), Section (2.2). On this basis, the cancelation of the frequency-domain samples within the main lobe only are necessary to completely cancel the corresponding partial. Figures 3.7 and 3.8 support this statement, with, in the former, and illustration of the main-lobe-only cancelation, and in the latter, the corresponding waveform, before and after cancelation. The reader can see that the highest non-zero time sample, after cancelation, is less than -40dB, and is situated, not coincidentally, at the edge of the analysis segment. In the Phase Vocoder scheme, a second windowing occurs after the

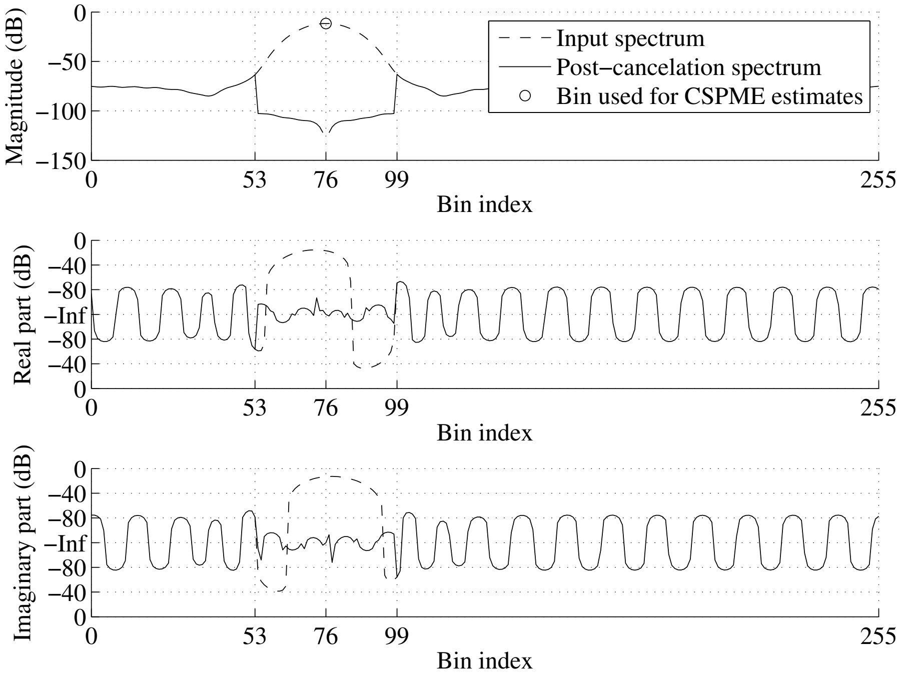

Figure 3.7: Spectrum before (dashed lines) and after (solid lines) cancelation, for minimal-sidelobe cosine window of order 2. Zero-padding was used here for visual purposes, but is normally unnecessary.

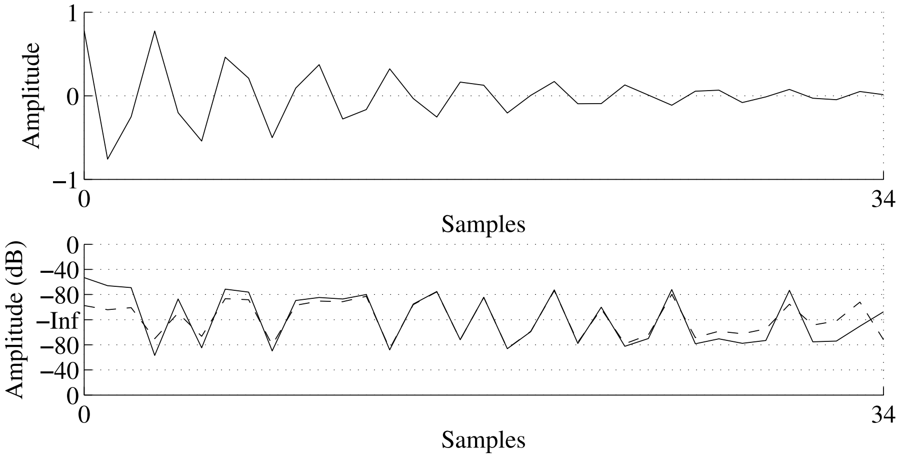

Figure 3.8: Time-domain segment before (top) and after (bottom) frequency-domain cancelation of the main lobe. In the lower plot, the dashed line is the time-domain segment after cancelation *and* windowing.

inverse transformation, to smoothen the edges of the grains post-processing. This additional windowing is not taken into account here, but if it were, the highest non-zero sample of the canceled string partial would be found near the middle of the segment, and be around -80dB.

The window used for this experiment was a minimal-sidelobe window of order two. In terms of computational savings, for a window length N and an FFT length M, the number of samples to synthesise reduces from N down to $\lfloor BM/N \rfloor$. In a typical case, this is a reduction by two to three orders of magnitude.

Dealing with overlapped partials

In strings with non-negligible inharmonicity, the phantom partial – longitudinal vibrations driven by the transverse vibrations – do not coincide with the normal partials throughout the entire series. In general, they are indiscernible below harmonic index 10, and largely discernible, and well-resolved, beyond index 30. In between, however, there is, inevitably, some indices where phantom and transverse partials overlap. In all our frequency analyses so far, we assumed that partials were well resolved enough for cross-partial interferences to be negligible. Now, in places, the overlap may be such that the lowest of the overlapping partials may not appear as a peak anymore, but as a mere protuberance, or *bulge*, which makes the finding of such partials even more delicate.

Informal tests taken by the author showed that so long as a partial, in spite of the interference of another partial, remained a peak and did not subside as a bulge, the CSPME measurements obtained from the frequency data of the main lobe were accurate enough for the cancelation process to work appreciably well. However, no such guarantee can be taken for mere bulges. In the case where two partials overlap, one shows as a peak and the other as a bulge, our solution is to measure the greatest peak and subtract it from the spectrum, and then, measure the lowest – which after the subtraction of the dominating peak has become a bulge – and subtract it in turn. We illustrate the process in Figure 3.9.

In terms of computation, it is evident that it is cheaper to subtract the partials directly in the frequency domain than in the time domain, where the waveform would need to be Fast-Fourier-Transformed after the subtraction of the dominating partial, requiring an extra $N \log N$ operations. It is also worthwhile mentioning that the frequency-domain process is more straight-

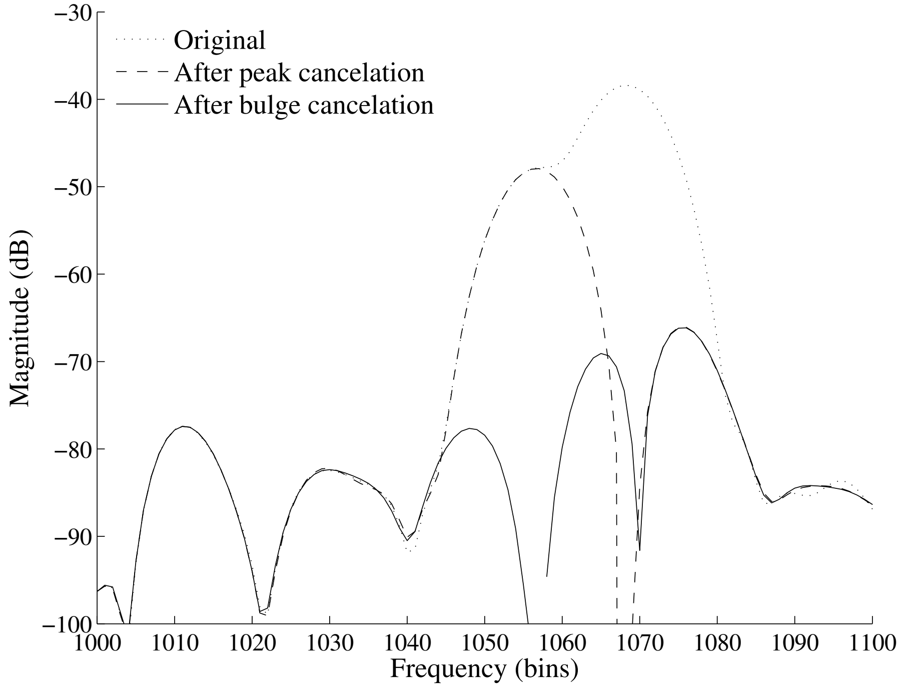

Figure 3.9: Frequency-domain cancelation of overlapping partials. The dominating peak (dotted line) is measured and canceled first. Previously mere bulge, the dominated partial has now become a peak (dashed line), and can be, in turn, measured and canceled.

forward to encode.

We have, in this section, described how the partials were canceled, and the advantages of doing so in the frequency domain as opposed to the time domain. In our spectra, the partials were mostly well resolved, and sometimes, due to the near-parallelism between the main transverse series and the phantom series, overlapped. In both cases, the partials analysed were cosine-windowed components of constant frequency and exponential amplitude, and the corresponding frequency-domain representation was given in

Section 2.5.2. The band-pass filtering properties of the cosine windows made the frequency-domain interference negligible, or at least, in the case of some transverse-phantom overlap, manageable. Yet, we have not been considering so far the “ends” of the input, where the string is attacked, and where it is muted. At these points, the string partial model (3.13) is broken. In the next section, we see what kind of model must now be assumed, and what its frequency-domain properties are.

3.2.2 Dealing with the “ends” of the input

The string partial model $s_k = \exp[\log A_k + \phi_k + j(\omega_k + \gamma_k)n]$ is the steady-state of the model. In reality, such model does not extend indefinitely, but is initiated at the moment of the attack, before which the string vibrations can be assumed to be nil, and is stopped when the string is muted. The moment when the string is excited is more important to us : first, the string's energy decays in an exponential manner, becoming less and less audible, rendering imperfections in the string extraction process synchronously less and less audible too. Second, and most importantly, the response of the instrument's body decays much more rapidly than the string vibrations. So as to appreciate the body response fully, it is therefore critical to cancel the string as early as possible into the sound. For these reasons, and to simplify the discussion, we are, in the following, going to focus on the attack of the string.

Unit-step windowing

Let us say that the string is attacked – plucked or hit – at sample index ν. The input x, generalised so as to account for the time before the attack, is now

$$x[n] = s[n]{}^{\nu}h[n] + \epsilon[n],$$

where the prefix before h denotes a time delay,

$${}^{\nu}h[n] \triangleq h[n-\nu],$$

and h, the *unit step* function (also known as the *heaviside* function, explaining the use of the letter h), is 0 when its argument is lesser than 0, and 1 when its argument is equal to or greater than 0, i.e.

$$h[n] = \begin{cases} 0 & n < 0 \\ 1 & n \geq 0. \end{cases}$$

All the partials s_k making up the string s undergo, of course, the same product. We illustrate one of them in Figure 3.10.

In terms of frequency analysis, the situation is not much complicated, and a development very similar to that used in Section 2.5 can be followed to demonstrate how the four parameters of our model (A, γ, ϕ and r) can be derived.

Let us recall that the estimation of these parameters follows two steps : first, the estimation of the first-order amplitude and phase terms γ and r

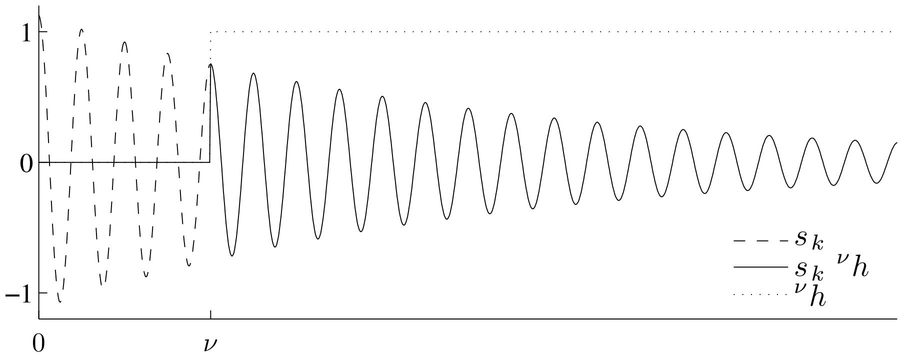

Figure 3.10: Unit-step-windowed string partial, modeling the attack of the tone.

with the CSMPE method, and subsequently, the estimation of the zeroth order terms A and ϕ. To avoid confusions, it must be made very clear that the signal under analysis is looked upon in different ways in each of these two steps. As a starting point to this clarification, let us express the signal to be Fourier-transformed as

$$x[n] = x'[n]g[n]\,{}^{\nu}h[n]\,{}_{\varsigma}w[n], \tag{3.15}$$

product of the constant-amplitude, constant-frequency sinusoid already seen in (2.51),

$$x'[n] = Ae^{j\phi}e^{jr2\pi n/N},$$

the exponential-amplitude envelope

$$g[n] = e^{\gamma n}, \tag{3.16}$$

a cosine window ${}_{\varsigma}w[n]$ and the ν-delayed unit-step window ${}^{\nu}h[n]$.

First-order terms

For the purpose of evaluating the parameters first-order amplitude and phase coefficients γ and r, we "group" the terms in (3.15) as follows,

$$x[n] = s[n]hw[n], \tag{3.17}$$

with $s[n] \overset{\Delta}{=} \exp[\log A + j\phi + (\gamma + j\omega)n]$ being our steady-state string model, and $hw[n] \overset{\Delta}{=} {}^{\nu}h[n]_{\varsigma}w[n]$, a "unit-stepped" cosine window.

Thereon we define

$$y[n] \overset{\Delta}{=} s[n-1]hw[n] \tag{3.18}$$

and state the relation

$$y[n] = x[n]e^{-\gamma - jr2\pi/N},$$

property already seen in (2.52), and which allows, in the frequency domain, the obtention of γ and r with a CSPME approach. It is important to note the consequence of lumping the unit-step function with the window. The one-sample delayed signal y only involves a delay of the steady-state string vibrations s, not of the attack-modeling function h or the window. As a result, even if a "natural" unit-step windowing is already featured in the original sound at the moment of the attack, a synthetic unit-step windowing will have to be forced upon the signal, and slid by one sample to get y, just like the window ${}_{\varsigma}w[n]$ itself has to be slid.

Zeroth-order terms

Contrastingly, the estimation of the zeroth-order amplitude and phase terms, A and ϕ, requires that the signal x is grouped otherwise,

$$x[n] = x'[n]gh[n]v[n], \tag{3.19}$$

where $v[n]$, as in Section 2.5.2, is the unbounded window seen in Equation (2.65),

$$v[n] = \sum_{p=0}^{P}(-1)^p a_p \cos(p2\pi n/N),$$

and the function $gh[n]$ combines in a product the exponential-amplitude rectangular window of Equation (2.57) with the unit-step, i.e.

$$gh[n] = \begin{cases} e^{\gamma n}, & n \in [\nu, N), \\ 0, & n \notin [\nu, N). \end{cases} \tag{3.20}$$

Lumping the unit-step function with the exponential-amplitude rectangular window makes easy the formulation of the Fourier transform of x, $\mathcal{X}(\omega)$. Indeed the spectrum $\mathcal{GH}$ is hardly more complicated than the spectrum $\mathcal{G}$, seen in Equation (2.64), of the exponential-amplitude rectangular window alone :

$$\mathcal{GV}(\omega) = \frac{1}{\gamma - j\omega}\left(e^{N(\gamma-j\omega)} - e^{\nu(\gamma-j\omega)}\right). \tag{3.21}$$

Then the convolution of $\mathcal{GV}$ and the Fourier transform of v, $\mathcal{V}$ (c.f. Equation (2.66)), which we may denote as $\mathcal{GHV}$, is, similarly to (2.67), quite simply

$$\mathcal{GHV}(\omega) = \frac{1}{2}\sum_{p=0}^{P} a_p(-1)^p \left[\mathcal{GV}\left(\omega - p\frac{2\pi}{N}\right) + \mathcal{GV}\left(\omega + p\frac{2\pi}{N}\right)\right], \qquad (3.22)$$

and thereon, following the same reasoning as in Section 2.5.2, we find that

$$Ae^{j\phi} = \frac{X_N[b_0]}{\mathcal{GHV}(2\pi b_0/N - \omega_r)}, \qquad (3.23)$$

where b_0 is the integer closest to the frequency coefficient r, and $\omega_r = r2\pi/N$.

Following the ideas of equations (2.62) and (2.63), an FFT-computed spectrum GHV_N can be used instead of $\mathcal{GHV}$. However, for the lobe of GHV_N to be frequency-aligned with the targeted lobe, the time-domain synthesis of an N-length complex exponential of frequency r is needed. Furthermore, the FFT is generally implemented in libraries in such a way that it is not possible to be returned a specific frequency-domain sample only. Rather, it returns the entire set of M samples, while we have shown in the development above that one such sample only is necessary for the estimation of the constants A and ϕ. Our analytical approach is thereby not only cheaper processing-wise, but also in terms of memory.

Now regarding the *cancelation* of the partial, it was pointed out in Section 2.5.2 that $\lfloor 2(P+1)M/N \rfloor$ frequency-domain samples only, centered on the frequency bin b_0, needed to be synthesized for the cancelation of the measured partial. This was on the basis that frequency-domain energy beyond this interval was negligible. Unfortunately, as can be seen in Figure 3.11, the property that the non-negligible part of the energy is concentrated in the main lobe is lost on attack-overlapping spectra. A greater number of

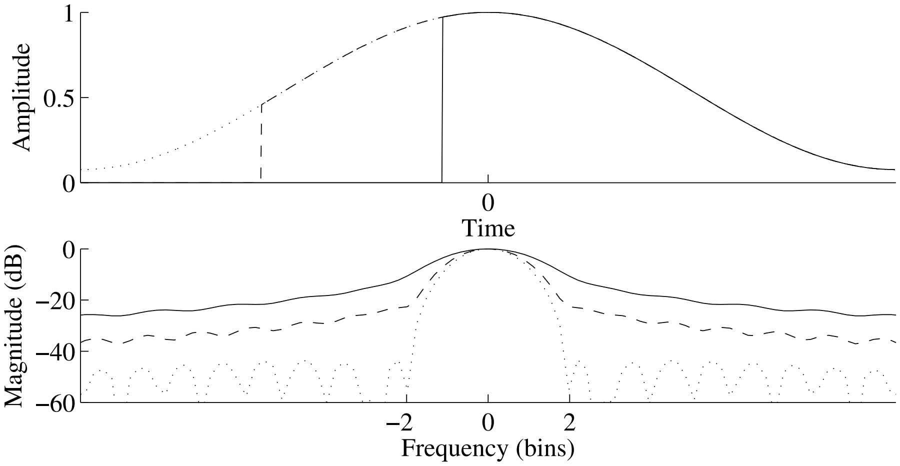

Figure 3.11: Unit-stepped hamming window. The dotted-line, standard window exhibits the expected spectrum, with its minimal, well-dented sidelobes. As the window goes unit-stepped, however (dashed line, solid line), it looses its optimal spectral properties, with much higher sidelobes.

frequency bins should therefore be synthesized for subtraction, and it may then become preferable, because computationally cheaper, to resort to the FFT approach.

The greatest problem regarding this widening of the main lobes nevertheless remains the cross-interference that it causes in the spectrum. The assumption that the energy of a spectral peak at the frequency corresponding to a magnitude local maximum is that of the targeted partial alone can be compromised, and the measurements, biased. Let us note, however, that this bias directly affects the frequency and amplitude modulation measurements only. If these were estimated correctly, and independently on how important the spectral leakage is, the spectrum $\mathcal{GHV}$ would still be synthesized correctly, exact estimates of the amplitude and phase constants ensue, and the partial cancelation would be successful. In future work, efforts should therefore concentrate on improving the exactness of the parameters γ and r.

Identifying the "attack sample"

Another problem that comes with the unit-step model of the attack of the tone is the finding of the first non-zero sample. The unit-step model is a simplification of the attack, which in some cases is fairly faithful, but in others, less so. In Figure 3.12, we show two examples. In the upper plot, an acoustic guitar is plucked with a plectrum, and in a lower plot, we have the example of a viola *pizzicato*. In the case of the acoustic guitar, we identified the attack sample ν manually. The vibrations following this index seem,

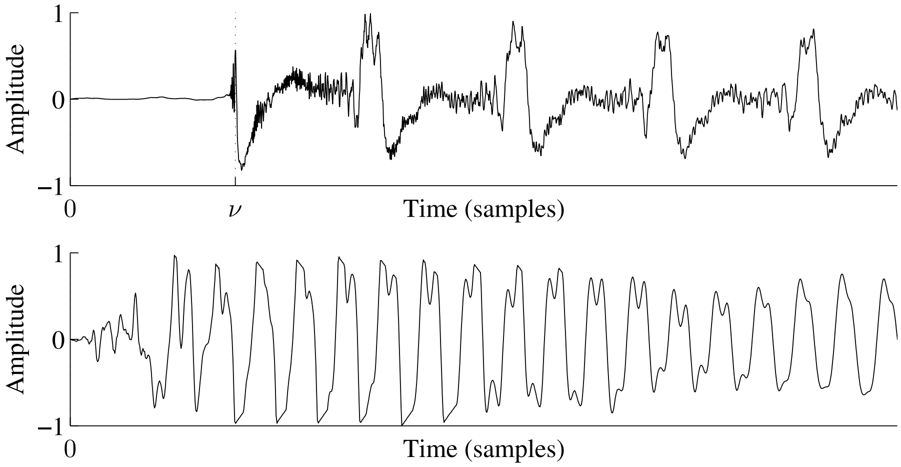

Figure 3.12: Top: Acoustic guitar E2 (MIDI note 40). Bottom: Viola (*pizzicato*) G5 (MIDI note 79).

indeed, to be steady-state. In the case of the viola, however, we observe a finite-length *attack time*, during which the energy builds up. Such a build up is not accounted for in our physical analysis, and we do not possess a model for this part of the sound. It is possible to "force" the input to the unit step model, by setting ν to some sample where the build-up phase is over. The inconvenience here is that, even during the build-up, some periodicity appears. Hence, setting ν too far into the sound would result in discarding a non-negligible segment of the string's vibrations, which would escape the cancelation process, and remain into the part of the sound that is considered to be the instrument's response to the attack. On the other hand, setting ν too early would mean analysing a signal that is, to some extent, in contradiction with the analytical model, thus impairing the exactness of the measurements and the effectiveness of the cancelation.

The research for an algorithm to decide on the attack sample ν is here left for future work. We are just going to mention that, to be transparent, such algorithm should return the first non-zero sample if fed with a synthetic, unit-stepped input, the likes of which was shown in Figure 3.10.

The cancelation of the partials in the frequency domain has been described, its advantages listed, and the difficulties inherent to a Phase Vocoder approach, outlined and discussed. Before being canceled, however, a partial must be found. Also, not all partials must be canceled, but only those pertaining to the string. The detection and identification of the partials of the string within the spectrum of a grain is the object of the next section.

3.3 Detection of string partials

Mainly, it is transverse partials that ought to be found, as they contain the very largest part of the string's energy. The phantom partials can nevertheless not be neglected. Often, the transverse partials, being the most significant, render the phantom partials inaudible, due to the masking effect [Moo04]. However, once the main series is canceled, the phantom partials may, left on their own, become audible.

The difficulty in the detection of phantom partials is that they are so small that they can hardly be identified, unless it is known where to find them. We saw, in Chapter 1, that their frequencies could be arranged in a pseudo-harmonic series, of identical fundamental frequency ω_0 to the main series, only with a quarter of the inharmonicity coefficient β. Only these two parameters are therefore needed to find the phantom partials. However, a precise estimate of these parameters can in practice only be obtained via the measurement of the transverse partials, which are much more conspicuous. We will therefore keep the focus on the main series, suggesting ways of evaluating the Fundamental Frequency and Inharmonicity Coefficient along the way. Then, once these estimates are available, we will be able, by the end of this section, to find the phantom partials.

Let us recall here the expression for the frequency series of the normal partials, already seen in Equation (1.33), Section 1.33:

$$\omega_k = k\omega_0\sqrt{1+\beta k^2}.$$

The partials of the transverse series are large and very recognisable early in the series. Yet, it is convenient to know where to look for some k^{th} partial right away, without having to infer, one way or another, its harmonic number. In this regard, we already have, on the one hand, some estimate ω_{AC} of the fundamental frequency, obtained with the autocorrelation algorithm (Section 3.1.1). On the other hand, the effect of inharmonicity is, in the lowest region of the spectrum, negligible, given that the order of magnitude of β is around 10^{-4}. Therefore,

$$\omega_k \approx k\omega_{\text{AC}} \quad \text{for small } k. \tag{3.24}$$

The first few partials, in general up to the tenth, can therefore easily be picked. For higher indices, the approximation (3.24) does not hold anymore; there, partials may deviate, under the "inharmonicity stretch", by several multiples of the fundamental frequency [HWTL09, HTL10]. Then, more than ever, it can be difficult, once a transverse partial is found, to tell what harmonic number it may be.

3.3.1 Median-Adjustive Trajectories

If the IC were known beforehand, finding the partials would be much easier. Estimating the IC without having to resort to a peak detection is actually a possibility. Galembo and Askenfelt developed several such methods. In [GA94], methods relying on the cepstrum and the Inharmonic Product Spectrum of the string tone are proposed, and in a later paper [GA99], a method using an inharmonic comb filter [GA99] is developed. These methods, however, are relatively expensive computationally, and it has been shown in more recent years that methods based on the measurement of the partials yields more accurate estimates, at lesser computational costs [RLV07, HWTL09]. The author contributed to this problem with an approach based on an analytical expression for the parameters ω_0 and β in terms of a pair of partial frequencies and their harmonic numbers. This method, called Median-Adjustive Trajectories (MAT), is explained now.

In the inharmonic series expression (1.33), we see that the frequency of the k^{th} harmonic depends on the number k and two unknowns, ω_0 and β. Two partials are therefore necessary to solve (1.33) for ω_0 and β, as well as their harmonic numbers:

$$\omega_0 = \sqrt{\frac{l^4\omega_k^2 - k^4\omega_l^2}{l^4k^2 - k^4l^2}}, \tag{3.25}$$

and

$$\beta = -\frac{l^2\omega_k^2 - k^2\omega_l^2}{l^4\omega_k^2 - k^4\omega_l^2}. \tag{3.26}$$

Given the negligible effect of inharmonicity on the first few partials, the MAT method can pick the first two partials based on such an approximation for ω_k seen in (3.24), using for example an autocorrelation-based fundamental frequency estimate. Thereon, a first estimate of the parameters is possible, and can be used for the detection of the third partial. After the third partial is found and measured, there are now three potential combinations of peaks, for three estimates. The median of these collected estimates can be used to progress upwards in frequency.[4] The process is repeated over and over again,

[4] The median is preferred to the average because it is much less sensitive to outliers [HWTL09].

until the last detectable partials, at the higher end of the series, are found.

This is a very robust and secure way of finding the partials of the series, and returns estimates that are extremely close to the true fundamental frequency and inharmonicity coefficient values. The downside of this method is its computational cost. Generally, for K partials found, $(K^2 - K)/2$ combinations are possible, and calculating the series' parameters for all of them is a process of computational complexity $O(K^2)$, notwithstanding the cost relative to the finding of the median value, which requires the sorting of the estimates.

Here, we propose a way of reducing the cost of this approach, by trying to select the combination of partials for which the estimate of the parameters is the least sensitive to error in the estimates of the partial frequencies. For the analysis of this error, it is more convenient to work with the square of the frequency,

$$\Omega \triangleq \omega^2.$$

We denote estimates with a hat, and equate them with their true values plus an error ε. As such, the partial frequency estimates become

$$\hat{\Omega}_k = \Omega_k + \varepsilon_k,$$

and the fundamental frequency and inharmonicity coefficient estimates,

$$\hat{\Omega}_0 = \Omega_0 + \varepsilon_0$$

and

$$\hat{\beta} = \beta + \varepsilon_\beta.$$

Doing all the necessary substitutions, the errors in fundamental frequency and inharmonicity coefficient estimates can be expressed as

$$\varepsilon_0 = \frac{l^4 \varepsilon_k - k^4 \varepsilon_l}{l^4 k^2 - k^4 l^2} \tag{3.27}$$

and

$$\varepsilon_\beta = -\frac{l^4 \beta \varepsilon_k + l^2 \varepsilon_k - k^4 \beta \varepsilon_l - k^2 \varepsilon_l}{l^4(\omega_k^2 + \varepsilon_k) - k^4(\omega_l^2 + \varepsilon_l)}. \tag{3.28}$$

Both errors (3.27) and (3.28) consist of ratios of fourth-order polynomials in k and l. It is not straightforward to find for what combinations of these indices the errors are the least, all the more that ε cannot be known *a priori*. Yet a few clues can be taken by observing the limits of these ratios of polynomials. For one thing, it seems unwise to use a combination of partials that are close together, as

$$\lim_{k \to l} \varepsilon_0 = \lim_{k \to l} \varepsilon_\beta = \infty, \tag{3.29}$$

if we consider that $\lim_{k \to l} \varepsilon_k = \varepsilon_l$. On the other hand, the effect of ε_k becomes negligible if k is very large, as

$$\lim_{k \to \infty} \varepsilon_0 = \frac{\varepsilon_l}{l^2}, \tag{3.30}$$

and

$$\lim_{k\to\infty} \varepsilon_\beta = -\frac{\beta\varepsilon_l}{l^4\omega_0^2\beta + l^2\omega_0 + \varepsilon_l}. \tag{3.31}$$

(3.30) tells us that the error in the fundamental frequency estimate is inversely proportional to the index of the other partial, l^2. It is true that the frequency measurement of partials is more subject to error as the partial number increases, especially because partials of higher indices are lesser in magnitude, and hence more exposed to noise. But this increase in sensitivity is at best linear, which leaves the fundamental frequency estimate error ε_0 still inversely proportional to the partial index l. Regarding the error in the inharmonicity coefficient estimate, ε_β, this decay trend is inversely proportional to l^4, which is a very fast decay.

Considering this statement, it would be tempting to pick the largest possible partial number l too. But then, condition (3.29) also has to be remembered. It therefore seems wisest to pick the largest k possible, but thereafter, to choose l somewhere halfway, for example by choosing $l = \lceil k/2 \rceil$.

An economical reduction of the MAT method is available here. Each new partial estimate avails of one new estimate for both ω_0 and β. To keep the algorithm robust, the estimates are to be collected, and for the detection of the next partial, the median of all available estimates should be used. Now the computational complexity of the peak detection, fundamental frequency and inharmonicity coefficient estimate is linear in K, the number of partials found.

3.3.2 Local linear approximation and Linear Least Squares

In this section, we present yet another method for the detection of the peaks and the evaluation of the fundamental frequency and inharmonicity coefficient. This method relies on the fact that, over a frequency interval of a few peaks only, the partial frequencies can be approximated as a linear function. The Linear Least Square (LLS) fit of a line can therefore be performed in partial frequencies over this restrained interval, and be used to predict the frequency of the next partial. Also, we will show that the coefficients of this linear function can thereafter be used to estimate the actual fundamental frequency and inharmonicity coefficient. The detail of this method follows.

In [FR91], a cubic polynomial was already used to approximate the entire inharmonic series. Here, we reduce this approximation to a first-order Taylor series approximation, valid over an interval of a few partials,

$$\omega(k) \approx \omega'(k_0)(k - k_0) + \omega(k_0), \quad \text{for small } |k - k_0|. \tag{3.32}$$

Let us call K_{fit} the number of partials to use in the linear fit. These partials are partials that are already measured and identified, and should be as close as possible one to the other – adjacent whenever possible – to reduce the interval of the linear approximation. Then, a partial number k_0, around which the Taylor approximation takes place, should be chosen among the selected partials. Now the matrix $\mathbf{K}$ can be constructed,

$$\mathbf{K} = \begin{bmatrix} k_1 - k_0 & 1 \\ k_2 - k_0 & 1 \\ \vdots & \vdots \\ k_{K_\mathrm{q}} - k_0 & 1 \end{bmatrix},$$

as well as the vector $\omega_\mathbf{k}$ of corresponding measured frequencies, where

$$\mathbf{k} = \left[k_1,\ k_2,\ ...,\ k_{K_\mathrm{fit}}\right]^T.$$

The vector containing the polynomial coefficients of (3.32) can now be obtained with a least-square fit,

$$\boldsymbol{\omega} = \begin{bmatrix} \omega'(k_0) \\ \omega(k_0) \end{bmatrix} = \left(\mathbf{K}^T\mathbf{K}\right)^{-1}\mathbf{K}^T\omega_\mathbf{k}. \tag{3.33}$$

It is rather simple to incorporate this fit into a peak detection. Again, the first K_fit partials, at the bottom of the series, can be found with the help of the autocorrelation fundamental frequency estimate. Then, the linear function (3.32) is fitted, and is used to find the next partial. This partial as well as the previous $K_\mathrm{fit} - 1$ is, in turn, used in the linear fit, to get updated $\boldsymbol{\omega}$ coefficients. The process can be repeated until it is estimated that the higher end of the series has been reached.

This method is intuitive, and, provided that the pseudo-harmonic series is well behaved, works well. However it was found that often the harmonic series, even for string tones issued from professional instruments and recorded in optimal conditions, could locally behave in unexpected ways. In Figure 3.13, for example, we use a linear fit over five partials (circled peaks). In

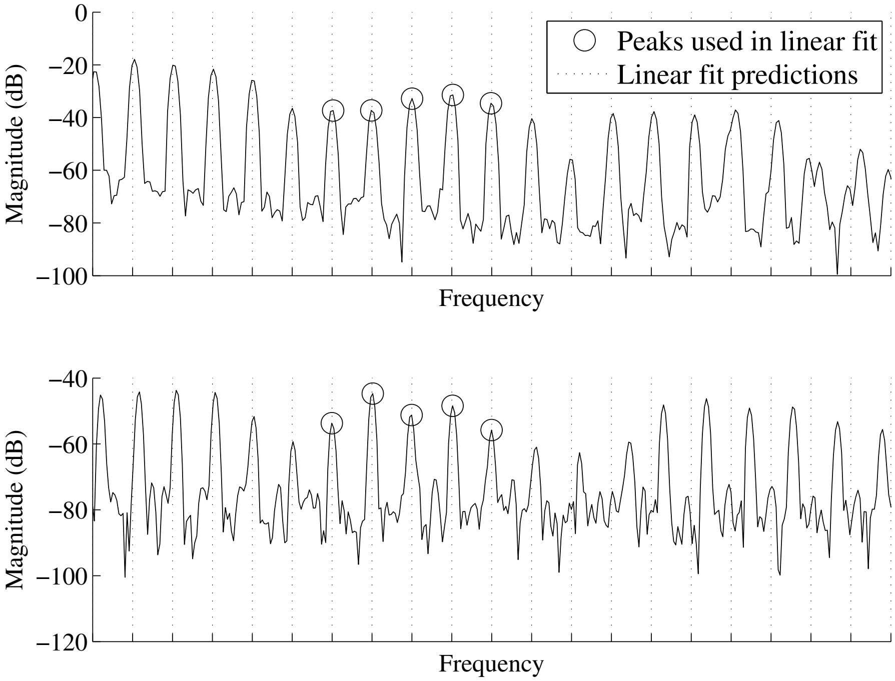

Figure 3.13: Detecting the transverse partials in an inharmonic spectrum with a local linear approximation. The frequencies of the circled peaks are used in a Linear Least Square fit. In the upper plot, the series behaves as expected, nearly harmonically. In the lower plot, however, an unexpected behaviour is visible, rendering the linear approximation (dashed lines) unsuitable.

the upper plot, the series is well behaved. The frequencies predicted by the linear fit, shown in vertical dotted lines, are consistent with the series over a certain number of partials, and then, on either side of the fit, a discrepancy progressively appears. The lower plot, in turn, is a typical illustration of a place in the spectrum where the linear-fit detection might fail. The three partials after the five circled peaks are suddenly and unexpectedly high. Thereafter, the series seems to regain a normal behaviour, but because the linear approximation is only valid about few partials, they are beyond reach, the discrepancy being already too large. A quadratic fit has been tried, to widen the interval of good approximation, but this approach fails even more easily if an outlying partial is confused for a transverse partial, or even when such unexpected behaviour as seen in Figure 3.13 makes the partials deviate dramatically.

For now, the MAT approach is far more reliable than the LLS approach, for the reason that the MAT approach is based on the model $\omega_k = k\omega_0\sqrt{1+\beta k^2}$, which applies to the *entire* series. Thence, the median statistics gathered throughout the entire search are relevant to any place in the spectrum, and local deviation from the expected behavior has practically no effect on the ω_0 and β estimates. In this regard, though, the LLS fit can be used in a spirit equivalent to the MAT approach, as these FF and IC estimates can also be obtained from the coefficients of the fit, could therefore be collected just the same, and the median of all these estimates could be used to estimate the frequency of the partials yet to be found. Indeed, the derivative for the pseudo-harmonic series at the index k_0 is

$$\omega'(k_0) = \omega_0 \frac{1+2\beta k_0^2}{\sqrt{1+\beta k_0^2}},$$

of which the first coefficient of the vector $\boldsymbol{\omega}$ offers a good estimate. The second coefficient of this vector is $\hat{\omega}(k_0)$, which is an approximation of the value of the series at index k_0, $\omega(k_0) = k_0\omega_0\sqrt{1+\beta k_0^2}$. From the value of the series at a given index and the derivative of the series at this same index, the fundamental frequency and inharmonicity coefficient can be evaluated, with

$$\omega_0 = \frac{1}{k}\sqrt{\omega(k)\left(2\omega(k) - k\omega'(k)\right)} \tag{3.34}$$

and

$$\beta = \frac{k\omega'(k) - \omega(k)}{k^2(2\omega(k) - k\omega'(k))}, \tag{3.35}$$

for all k. The coefficients of the linear fit, $\hat{\omega}'(k_0)$ and $\hat{\omega}(k_0)$, can thus be substituted into (3.34) and (3.35) to get the required estimates. These are, theoretically, as valid as the estimates based on pairs of partial frequencies shown for the MAT approach in equations (3.25) and (3.26), except that here, there is no question as to which combination of partials should be picked. Yet, an unexpected trend emerged in the implementation of this approach: the inharmonicity coefficient estimate tended to get bigger when partials of higher order were used in the fit. This could be due to the increasing rate of change in the derivative of the series,

$$\omega''(k) = \omega_0\beta k\frac{3+2\beta k^2}{(1+\beta k^2)^{3/2}},$$

affecting the linear approximation. Whichever the reason, such a trend renders the use of statistics for the estimation of the series' constants futile. In contrast, the MAT approach as described in Section 3.3.1 works fine, is cheaper computationally, and is easier to implement. It was nevertheless deemed that the LLS approach should be discussed here, at once because it is an intuitive approach that one might think of and because is emphasizes, by comparison, the appropriateness of the MAT method.

3.3.3 Accounting for phantom partials

We finally address the problem of accounting for the phantom partials in the string cancelation process at the time scale of a frame. Following the physical analysis of Chapter 1, we know that these partials are ordered in frequency according to the series

$$\omega_k^{\mathrm{L}} = k\omega_0\sqrt{1 + {}^{1}\!/\!_{4}\beta},$$

already seen in Equation (1.56).

A typical feature of phantom partials is that they are generally of lesser amplitude than the transverse partials, and are at times so low, that they are invisible, drowned in the surrounding noise. This phenomenon can be witnessed in the Spanish guitar and grand piano spectra of figures 1.11 and 1.12. When they are so low, it becomes superfluous to try and cancel the partials. Augmenting the string cancelation process with an estimator that decides when a partial should be treated or not, depending maybe on its magnitude, or its audibility, could help improve the computational efficiency of the string extraction process.

Not to waste computational resources, a more important aspect of the cancelation is the sequence according to which transverse and phantom partials must be canceled. We already suggested, in Section 3.2.1, a way to overcome the problem of cross-interference in the case of overlapped transverse and phantom partials. It was proposed there to measure and cancel the dominating peak first, and the dominated partial, second. If phantom partials were always of lesser amplitude than the neighbouring transverse partials, then a first detection-cancelation pass could be run to clear the spectrum of the transverse partials, and then a second pass, to clear the phantom partials. But it can happen that some phantom partial might be greater and actually dominate a transverse partial. Because, in a situation of overlap, the greatest partial must be canceled first, this uncertainty makes the cancelation process more delicate. A brute-force approach would be: in a first pass, to see which transverse partials are dominating and which are being dominated, and cancel all the dominating partials ; in a second pass, to move on to the phantom partials, which at that stage must all be dominating and can therefore all be canceled ; and in a third and final pass, to eliminate the last few transverse partials. Instead, we propose here an algorithm which guarantees the cancelation of all peaks in one pass only. Before we present this algorithm, an introduction to the concepts of *peaks*, *bulges* and *caves* is necessary.

A peak is, in a magnitude spectrum, any local maximum, that is, a sample on either side of which the magnitude is lesser. A bulge, in turn, is a positive local minimum in the curvature of the magnitude spectrum. It should be noted that a peak is necessarily a bulge, but that a bulge is not necessarily a peak. Rather, we should say that a peak is necessarily *associated* with a bulge ; this nuance is important because the sample of minimal curvature does not necessarily exactly coincide with the place of maximal magnitude. In Figure 3.14, for example, it is visible that the peak is to be associated, in this case, with the curvature minimum that is closest in frequency. On the other

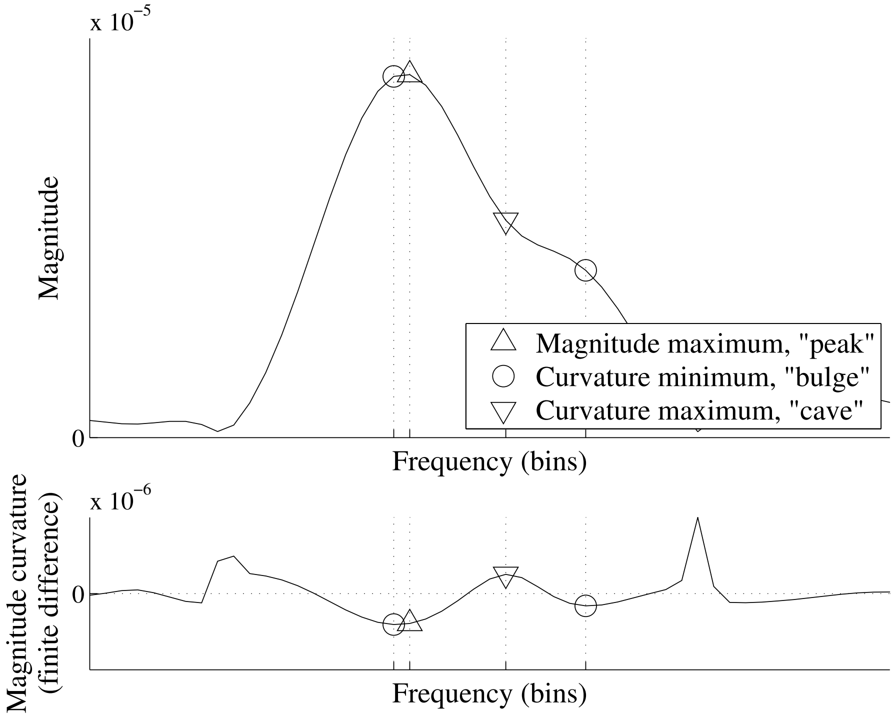

Figure 3.14: *Peaks*, *bulges* and *caves*: three concepts necessary to decide securely which of the two overlapping partials should be canceled first. Even when they are the effect of one same partial, a curvature minimum (bulge) does not always coincide exactly with a magnitude maximum (peak).

hand, the other bulge, clearly the effect of an overlapping partial of lesser magnitude, should not be associated with any peak. The problem addressed now is how to make the distinction between the bulges that are peaks and those that are not. After various experiments, it was found that the most reliable way of doing so was to check whether, between the examined bulge and the nearest peak “on the way up” in magnitude, there was a *cave* or not. In mathematical terms, a cave is a positive local maximum in curvature. The algorithm should scan the magnitude spectrum, starting from the bulge that is being examined, following the direction where the magnitude goes up, and declare that it is a peak if a peak is found first, or only a bulge if a cave is found first. Note that it is not impossible, in the definition laid above, that a peak is also a cave, counter-intuitive as it may seem. In such case, of course, it should be considered to be a peak.

This basis laid, we can now introduce the detection-cancelation algorithm. Basically, the search goes from the lowest to the highest harmonic index for the transverse and phantom series at once, trying all harmonic indices. Peaks are canceled as soon as they are found, and transverse/phantom bulges which are dominated by phantom/transverse peaks wait for the latter to be canceled, and are thereafter canceled in turn. The algorithm is as follows:

1. Set k and l, the harmonic indices of the transverse and phantom series, to 1. Allocate two boolean values, `transverse` and `phantom`, which will become useful later on. Initialise the fundamental frequency with such an estimate as that returned by the autocorrelation method, and set the inharmonicity coefficient to zero.

2. Repeat the following steps until all string partials have been detected and canceled.

3. Find the bulge closest to ω_k and ω_l^{L}, referring to equations (1.33) and (1.56), respectively, with the current fundamental frequency and inharmonicity coefficient estimates.

4. For each detected transverse and phantom bulges, find out whether

they are peaks or not, according to the method described above.

5. If the transverse and/or the phantom peaks/peak are/is marked as canceled, increment k and/or l by 1 and return to step 3.

6. Now it must be determined whether it is a transverse or a phantom peak that is going to be canceled, if any. This depends on the result to the previous step, of which four cases are possible:

   - Both the transverse and phantom partials are peaks, and they are *distinct* peaks. In this case, select the peak of least frequency for measurement and cancelation. Set `transverse` and `phantom` to `true` or `false` accordingly.

   - Both are peaks, but they turn out to be the same, which should be what happens for the first few partials. Select this peak for measurement and cancelation, and set *both* `transverse` and `phantom` to `true`.

   - Only one of the two is a peak. Select that which is a peak for measurement and cancelation, and set `transverse` and `phantom` to `true` or `false` accordingly.

   - None of the two bulges is a peak. There is no need to select a peak, because no measurement and cancelation will take place. Set both `transverse` and `phantom` to `false`.

7. If either or both `transverse` and `phantom` are true, proceed to the measurement-cancelation part of the selected peak.

8. If `transverse` is `true`, use the frequency measurement and current transverse harmonic number to refine the fundamental frequency and inharmonicity coefficient estimates.

9. Now the harmonic numbers k and l must be incremented appropriately:

   - If `transverse/phantom` is `true` and `phantom/transverse` is `false`, increment k/l, and if the transverse/phantom peak is higher in frequency than the peak that dominates the phantom/transverse bulge, increment l/k. This last step ensures that the phantom/-transverse bulge is not left behind, waiting unnecessarily to become a peak.
   - If neither or both `transverse` and `phantom` are `true`, increment both k and l.

These steps should be repeated until all partials that should be canceled are canceled. For the purpose of deciding when to stop, it is useful to set a frequency threshold. One way to do so is to detect all the peaks, or local maxima, of the magnitude spectrum before the detection-cancelation process is started, and set the threshold to the highest (in terms of frequency) peak above a magnitude threshold. This magnitude threshold can be -96dB, which is the Signal-to-Noise Ratio of a 16-bits Analogue-to-Digital Conversion [OSB99], or greater, for noisy recordings. Alternatively, Fletcher-Munson equal-loudness contours [FM33] may be used to estimate which peak is the last to be audible. Of course, the frequency threshold should not go beyond the Nyquist frequency, half the sampling rate, where the magnitude spectrum is mirrored. Ultimately, this can be set as a frequency threshold, but most often the frequency-domain decay rate of the magnitude of the string partials is such that any measurable energy vanishes well before this frequency.

Conclusion

This chapter proposed a Phase Vocoder process for the virtual extraction of a string in monophonic, plucked or hit string tones. The Phase Vocoder scheme is based on a Short-Time Fourier Transform (STFT) time-frequency representation of the signal, where the input is first decomposed into a succession of overlapping *grains*. It was therefore logical to begin, in Section 3.1, with a complete formulation of this granulation process, which was to satisfy two conditions: first, that the scheme is *transparent*, and second, that the window used in the granulation process shows optimal frequency-domain properties for the reduction of the cross-interference during the estimation of the parameters of the string partials.

Once the time-domain arrangements were described, the discussion could move on to the short-time spectra obtained from the Fast-Fourier Transformation of the grains, where the detection and cancelation of the string partials take place. It was chosen to describe the cancelation of the partials first, in Section 3.2. The linearity of the cancelation process and the Fourier transform makes it equivalent that the cancelation takes place in the time or frequency domain. It was nevertheless shown that a frequency-domain approach has significant advantages, essentially regarding the computational savings that it offers. This section was also the appropriate place where to mention the difficulties of such cancelation process, independently of whether it is done in the time or frequency domain, when it comes to dealing with the attack of the tone and overlapping transverse and phantom partials.

Finally, the problem of detecting the string partials was approached, in Section 3.3. From the presence of phantom partials were shown to arise situations of strong overlap, where the smallest of the two partials cannot emerge as a peak, but is still discernible by the bulge it causes in the dominating partial. To account for this phenomenon, an algorithm was introduced, where, in situation of overlap, the largest partial was canceled first, clearing the second partial which, in turn, could become a peak, be measured and canceled. The algorithm was designed so as to require one "pass" only, and it makes use of the linearity of the frequency-domain subtractive cancelation process that this thesis has introduced. Regarding the detection and identification of the partials, the method that was found the most secure was the method of Median-Adjustive Trajectories (MAT), introduced by the author in [HWTL09]. Yet new material since the introduction of the method was presented in Section 3.3, showing how the square complexity of the MAT could be reduced to a linear complexity. Also, an alternative, intuitive approach of locally approximating the pseudo-harmonic series with a linear series was explored. But such approach does not feature such robustness as the MAT does, and the reasons why were identified.

Chapter 4

Tests, experiments and results

In this chapter, we test the string extraction method described throughout this thesis onto a variety of string instruments of contrasting character, both professional and custom recordings. A list of the featured instruments follows:

- Steel-string acoustic guitar
- Nylon-string acoustic guitar
- Spanish guitar
- Martin acoustic guitar, recorded in the NUI Maynooth Music Technology Laboratory by the author
- Stratocaster electric guitar
- Harpsichord
- Grand piano
- Double-bass (plucked Jazz-style)

- Viola (*pizzicati*)

All samples, except for the Martin and Spanish guitar samples, are professional-quality Yamaha samples [A5000 Professional Studio Library]. The exact brand of the instruments involved is unknown, except for the Fender Stratocaster electric guitar, of which it can be expected that it is American-built. Likewise, it is probable that the grand piano is a Steinway, the reference brand. The Spanish guitar samples were obtained through the use of Yellow Tools' *Independence* software sampler http://www.yellowtools.com/ (latest access: October 20th, 2011).

Evaluation criteria

The string extraction examples are evaluated after two general criteria: aural quality and computational efficiency. Having said that, it will become obvious to the reader that our main focus will be aural quality. Although our approach to the problem of string extraction has been striving to facilitate real-time implementation, the method could still be found to have some uses even if, at the end of the day, it could not run in real-time (for example, as a pre-processing unit for samplers). On the contrary, a real-time application that does not produce results of respectable aural quality would probably not be of much use.

Another point that makes us give priority on quality over performance is the impracticality of testing the actual computation time in a non-optimal implementation. Our method is currently implemented as a Matlab prototype, while ultimately, a real-time music application should be written in C++. Matlab code is far more concise and straightforward than C++ code, making prototyping and experimentation convenient, but it is also much slower, and does not support real-time input or output. Within the scope of this thesis, we therefore cannot test comprehensively how well the method can run in real-time. This possibility aside, encoding efficiency was mostly disregarded for the benefit of code conciseness readability. The only sort of computational efficiency testing that we could include in this thesis is *comparative* testing, such as in Section 4.3, where the cancelation of a partial through Fourier-series approximation of its main lobe is compared with an FFT approach. Only then was the effort made to write some code that requires the least possible computation time, so as to compete with the computation power of the FFT.

Aural and visual quality evaluation

Coming back to quality, the means in our possession in this thesis to evaluate the success of string extraction are aural and visual. Whether an input that has undergone string extraction "sounds good" or not is certainly the test that matters most, as eventually all the efforts put in this thesis are intended to be useful in musical applications. A thesis document can nevertheless only support writing and graphics, but the sound examples alluded to in this chapter can be found on the Compact Disc accompanying this document, as well as on the Web page http://www.cs.may.ie/~matthewh/ThesisExamples.html. Each example in this chapter was given a unique number which will be used as reference.

An acute human ear is the best judge of the quality of a tone, but in our case the spectrogram and waveform of the input and output can convey useful information. The examination of a waveform after extraction might indicate whether an instrument's body response is very reverberant or not, or if some sinusoidal energy remains, or else if there are traces of background noise(s). A spectrogram, in turn, says a lot about the cancelation process. For example, it can happen, in complex spectra such as a grand piano's, where a lot of overlapping between transverse and phantom partials occur, that our method cancels a given peak in a given frame, and spares it in the following frame, to cancel it again the frame after. This phenomenon, referred to as *burbling* in [IS87], is very visible in the spectrogram of the tone *after* cancelation but *before* re-synthesis. This is another interesting point, that due to the overlapping of the analysis frames in the Phase Vocoder scheme, the spectrogram of the waveform after re-synthesis will not be the exact same as that of the input after frequency-domain string extraction. We illustrate this phenomenon in Figure 4.1. Here is visible that the spectral energy of each frame is smeared upon re-synthesis, both horizontally and vertically. The overlap factor used in this figure is 5, which means that a given window will reach into the two windows before it and the two after. In accordance, the energy in frame of index 0, in the upper spectrogram, can be seen in the lower spectrogram to leak into the two frames before, where the energy was originally nil.

As a recapitulation of the various means of assessing the quality of the extraction process, let us examine the string extraction example of a harpi-

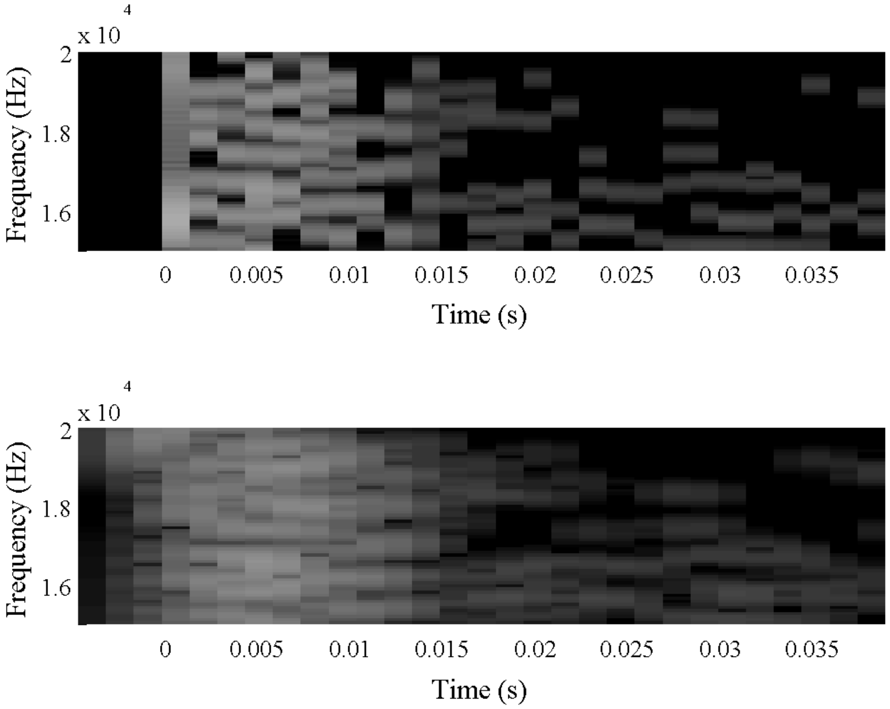

Figure 4.1: Frequency-domain synthesized string spectrogram (top) and spectrogram after time-domain re-synthesis (bottom). Due to the overlap (here, 5) of the analysis frames, these are not equal. Notice the smearing, both horizontal and vertical, of the spectral energy.

chord A4 (MIDI note 69). The waveform before (lighter shade) and after (darker shade) processing are superposed in Figure 4.2. In a similar arrangement, the spectrogram of the harpsichord tone before processing is seen in the upper panel of Figure 4.3, and after processing, in the lower panel of the same figure. Finally, the corresponding audio waveforms (here, 03.wav and 04.wav) can be downloaded from the webpage mentioned earlier. A burst of energy is visible in the processed waveform, and the processed spectrogram indicates that this burst of energy is broadband. The processed waveform

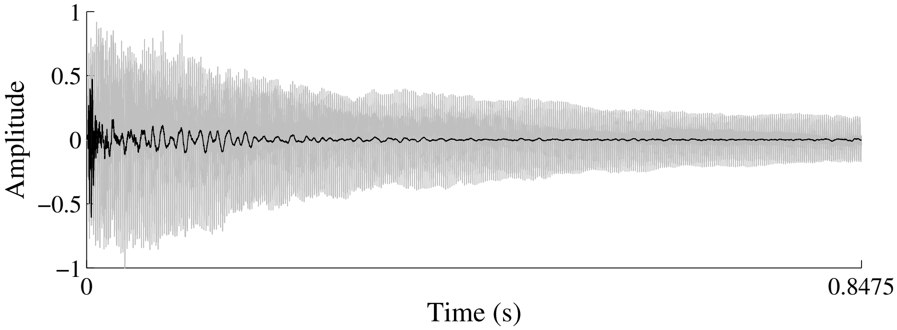

Figure 4.2: Harpsichord A4 (MIDI note 69) before (lighter shade) and after (darker shade) string extraction.

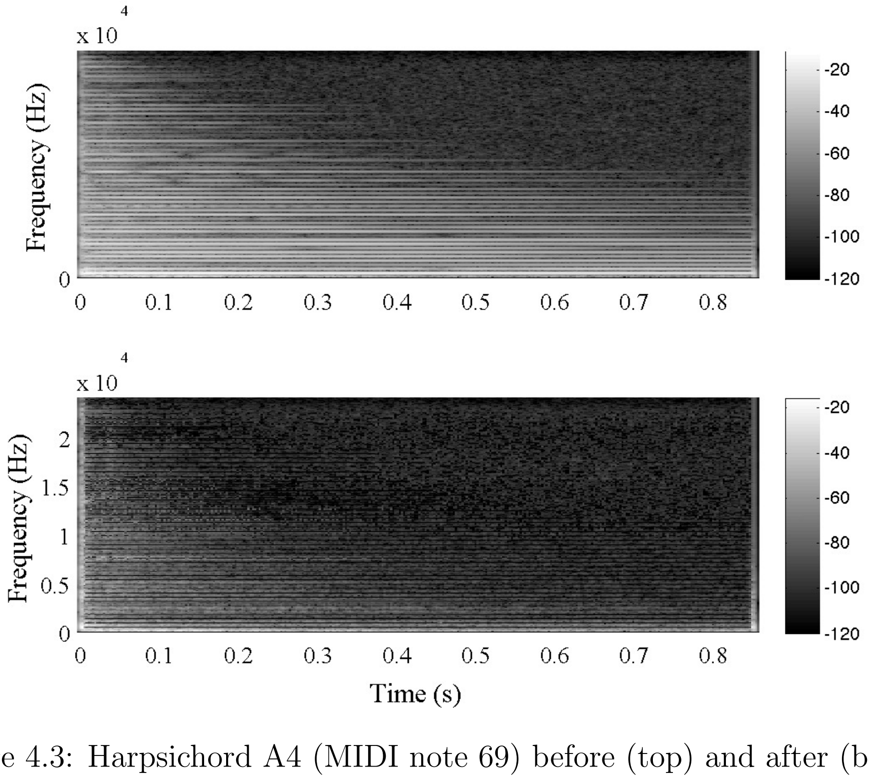

Figure 4.3: Harpsichord A4 (MIDI note 69) before (top) and after (bottom) string extraction.

also features some residual, low-frequency energy, which is difficult to locate in the processed spectrogram. However, the partials can be seen there to have been canceled correctly, which suggests that this residual is due to some body resonance. But this hypothesis, however, will really be confirmed or rejected upon comparative listening of the sound file before processing, to identify the pitch of the original tone, and after processing, to see whether this resonance can be part of the extracted string's original series or not. These are the kind of commentaries that our graphics and sound files may inspire.

Organisation of this chapter

The string extraction method we developed throughout this thesis is innovative in some essential respects: the use of Median-Adjustive Trajectories to detect partials, the frequency-domain cancelation of partials through subtraction and the resulting subtractive approach to the handling of overlapping transverse and phantom partials, as well as the unit-step model for the analysis of Phase Vocoder time segments that overlap with the onset of the tone. First, general examples will be given, where the overall method, including all the above-mentioned features, will be tested. The aim there is to convince the reader of the potential of the method, and also to give the opportunity to pinpoint the concept of *string extraction* (as opposed to *excitation* extraction) advanced in this thesis. Not to overload this document, the examples there will be relatively few, but of strategic importance to nourish the discussion. More examples can be found at http://www.cs.may.ie/~matthewh/ThesisExamples.html.

Thereafter, we will move on to testing the respective contributions of the techniques original to this thesis to the overall quality of the method. We choose, rather arbitrarily, to approach them in their order of appearance in the body of this work: the Complex Spectral Phase Magnitude Evolution (CSPME); the approximation of the partials' main lobes with the amplitude-modulated cosine window spectrum; the modeling of the tones' onsets with the product by a unit-step window; and the handling of phantom partials. Limitations of the method in its present state will arise in the various contexts of this discussion, and will be acknowledged in due time.

4.1 Testing the method overall

The process of *string extraction*, defined in the introduction to this thesis, is the process of "removing", virtually, all vibrational component from the input that a direct transduction of the object string's vibrations (i.e. the string that the process aims at extracting). Indirectly, the instrument's body is going to respond to the excitation of the string. This is the part of the sound that remains after string extraction that is commonly referred to as the *excitation* [LDS07]. However, unless the input sound was very carefully recorded, and where the noise floor is negligibly low, this is not the only remainder. Open strings of the instruments that are not carefully muted are also responsive, although in a damped manner, to the excitation of the processed string, which is transmitted through the bridge of the instrument. Also, the recording might comprise other accidental sound sources, such as the breathing of the performer, or the buzz of an electric guitar's amplifier. The cancelation process is not going to account for these, which, as a result, are still going to be present in the output. In such cases, the term *excitation extraction*, i.e. the isolation of the response of the body from all the rest, is deemed inappropriate. Rather, it is the string which is isolated from all the rest, hence our preference for the term of *string extraction*.

There are nevertheless cases where the extraction of the string is equivalent to the extraction of the excitation, that is, in the cases where only the string's energy and the energy of the body response are non-negligible. This is the case of the viola pizzicato example, in Figure 4.4. Listening to the corresponding sound files (01 (input) and 02 (output), one realises that the residual waveform visible in the figure is the response of the body alone, as no trace can be heard of sympathetic vibrations from other strings or ambient noise. This is not the case, however, of the Stratocaster example seen in Figure 4.5 (sound files 11 and 12). There, the response of the body is much shorter-lived (which is not surprising, given that an electric guitar amplifica-

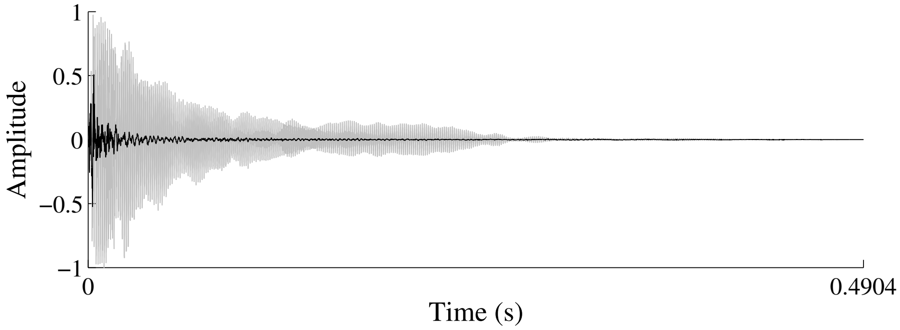

Figure 4.4: Viola G5 (MIDI note 79), played *pizzicato*.

tion system does not rely on an acoustic, resonant body), but it is visible that the waveform then settles to a steady noise, the electric buzz of the amplifier. A similar comment can be made about the Martin E5 example of Figure 4.6 (sound files 13 and 14), where the waveform shows a noise floor of about -60 decibels (which, as a way of comparison, is approximately the noise floor of analog audio tapes [Ear03], common before the advent of the CD in the early 1990s). The variations in the shape of this noisy waveform (after the body response has faded out) are revealed to be frictions caused by movements of the performer in the recording room (certainly not a professional!). Finally, another typical situation where the remainder of a successful cancelation process may not be the response of the body alone is exemplified in Figure 4.7 (sound files 09 and 10), where an acoustic guitar open D3 (MIDI note 50) is processed. There the other open strings were not carefully muted, and their ringing can be distinctly heard after the extraction of the object string.

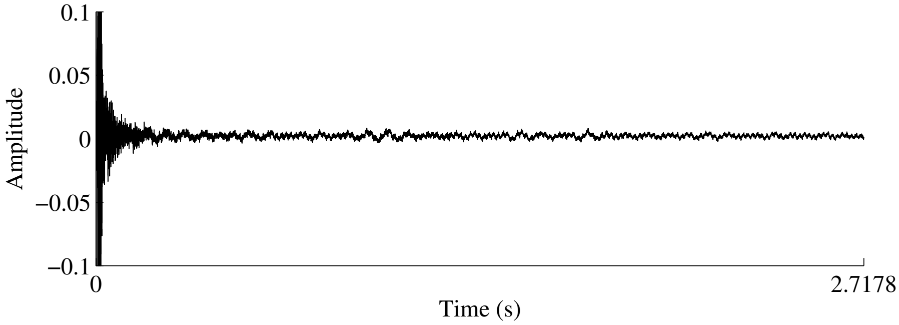

Figure 4.5: Stratocaster D3 (MIDI note 50): example of noise which is neither string nor excitation

In the Stratocaster, the Martin and the acoustic guitar examples above, one cannot talk of *excitation extraction* as such. For actual excitation ex-

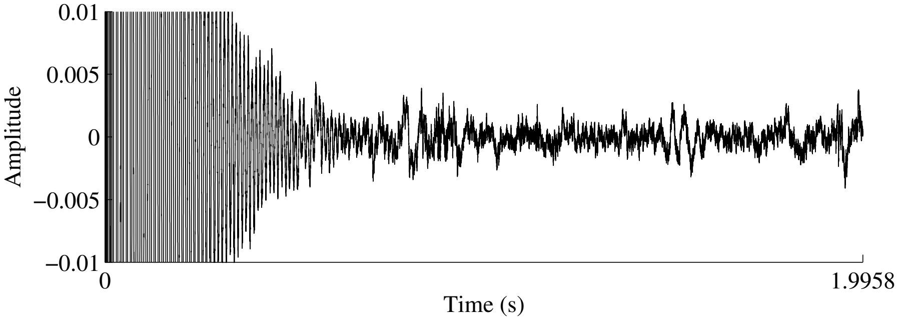

Figure 4.6: Martin E5 (MIDI note 76) after string extraction. The “events” in the noise floor are movements of the performer during recording, only audible after string extraction.

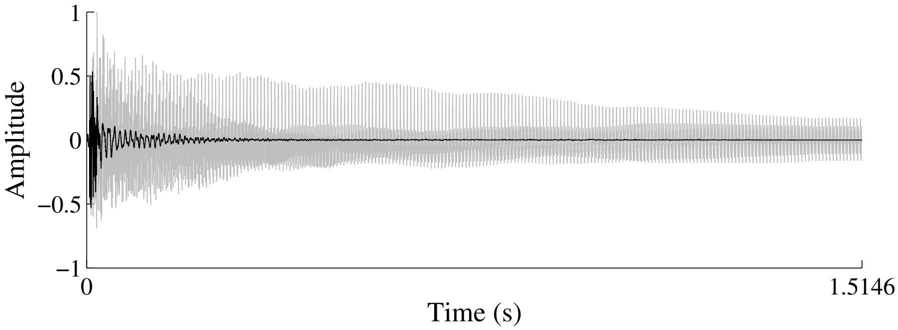

Figure 4.7: Acoustic guitar open D (MIDI note 50). In this example, the sympathetic vibrations of the other open strings are clearly audible in the processed sound.

traction, for such use as in Commuted Waveguide Synthesis [KVJ93, Smi93], the method would have to take into account the potential sympathetic vibrations of other strings, as well as potential background noise. Else, for our method to be useful to excitation extraction, there should be the requirement on the input that no sympathetic vibration or noise be present, or in negligible quantities. But this goes contrary to the orientation of this thesis’ work, which aims at facilitating real-time implementation, for real-time musical situations. Such constraint, on the contrary, really can only be met in carefully arranged, studio conditions.

The results presented so far have, in addition of demonstrating the potential of the method, inspired a discussion which helped refining the definition of this thesis' work. This done, we can now evaluate the respective contribution of the various methods found in this thesis.

4.2 The CSPME exponential-amplitude generealisation and its contribution to the cancelation process

The CSPME method, introduced in Section 2.5.1 of this thesis, returns the amplitude modulation and frequency information of an exponentially-decaying sinusoidal component. These measurements can be used to synthesize a cosine window's spectral lobe as modified by such time-domain modulation, and cancel efficiently the analysed partial. This method was deemed most suitable for our purpose in Section 2.4.1, where we took an informed guess as to whether a constant-amplitude model was good enough, over the short period of time of a windowed grain, to cancel string partials satisfactorily. There, we compared the constant-amplitude model with a linear-amplitude model, which was deemed to be a good approximation of the actual exponential model over such a short period of time. The gradient of the linear-amplitude model was derived from magnitude measurements of partials taken throughout the overall tone. The results, summarised in Table 2.3, led us to the conclusion that bright instruments (such as the piano or the harpsichord) had so slow decay times that a constant-amplitude approximation could be enough, but that duller instruments (such as the double-bass), with their far more rapidly decaying partials, seemed to require a CSPME approach.

For this reason, the contribution of the exponential-amplitude generalisation in the measurement and cancelation processes was first tested upon some double-bass samples. Surprisingly, the output was found to be of identical quality whether the amplitude modulation information returned by the CSPME was used in the cancelation or simply disregarded. The difference between the waveform rose to -60 decibels, so the outputs were at least not identical, but neither aurally nor visually could a distinction be made. This example is given in Figure 4.8 (sound files 24 and 25), but only for the case where amplitude modulation was taken into account, to avoid redundant graphics.

On the other hand, it was equally surprising to find that, for the lower partials in the lower range of much brighter instruments, it really did not make a difference whether the amplitude modulation was accounted for or not. With slower-varying amplitude envelopes, our conclusion of Section 2.4.1 was that a constant-amplitude approach might have yielded results of satisfactory quality. Figure 4.9 shows, however, that a trace of the partials remains when a constant-amplitude model is used, while the cancelation is visibly more successful when the exponential-amplitude constant returned by

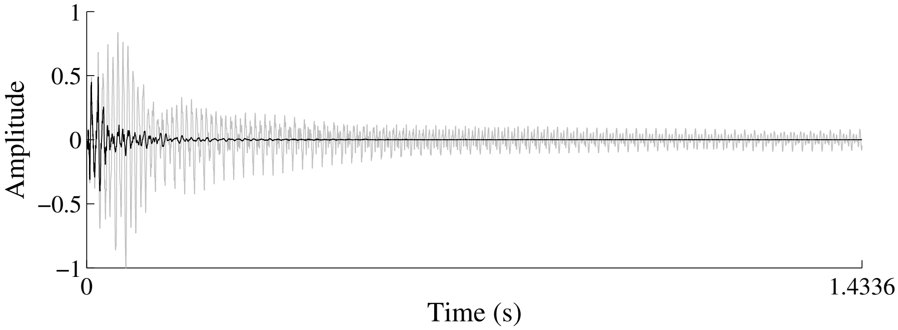

Figure 4.8: Double Bass A2 (MIDI note 45) before (lighter shade) and after (darker shade) string extraction: a constant-amplitude model, for this heavily damped tone, yields a result of equivalent quality.

the CSPME is taken into account. The difference is as much audible in the waveforms (sound files 15 and 16) as it is visible in the spectrograms. In the constant-amplitude case, for instance, the pitch of the original sound is still clearly recognisable, while hardly audible in the CSPME case.

4.3 Fourier-series approximation of a partial's main lobe for synthesis and cancelation

Following the introduction, in Section 2.5.1, of the CSPME for the obtention of the frequency and amplitude modulation constants r and γ of the measured partial, a complementary method to the subsequent obtention of the amplitude and phase constants A and ϕ was described. This method is equivalent to those found in [SG06, Zö02], but generalised to amplitude-modulated spectra. In short, r and γ serve to derive an "intermediary spectrum", whose values can be used to divide the corresponding values of the analysed spec-

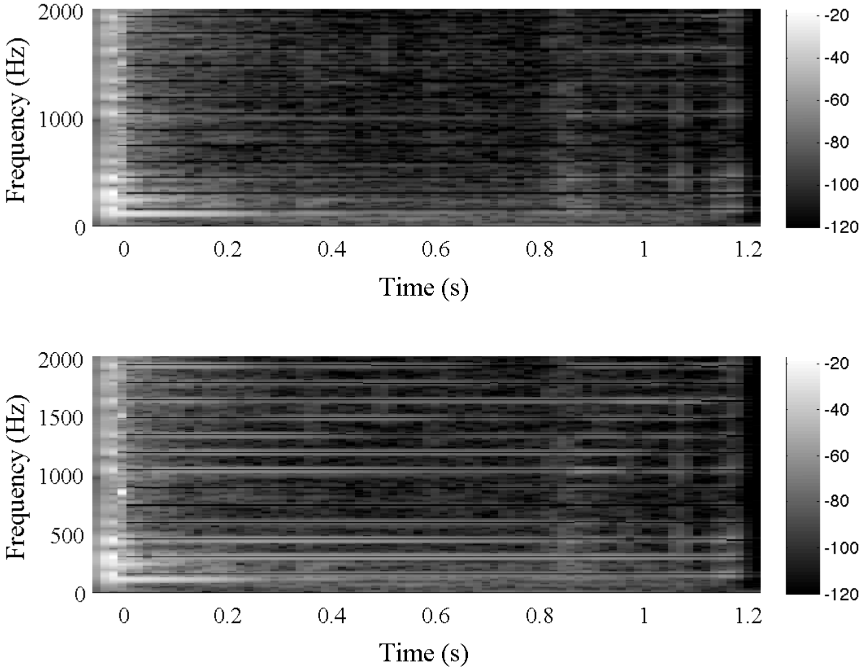

Figure 4.9: Nylon-string acoustic guitar D3 (MIDI note 50) with CSPME (top) and a contant-amplitude (bottom) models for cancelation.

trum. From this quotient can then be extracted the amplitude and phase constants. In our subtractive partial cancelation method, these constants are then used to scale and phase-increment the intermediary spectrum, which then becomes a faithful copy of the original spectrum.

This intermediary spectrum therefore has two roles: the estimation of the amplitude and phase constants, and the cancelation of a partial through its frequency-domain synthesis and subtraction. For the former task, only one spectral sample of the intermediary spectrum is necessary – preferably, that of maximal magnitude, because it is generally the least exposed to the cross-interference from other partials. For the purpose of subtractive cancelation, all spectral samples that are not negligibly small should be synthesized. For a cosine window, where the essence of the energy is concentrated in the main lobe, this may be the spectral samples of the main lobe only, which in general amount to about 10. The question then is: how to synthesize these samples? The easy option is to take the Fast Fourier Transform of the product of a complex exponential whose frequency is r, decay rate γ, and the window used in the analysis. This problem has the drawback of generating M samples (for a length-M analysis) while only about 10 of these are necessary, and also the necessity of synthesizing N samples of a windowed complex-exponential is still present. Instead, could the synthesis of the useful samples using the analytical expression of the intermediary spectrum (2.67) be more economical? This is what we aim at testing in this section, along with the assumption that the sidelobes of a partial can be spared in the cancelation without compromise of quality.[1]

Two versions of the string extraction method that differed only in the means of evaluating the above-mentioned spectral samples were implemented: in the first, the analytical approach was used, in the second, the FFT approach. Time counters were introduced before and after the lines of code where these methods are implemented to compare the computation times.

[1] It should be pointed out, however, that even if the tests were satisfactory, that the analytical approach to synthesis has the drawback of requiring an analytical expression for the spectra of our partials. (2.67) is for steady-state, cosine-windowed partials. For windows overlapping with the onset of the tone, an analytical expression for the spectrum of the product of the signal, the window and the unit-step function could not be found. If this method were opted for, it could therefore apply to steady-state windows only, and if onset-overlapping windows were to be treated in the same subtractive way, a switching from FFT synthesis to analytical synthesis should be included in the implementation.

The analytical approach, on average, requires around 20% of the time required for the FFT approach. In the worst cases, it still takes less than 50% of the FFT time, and in the best cases, less than 5%. It is yet unclear, however, why these percentages can show such variations.

The quality of the outputs (which might differ because, in the analytical approach, a Fourier-series approximation is used and only the main lobe is subtracted) was also compared, by the examination of the spectrograms and the waveforms, visually and aurally. The tests were first run using a second-order, continuous cosine window (see Section 2.2.3 and Figure 2.4 in Section 2.2). In some cases (for example, in the double-base case), the output was of identical quality whether the entire set of FFT samples was used for subtraction or only the main lobe of the analytically-synthesized spectrum. In further tests, on instruments of brighter spectrum, an audible, low-pitched buzz could be discerned with the analytical approach case, that was not in the FFT case. The reason for this potential artifact is that the spectral energy of the sidelobes of a second-order continuous cosine window is, against our assumptions of Section 2.5.2, not completely negligible, and actually may contribute to some audible extent to the re-synthesized waveform. To remedy to this problem, the nearest side-lobes can be taken into account in the resynthesis, but this is not very elegant. Else, the order P of the window can be increased to 3. Then, a yet greater proportion of the spectral energy would gather in the main lobe (i.e. compare the bottom left and right subplots of Figure 2.4, Section 2.2), and now the main-lobe only approximation might be enough. Comfortingly, the tests then gave entirely satisfactory results. The use of a greater window order, however, is a computational drawback: the main lobe is now wider and thus comprises more spectral samples, and the synthesis of each sample is also more costly, given the convolution with more cosine window components in (2.59). Yet, the computation cost comparison with the FFT still returned a similar gain of time in favour of the analytical approach.

4.4 Behaviour of the cancelation process in onset-overlapping frames

Throughout an input of regular length (i.e. at least a few tens of a second), most of the Phase Vocoder frames are fully into the signal. Over such periods of time, the signal can be said to be *steady-state*. However, there are a few frames (their number depends on the overlap factor) which overlap with the onset (e.g. the moment of release of a plucked string) of the tones, before which the waveform can be considered to be nil, and after which it is non-zero. We call such frames *onset-overlapping frames*.

Essentially, the problem encountered in onset-overlapping frames is the impaired resolution of the partials. Cross-interference might then become too big for the assumption that a magnitude local maximum is the effect of one harmonic only to be valid. In this section, we will check whether the cross-interference in overlapping frames is too big, or if the CSPME estimates could still be accurate enough for the cancelation in such frames to have some positive effect. Indeed, tests were occured where anomalies due to excessive cross-interfence were observed. Figure 4.10 illustrates such anomaly. In the

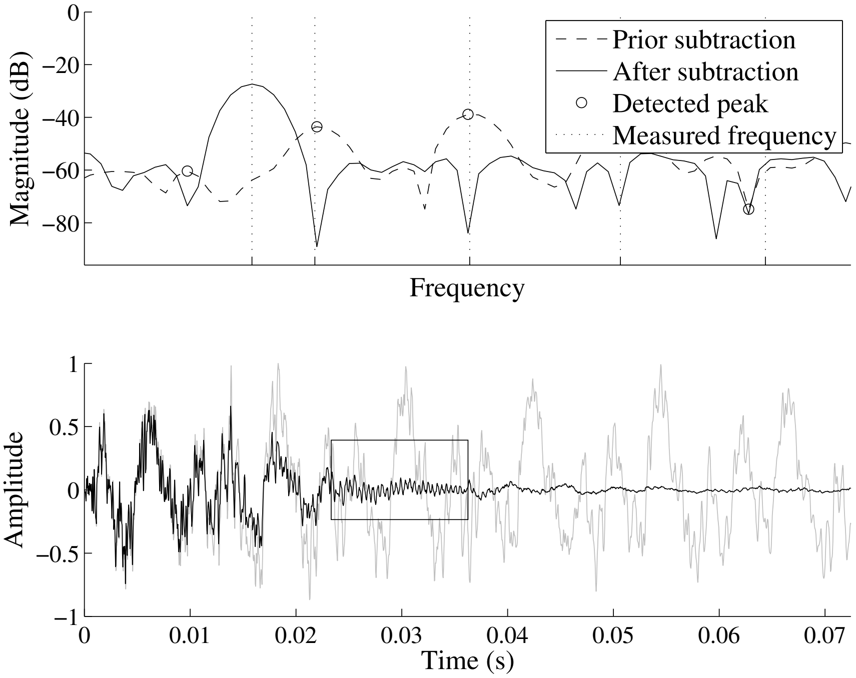

Figure 4.10: Acoustic Guitar E2 (MIDI note 40): Spectrum after subtraction (upper plot, solid line) should be lesser than before (dashed line). CSPME measurement of leftmost circled peak largely erroneous, responsible for added energy. In time-domain output (lower plot), results in outstanding sinusoidal grain.

upper plot, a magnitude spectrum of an onset-overlapping frame is shown. The dashed line represents the state of the spectrum before cancelation, and the solid line, after. The circles denote the places of maximal magnitude that were identified, during the peak detection process, as harmonics to cancel. The vertical dotted lines, in turn, indicate the CSPME frequency estimate of each of these harmonics. The reader should observe that the two closest circles to the left of the legend's box are each very near a vertical dotted line, which means that the the CSPME frequency measurement is at least reasonably exact. However, the leftmost vertical dotted line is much higher in frequency than the peak it was supposed to indicate the frequency of – the dashed peak where the leftmost circle sits onto. The CSPME frequency measurement there was excessively biased. Notwithstanding this, the measurement was automatically used for the synthesis of a main lobe, but upon subtraction, instead of canceling the targeted peak, it therefore added a lobe that was not present originally. This is responsible, in the time-domain waveform resulting from the (mis-)processed spectrum, for the appearance of a short burst of unexpected sinusoidal energy (outlined by a rectangle on the figure). At times, such bursts can be very audible.

The amount of cross-interference in the measurement of a partial depends on the "health" of this partial in the frequency domain (i.e. notice that the peak that was not measured correctly is smaller than the peaks that were) and on the amount of overlap of the analysis window with the onset in the time domain (as shown previously in Figure 3.11, Section 3.2.2), along with the order of the cosine window, because a cosine window of higher order has its energy concentrated nearer to its center (see Figure 4.11).

It can also be expected that windows which do not overlap too much with the onset might still be suitable candidates to the cancelation process. The numerous tests run confirmed in this hypothesis, although in some cases some sinusoidal bursts could occur. To present these results in an informative manner, an indication of the amount of overlap must be given.

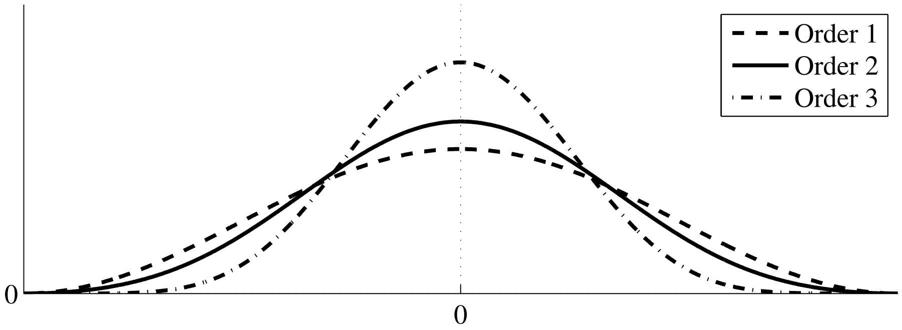

Figure 4.11: How cosine windows of higher order gather their energy closer to their center.

The window used by default in these tests is a second-order cosine window (solid line in Figure 4.11), and the Phase Vocoder process was set to use a minimal overlap, which, for a cosine window of order $P = 2$ is 5 (see Section 2.2.3). The time sample deemed closest to the moment of the attack, ν (Section 3.2.2), was determined manually, and the granulation of the input was synchronised to ν, so that the first sample of the first window not to overlap with the onset was ν. (The idea is illustrated in Figure 4.12.) The number of windows overlapping with the onset that are used for cancelation may be denoted along with the overlap factor by a fraction, such as 2/6 if the last two windows are used in a situation where the overlap is 6.

It was found that for high-pitched tones, an onset overlap of 0 (i.e. only beginning the cancelation process in the frames fully into the sound) is sufficient to yield aurally satisfying results, but that lower tones require some overlap, without which some sense of pitch persists at the moment of the attack, dissimulating, or at least altering, the true character of the response

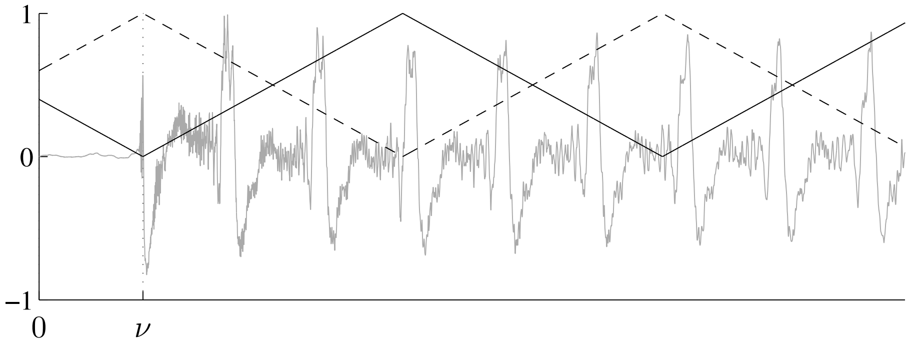

Figure 4.12: For testing, the windowing was synchronised with the attack sample, ν. The shape (triangular) an line style (solid and dashed) of the windows are for visual purpose only.

of the body. In general, even for the lowest-pitched tones, an onset overlap of 1/5 yields satisfactory results, although it must be said that increasing the onset overlap to 2/5 contributes to yielding body responses of exciting quality – provided, of course, that the cross-interference then was not so big as to get erroneous CSPME measurements and cause audible sinusoidal bursts. We give an example of this statement in Figure 4.13 (sound files 17, 18 and 19), where an acoustic guitar open bass E is processed. In the uppermost panel, the onset overlap is 0/5, and the sinusoidal energy remaining early in the early part of the processed wave (darker shade) gives a sense of the pitch of the original wave (lighter shade), even if slightly. In the middle plot, the onset overlap is augmented to 1/5, and it can be said that the sense of the original pitch, upon audition of the processed wave, has disappeared. However, the bottom-most example, where the onset overlap is pushed to 2/5, gives a very clear impression of the sharp response of the body to the string's sudden release.

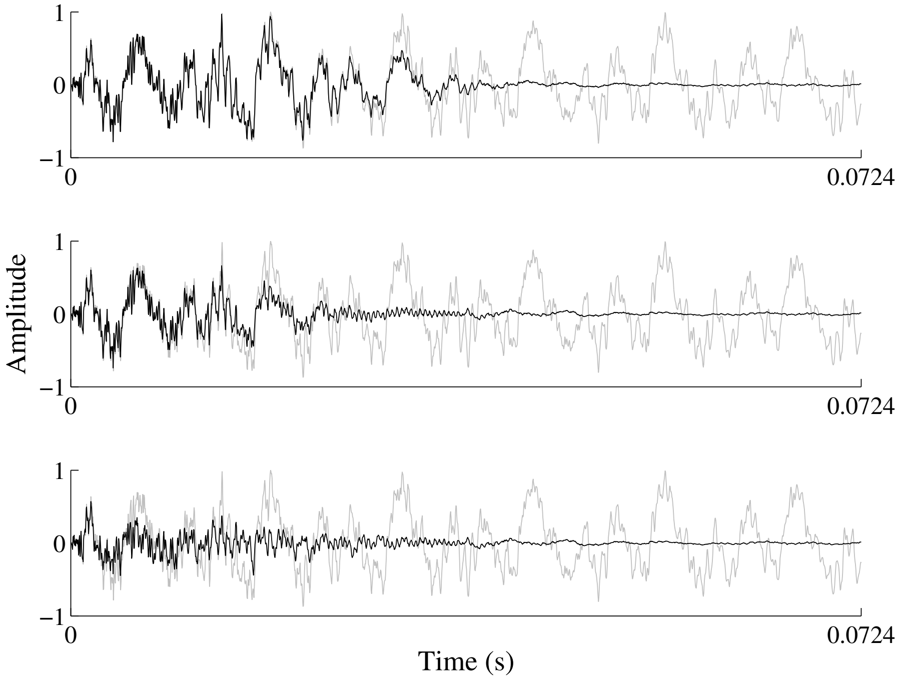

Figure 4.13: Acoustic Guitar E2 (MIDI note 40) with onset overlap 0/5 (top), 1/5 (middle) and 2/5 (bottom)

The benefit of dealing with onset-overlapping frames is thereby clear, but more research is needed to make the cancelation process then less susceptible of adding energy rather than taking some away. We succintly propose here a few ideas that could lead to more permanent solutions, but this list should not be considered exhaustive. To begin with, a brute-force approach is to evaluate the total spectral energy of the spectrum before the cancelation, and re-evaluate it after. It should be lesser, but if it were not, then the cancelation process could, by its linear nature, easily be undone.

Another idea can be inspired from the observation, on Figure 4.10, that the discrepancy between the frequency of the peak and the measured frequency is unreasonably large. Some threshold, decided heuristically, might therefore be useful in deciding whether a partial was measured properly or not. This method, however, along with the brute-force method mentioned above, has the inconvenience that some partial that should be canceled might be spared only on the basis that it could not be measured properly. An approach that does not have this inconvenience would be to resort to a quadratic-fit-based estimation of frequency, which, logically, cannot give an error that is greater than half the frequency resolution of the analysis if a magnitude peak is symmetrical about the frequency of the underlying partial. The inconvenience then is mainly the necessity to switch methods depending on whether there is onset overlapping or not. Also, the exponential-amplitude constant cannot be evaluated with a quadratic fit approach or, consequently, used in the cancelation process.

Time-domain implications of the unit-step model

Before the end of this section is reached, it should be pointed that the modeling of the region of the attack with unit-step windowing has another distinct benefit, that of avoiding any form of *pre-echo*. This phenomenon is mentioned in [LDS07], where a Phase-Vocoder-based method is described for excitation extraction. In our case, the original spectrum is that of a unit-stepped windowed waveform (refer to Section 3.2.2), and the synthesized spectrum that is subtracted from the original spectrum is also unit-stepped, which guarantees that all samples before ν are, indeed, going to be zero. It should be mentioned, however, that care must then be taken to avoid a discontinuity in the waveform between samples ν and $\nu - 1$, if, in the input waveform, the samples before ν were not really zero – which is generally the case. A simple cross-fade can then be operated, starting at ν and for a few tens of samples thereafter. Figure 4.14 gives an illustration of how cross-fading the input and output prevent the click of a unit-stepped attack.

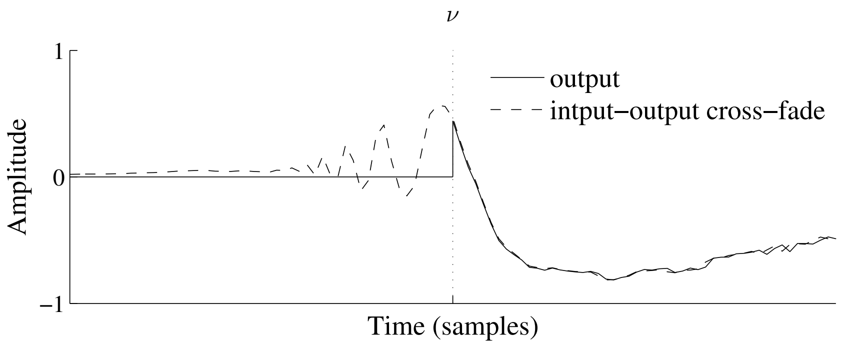

Figure 4.14: To avoid the discontinuity inherent to the unit-step modeling of the attack, the output should be cross-faded with the input over a few samples after ν.

4.5 Phantom partials: the issue of overlapping partials; their cancelation.

In this last section on the evaluation of the string extraction method, we compare the approach that accounts for the phantom partials by the means of the algorithm introduced in Section 3.3.3, with a standard Median-Adjustive Trajectory (MAT) peak-picking approach. We recall that the MAT is a bottom-up peak detection approach for string tones showing non-negligible inharmonicity. As the detection progresses upwards in frequency, the number of measured peaks increases, and thereby the number of estimates of the Inharmonicity Coefficient (IC) and Fundamental Frequency (FF) that can be made. After each partial measurement, the sets of IC and FF estimates is augmented, and the median of each sets is updated and used to direct the search for the next peak – hence the name of the method.

Overall, the results in this section will show that our method handles well phantom partials that are sufficiently resolved. These tests were nevertheless useful in pointing out, however, that not all situations of overlap had been foreseen in our algorithmic design. Our algorithm looks for the bulge caused by the presence of a partial, as opposed to looking for a peak. This allows the detection of a partial even when, in a situation of overlap, it is dominated by another partial of greater magnitude and for this reason cannot emerge as a peak. However, there are situations where partials are even closer, and a bulge appears for the dominated partial only periodically, in a pattern that repeats every two frames or more. An example of this is given in the upper left spectrogram of Figure 4.15, spectrogram of a Spanish guitar bass E after spectral processing. It can be seen that a couple of overlapping partials have been alternately recognised as the one same partial (in the frames where two sidelobes can be seen) or two distinct partials (in the frames where cancelation is successful). The reason for this alternating shape of the magnitude "mound" formed by the two overlapping partials is their phase difference from one frame to the next, which, given their closeness in frequency, is slowly incremented. Figure 4.16 demonstrates this phenomenon.

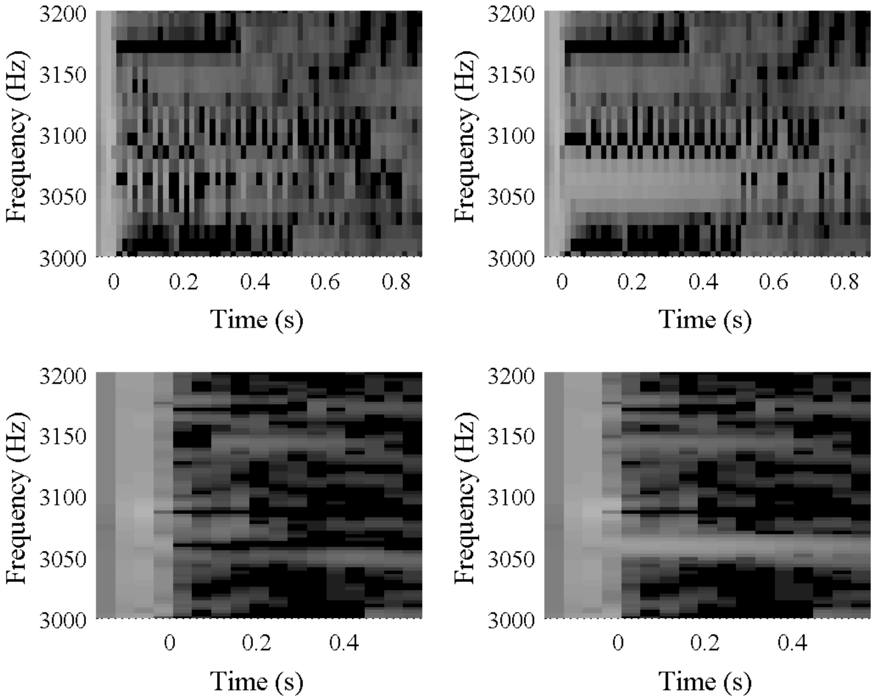

Figure 4.15: Spanish guitar open bass E (MIDI note 40). Left column: both transverse and phantom partials are sought and canceled; right column: only transverse partials are canceled. Upper row: the window length is set to the minimal length; lower row: the window length is thrice the minimal length. The “burbling” in the upper row is a confusion of our algorithm, caused by a misleading situation of overlap.

Here, two partials were synthesized, one (the leftmost) half the amplitude of the other. The difference of their frequencies is only a quarter of the minimal required for good resolution (c.f. Section 2.2). Their sum is windowed and Fourier-transformed eight times (panels numbered from 1 to 8), and each time their phase difference (at the time sample corresponding to the centre of the analysis) is incremented by $\pi/2$, beginning with a phase difference of

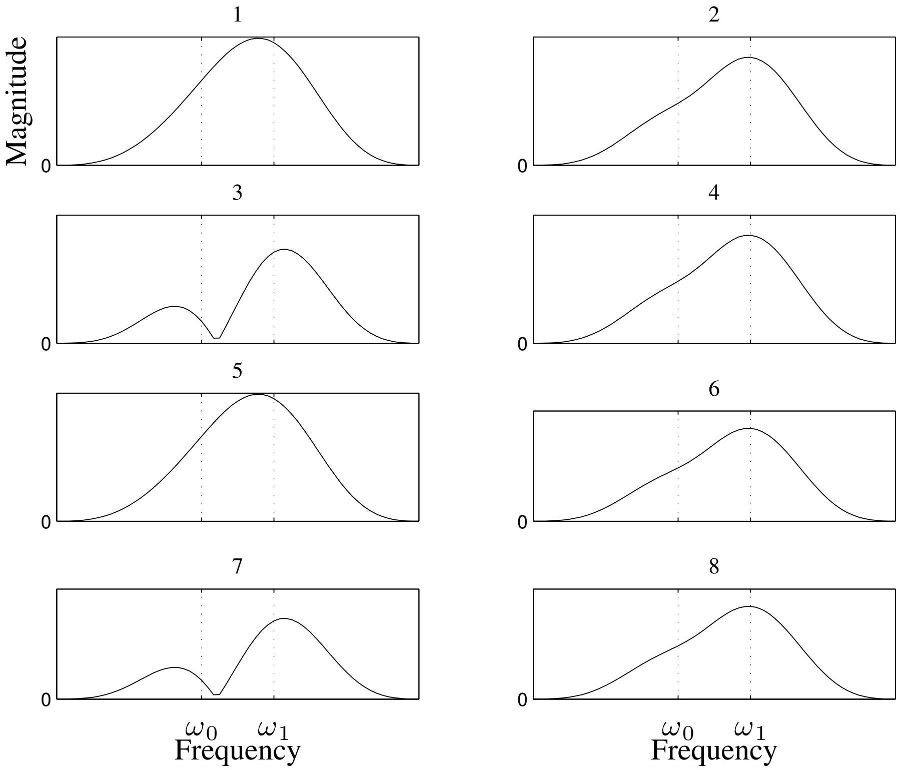

Figure 4.16: "Burbling", caused by the slowly-incrementing phase difference in the sliding analysis of overlapping partials, of frequencies ω_1 and ω_2. In subplots 1 and 5, the phase of the two partials is equal at the centre of the analysis, and in 3 and 7, is is opposite.

0 (frame 1). This figure gives an impression for the difficulty introduced by overlapping partials in our FFT-based detection and cancelation of partials. In the frames where the phase difference is nil, the two partials merge so well that it is even impossible to detect a bulge for the partial of lesser amplitude. This bulge emerges when the phase difference is of $\pi/2$, and then becomes a peak when the partials are opposite in phase. Periodically, our analysis will therefore alternate between modeling this magnitude mound as one or two

partials, causing the burbling seen in the topmost plots of Figure 4.15. Also, notice the oscillation in frequency of the magnitude peaks about the actual frequencies ω_1 and ω_2, illustrative of how unreliable frequency measurements might become.

Back to the broader picture of actual spectrograms, we have tried to use a peak detection and cancelation process that, contrastingly, seeks transverse partials only, thinking that a simpler approach might prove more robust in such delicate situations. The comparative spectrogram is found in the upper right panel. The reader will notice that the burbling is *identical* for the uppermost partial in either spectrogram, which shows that this problem affects straightforward peak detections in the same way, only it is more visible in the case of our algorithm because the cancelation of two partials is involved. In our opinion, the reason for the persistence of such difficulties is that we have been transposing a one-partial approach to detection and cancelation onto a situation where two overlapping partials make up the spectral data. A generalisation of the detection model to two overlapping partials might be a wiser idea than trying to fix some method that is, from the start, not completely appropriate. This idea is not novel, as research has recently emerged in this direction [Fab10].

The bottom two spectrograms of Figure 4.15 repeat the experiment, but with an increment of the window length to three times the minimal length (which is $2(P+1)$ times the fundamental period of the tone). The overlapping situation here has disappeared, and the burbling with it. Augmenting the window length nevertheless has the inconvenience that, for a same onset overlap ratio (here, 1/5), the pitch of the tone might become more audible at the instant of the attack (which will be a problem so long as the onset overlap problem explained in Section 4.4 is not solved in a manner that is satisfying in all quality and robustness respects), which is what happens in this example; the reader will find the sound files 27, 28, 29 and 30, corresponding to the upper left, upper right, bottom left and bottom right spectrograms, respectively. It should be noticed, however, that the string extraction is more thorough when phantom partials are accounted for, even in this Spanish guitar example, where the phantom partials are relatively scarce and faint. The guitar tone used for these spectrograms was in fact the same as that already used in Figure 1.11, Section 1.2.7. In that figure, the reader can see that phantom partials in acoustic guitar tones are yet stronger and more numerous. As a means of illustration, an example of the string extraction process with and without accounting for the phantom partials is given in figures 4.17 and 4.18, corresponding to sound files 20 (transverse and phantom cancelation) and 21 (only transverse cancelation). Finally, piano tones present yet a greater density of phantom partials, in their lower region as well as their mid-range, fact illustrated in figures 4.19 and 4.20 (sound files 22 and 23). Both spectrograms and waveform show that string extraction should, indeed, account for phantom partials. Without accounting for them, phantom partials visibly stand out in the output spectrograms, and the output waveforms still show significantly more sinusoidal energy than they should. The difference is also clearly audible in the sound files.

This chapter's first aim was to test the method of string extraction intro-

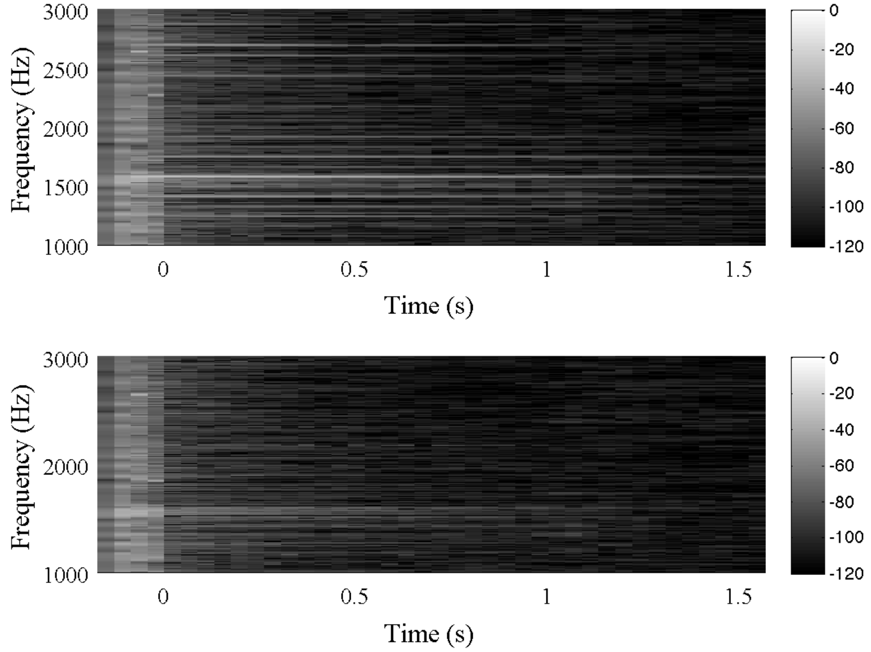

Figure 4.17: Acoustic guitar E2 (MIDI note 40) string extraction spectrograms omitting (top) and accounting for (bottom) the phantom partials.

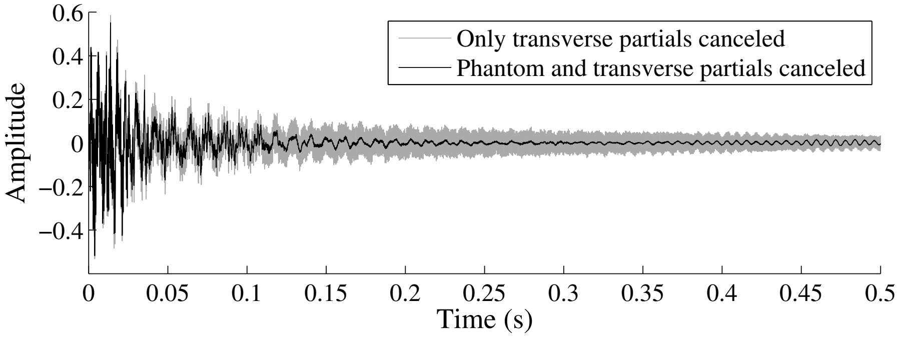

Figure 4.18: Acoustic guitar E2 (MIDI note 40) string extraction waveform, omitting (lighter shade) and accounting for (darker shade) the phantom partials.

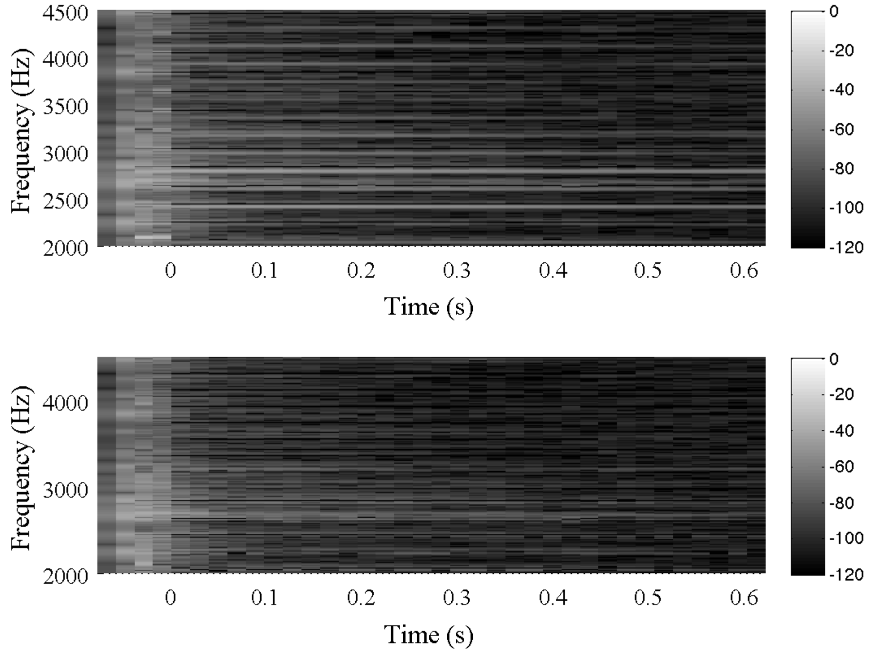

Figure 4.19: Grand piano F#4 (MIDI note 54) string extraction spectrograms omitting (top) and accounting for (bottom) the phantom partials.

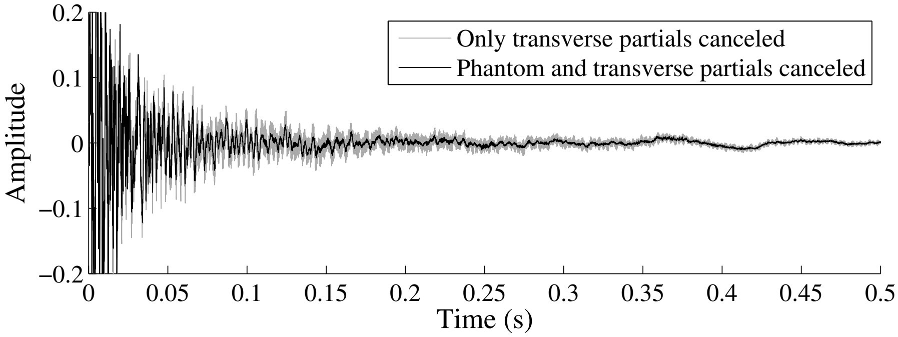

Figure 4.20: Grand piano F#4 (MIDI note 54) string extraction waveform, omitting (lighter shade) and accounting for (darker shade) the phantom partials.

duced and described in this thesis. Its potential was shown through successful examples, where no artifacts, trace of the original pitch or ringing of the object string could be heard in the output, and this, for instruments of contrasting character (viola *pizzicato*, electric and acoustic guitars, double bass, harpsichord). Another aim of this chapter was to show separately the contribution of the different original ideas spanning this work: the exponential-amplitude generalisation of the CSPE; the complete cancelation of a partial through the synthesis of only a few frequency-domain samples, possible thanks to the formulation of an analytical model; the modeling of the string's attack with a unit-step product; and the detection and cancelation of phantom partials, along with the main series of transverse partials, possible thanks to Median-Adjustive Trajectories.

Imperfections of the method were also highlighted and discussed, and the reader will observe that these all arise because of cross-interference of partials. This cross-interference can be due to excessive spectral leakage in onset-overlapping frames, or to phantom partials overlapping with the normal transverse partials. These issues are not trivial, and could not be solved within the scope of this thesis, although some ideas were suggested that might be explored more in depth in future work.

Chapter 5

Conclusion

This conclusion is the opportunity to give a recapitulation of the aims, organisation and contributions of this thesis, as well as indicating future work in the continuation of this thesis.

5.1 Aims

This thesis proposes a Phase-Vocoder, subtractive approach to monophonic string extraction.

String extraction is the sound processing paradigm that consists of decomposing the waveform produced by a plucked- or hit-string instrument into the resonances of its strings on the one hand, and all other sound components – stochastic and deterministic together – otherwise produced. These include the indirect responses of the instrument's body to the excitation of the strings, which in turn consist of bursts of energy, and potentially resonances of the body as well. Components other than the resonances of the string may also include environmental noises, such as the noise floor of the recording, the buzz of an electric guitar amplifier, or accidental noises.

The paradigm stated here is the string extraction paradigm in the polyphonic sense. The initial aim of this thesis was to propose an automated method for the excitation extraction of monophonic string tones for subsequent use in Commuted Waveguide Synthesis (CWGS) [KVJ93, Smi93, LDS07]. The contemplation of a real-time approach, progressively brought within reach by the achievements made throughout this thesis, eventually stimulated the formulation of the concept of string extraction. The FFT nature of this thesis' approach to monophonic string extraction, however, appears futile in the context of polyphonic string extraction, where it cannot be assumed that the partials are well resolved. Resorting to higher-resolution methods such as ESPRIT [RPK86, DBR06] seems not to be an appropriate response inasmuch, given the "blindness" with which the method models the sinusoidal structure of the input, preventing the use of our string's time-frequency model for extraction and for physics-based musical effects. In this regard, string extraction comes across as a paradigm whose monophonic reduction alone can be tackled with presently existing means. This much was shown in this thesis.

5.2 Organisation

The main body of this thesis was organised in five chapters. In Chapter 1, an analytical model of a string's vibrations was developed from physical analysis and empirical observations and modeling. The result of this model is summarised in Table 1.2. This model is indicative of the time-frequency sinusoidal structure of a string. The vertical aspect of this model is helpful in inspiring methods for the location and identification of partials in a short-time magnitude spectrum. The horizontal aspect, in turn, can be used to distinguish what aspects of a string's sinusoids it is necessary to take into account in short-time measurements and modeling.

The method for string extraction advanced in this thesis is based on a Phase Vocoder scheme, which uses a Short-Time Fourier Transform representation of the signal as an environment for analysis and processing. The sinusoids are identified, measured and subtracted in the frequency domain, where their properties are intrinsically dependent on the properties of the analytical window used during the granulation process. For this reason, Chapter 2 opened with an extensive discussion on windows, on their frequency-domain properties, and also on their time-domain properties. To complete the description of the frequency-domain image that the FFT was going to provide us with, the spectral properties inherent to discrete signals were also discussed. At that stage, a sinusoidal model for the string's resonances had been formulated, and the spectral representation of such a model had been described. The chapter could thereon close with the design of a novel, appropriated method for the sinusoidal analysis of string partials: the Complex Spectral Phase Magnitude Evolution (CSPME) method.

Following the low-level discussion of Chapter 2, Chapter 3 could initiate a higher-level description of the proposed string-extraction method. The Phase Vocoder scheme was described in its time-domain organisation. The idea of resorting to frequency-domain subtraction for the cancelation of partials could be introduced, and the novel possibilities that came with it: the frequency-domain approximation of an entire partial with the synthesis of the few bins of its main lobe with an analytical model, and the clearance of a phantom partial's lobe from a dominating, overlapping transverse partial. For the detection and identification of the partials, a linear-complexity simplification of the Median-Adjustive Trajectories method was given, and shown to be more robust than the intuitive approach of local-linearisation of an inharmonic series. Finally, the problem of frames overlapping with the attack of the tone was proposed a solution for with the modeling of the attack as a product of the string's steady-state with a unit-step function.

Chapter 4 was a discussion articulated around tests of various nature. General testing was first used to demonstrate the overall success of the method, and examples were also used as a means of discerning the subtle difference between string extraction and excitation extraction. The various lower-level ideas used in the overall string extraction method were tried separately, and all were shown to contribute in significant ways to its success, qualitatively or computationally: the exponential-amplitude generalisation of the CSPME, the Fourier-series approximation of the partials' spectra, the unit-step modeling of the attack, and the detection and cancelation of the phantom series.

5.3 Contributions

These sub-methods were inspired for the purpose of this very thesis. However, their applications range beyond the scope of string extraction. Median-Adjustive Trajectories return fundamental frequency and inharmonicity coefficient estimates of unprecedented accuracy [HWTL09], and throughout the numerous tests of this thesis have proven their reliability for the detection and identification of partials, both phantom and transverse, in inharmonic spectra. It therefore represents a valuable tool for such analyses of the inharmonicity of string tones as those carried in [FBS62]. In fact, the use for a method of such accuracy has been demonstrated within this very thesis, where it permitted to outline trends in the inharmonicity coefficient of tension-modulated tones that had not been accounted for to date. This then led to the development of semi-empirical, semi-physics-based exponential-plus-constant models for the time-evolution of fundamental frequency and inharmonicity coefficients [HTL10].

In the context of sinusoidal analysis, the CSPME, generalisation of the Complex Spectral Phase Evolution (CSPE) method [SG06], is a novel means of estimating the frequency and exponential amplitude of a sinusoid accurately which contrasts, by its simplicity, with other exponential-amplitude generalisations [AS05, MD08]. The decay-rate profile of a spectrum can thereby easily be obtained, and save a lot of computation and effort for the estimation of “gain spectra” for the calibration of digital waveguides [KVJ93].

These are the innovations of this thesis whose benefit to the audio signal processing community is evident. Beside such contributions, it is hoped that this thesis will be appreciated for the pedagogical concern which governed its writing. The essential of the material was learned and discovered by the author during the genesis of this thesis. The reader was therefore considered like a companion in this learning process. Developments were initiated from first principles which led to the various rules and formulae, mostly known in the literature but whose origin often remains buried in the literature, sometimes to the point that they can hardly be retrieved. For example, this approach allowed the unified formulation of the series of damped, stiff strings (1.29), of which it could be seen that the inharmonic-series expression generally seen (1.33) is an approximation. Elsewhere, the constant-sum property of power-raised cosine window was proven, and the basis for this proof yielded the minimum-overlap factor rule in terms of the order of the window and the power it is raised to. Likewise, an example could be made of the Fourier-series approximation of the string partials used for computationally-efficient cancelation, whose base stems from the extensive development of Chapter 2 on windowing.

5.4 Future Work

The implementation of the method in a programming language supporting real-time input, output and or real-time processing capabilities should probably come first in the list of priorities for future work. C++ seems like the most suitable language, all the more that it is the language the VST (Virtual Studio Technology) Application Programming Interface is written with. VST applications (*plugins*) can be imported in most Digital Audio Workstations to work in real time as virtual instruments or sound effects. To run in real-time, our method should be augmented with an onset-detection algorithm [BDA$^+$05, GLT11], and possibly a silence-detection algorithm as well. Also, an appropriate pitch estimator should be devised or chosen. The principle of autocorrelation-based pitch estimation was shown in Section 3.1.1, but numerous other approaches might be considered [Ger03]. The estimate does not have to be very accurate, as its sole purpose is to determine what minimal length the analysis window should be. However, so as to keep the latency of our string extraction method to a low, it should be reactive. Furthermore, so as to avoid combining the latency of the pitch detection and of the FFT analysis, a bank of Phase-Vocoder engines of different window lengths could be set in parallel, and upon the estimation of the input's pitch, the string extraction process could be assigned to the sliding analysis that turns out to have the most appropriate length. Say, to take a simple example, that the expected pitch range of the processed instrument went from 100Hz to 800Hz, sliding analyses a, b and c could run in parallel, with respective window lengths 1/100, 1/200 and 1/400 seconds. Then, the string extraction of a note which turns out to be of frequency 240Hz would be assigned to sliding analysis b, and that of a note of 700Hz, to sliding analysis c, and so on. How finely the pitch range of the instrument should be divided should take into account the computational cost of the string extraction process, the capabilities of the host machine, and the extent to which our method can successfully extract strings as the length of the window exceeds the minimal length required.

Roughing out excitation extraction, this thesis' work has uncovered, between deterministic-stochastic decomposition and physical modeling, the paradigm of string extraction. Its approach to undertaking monophonic string extraction was tailored to the problem's profile. The delineated arrangement of monophonic spectra, however, shall be disrupted in a magma of peaks, bulges and mounds when string extraction goes polyphonic, and all assumptions this thesis was building on shall crumble. From this thesis, indeed, has risen a paradigm which the most recent advances seem insufficient to tackle.

Appendix A

Four Essential Transforms

Fourier analysis tools are used extensively throughout this thesis in many situations and for many purposes. A unified syntax across all four transforms outlines their similarity. Essentially, the same idea of the projection of a function onto complex exponential basis vectors is applied to four categories of signals : continuous and infinite, continuous and periodic, discrete and infinite, and discrete and periodic. (We oppose infinite and periodic.) It is interesting to note that a signal that is of continuous/discrete domain transforms to a signal that is of infinite/periodic domain.

A.1 Fourier Transform

The Fourier transform takes a signal that is continuous and extends infinitely in time, to output a signal that is continuous and extends infinitely in frequency :

$$\text{FT}\{f(t)\} = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt = F(\omega) \tag{A.1}$$

$$\text{FT}^{-1}\{F(\omega)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{j\omega t}\,d\omega = f(t) \tag{A.2}$$

t is in seconds, and ω, in radians per second.

FT : $\mathbb{C} \to \mathbb{C}$. When $f(t) \in \mathbb{R}$, the Fourier transform is also susceptible of yielding a complex spectrum, however, with one specificity : the negative-frequency half is the complex conjugate of the positive-frequency half, i.e. $F(-\omega) = F^*(\omega)$. This is easily proven considering that $(f(t))^* = f(t) \iff f(t) \in \mathbb{R}$, which implies that

$$\begin{aligned} F(-\omega) &= \int_{-\infty}^{\infty} (f(t))^* \left(e^{-j\omega t}\right)^* dt \\ &= \left(\int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt\right)^* \\ &= F^*(\omega). \end{aligned}$$

This property is found in the three following transforms as well.

A.2 Fourier Series

The Fourier series looks at signal that is continuous in time and periodic in T seconds (i.e. $f(t) = f(\text{mod}(t,T))$).

$$\text{FS}\{f(t)\} = \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)e^{-jk2\pi t/T}dt = F[k] \tag{A.3}$$

$$\mathrm{FS}^{-1}\{F[k]\} = \frac{1}{T}\sum_{k=-\infty}^{\infty} F[k]e^{jk2\pi t/T} = f(t) \tag{A.4}$$

Its output is discrete in frequency, but of infinite extent (i.e. range(k)= $\mathbb{Z}$). k is the frequency *bin index* (unit-less) and t the time-variable, in seconds.

A times it is useful to look at the Fourier series of a segment of a function $f(t)$ which is not periodic in T. However, within the context of this transform, $f(t)$ somehow *becomes* periodic. To prove this point, let us express this last term side in terms of its inverse Fourier series :

$$\begin{aligned}
f(\mathrm{mod}(t,T)) &= f(t - T\lfloor t/T \rfloor) \\
&= \sum_{k=-\infty}^{\infty} F[k]e^{jk2\pi(t-T\lfloor t/T\rfloor)/T} \\
&= \sum_{k=-\infty}^{\infty} F[k]e^{jk2\pi t/T}e^{jk2\pi\lfloor t/T\rfloor} \\
&= \sum_{k=-\infty}^{\infty} F[k]e^{jk2\pi t/T} \\
&= f(t).
\end{aligned}$$

As can be seen here, this phenomenon is related to the process of summation over the integer k, as $k\lfloor t/T\rfloor$ is necessarily an integer and thus $e^{jk2\pi\lfloor t/T\rfloor} = 1$. Anytime, in the context of Fourier analysis, summation is used to transit from one domain to the other, the result is going to be periodic - this is true for the forward Discrete-Time Fourier Transform (DTFT) and both the forward and inverse Discrete Fourier Transform (DFT) as well.

A.3 Discrete-Time Fourier Transform

The Discrete-Time Fourier Transform (DTFT) takes in a signal that is discrete in time and of infinite extent, to output a signal that is continuous in frequency and periodic in 2π, i.e. $F(\omega) = F(\mathrm{mod}(\omega, 2\pi))$.

$$\mathrm{Z}\{f[n]\} = \sum_{n=-\infty}^{\infty} f[n]e^{-j\omega n} = F(\omega) \tag{A.5}$$

$$\mathrm{Z}^{-1}\{F(\omega)\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} F(\omega)e^{j\omega n}d\omega = f[n] \tag{A.6}$$

Now our time index, $n \in \mathbb{Z}$, is in *samples*. As the argument to the complex exponential must be in radians, ω is now in radians per sample.

Like the Fourier transform, the DTFT is symmetric in frequency about 0 when the analysed signal is real. This, combined with periodicity, makes $F(\omega) = F^*(-\bmod(\omega, 2\pi))$ (for $f[n] \in \mathbb{R}$)). We set to prove this here :

$$\begin{aligned} F(-\bmod(\omega, 2\pi)) &= F(-\omega + 2\pi\lfloor\omega/2\pi\rfloor) \\ &= \sum_{n=-\infty}^{\infty} f[n]e^{j\omega n}e^{-j2\pi\lfloor\omega/2\pi\rfloor n} \\ &= \sum_{n=-\infty}^{\infty} f[n]e^{j\omega n}. \end{aligned}$$

Now we follow the same idea as in section A.1 : if $f[n] \in \mathbb{R}$, then $(f[n])^* =$

$f[n]$, and

$$\begin{aligned} F(-\bmod(\omega, 2\pi)) &= \sum_{n=-\infty}^{\infty} (f[n])^* \left(e^{-j\omega n}\right)^* \\ &= \left(\sum_{n=-\infty}^{\infty} f[n] e^{-j\omega n}\right)^* \\ &= F^*(\omega). \end{aligned}$$

This also applies to the DFT.

The DTFT is equivalent to the Z-transform, $F(\omega) = \frac{1}{2\pi}\sum_{n=-\infty}^{\infty} f[n]z^{-n}$, when $e^{j\omega}$ is substituted in place of z. In this thesis, unless specified otherwise, we will always look at z as $e^{j\omega}$ and thus use A.5 for the Z-transform in place of the more general formulation.

A.4 Discrete Fourier Transform

The Discrete Fourier Transform (DFT) produces, from a signal that is discrete and time and periodic in N samples, a signal that, likewise, is discrete in frequency and periodic in N bins.

$$\mathrm{DFT}\{f[n]\} = \sum_{n=0}^{N-1} f[n] e^{-jk2\pi n/N} = F[k] \tag{A.7}$$

$$\mathrm{DFT}^{-1}\{F[k]\} = \frac{1}{N}\sum_{k=0}^{N-1} F[k] e^{jk2\pi n/N} = f[n] \tag{A.8}$$

Here, k is the frequency *bin index* (unit-less), and n, the sample number (in samples).

We may use DFT{.} to denote the DFT.

A.5 Transforms of delta function, complex exponentials and real signals

The table found in this section is a recapitulation of the various facts stated above. We also take advantage of this place and time to list two transforms recurrent in this thesis : those of the delta function and the complex exponential.

FT		FS	
$f(t)$	$F(\omega)$	$f(t)$	$F[k]$
$\delta(t-t_0)$	$e^{-j\omega t_0}$	$\delta(\mathrm{mod}(t-t_0,T))$	$e^{-jk2\pi t_0/T}$
$e^{j\omega_0 t}$	$2\pi\delta(\omega-\omega_0)$	$e^{j2\pi pt/T}$	$T\delta[k-p]$
N.A.	N.A.	$f(t)=f(\mathrm{mod}(t,T))$	N.A.
$\in\mathbb{R}$	$F(\omega)=F^*(-\omega)$	$\in\mathbb{R}$	$F[k]=F^*[-k]$
DTFT		DFT	
$f[n]$	$F(\omega)$	$f[n]$	$F[k]$
$\delta[n-m]$	$e^{-j\omega m}$	$\delta[\mathrm{mod}(n-m,N)]$	$e^{-jk2\pi n/N}$
$e^{j\omega_0 n}$	$2\pi\delta(\omega-\omega_0)$	$e^{j2\pi pn/N}$	$N\delta[k-p]$
N.A.	$F(\omega)=F(\mathrm{mod}(\omega,2\pi))$	$f[n]=f[\mathrm{mod}(n,N)]$	$F[k]=F[\mathrm{mod}(k,N)]$
$\in\mathbb{R}$	$F(\omega)=F^*(-\bmod(\omega,2\pi))$	$\in\mathbb{R}$	$F[k]=F^*[-\bmod(k,N)]$

Table A.1: Important transforms and transform properties

Appendix B

Convolution

Convolution is an operation whereby a signal f is issued from the manipulation of two other signals, g and h. f inherits properties of both signals, as will be seen later in the equivalence between time-domain convolution and frequency-domain pointwise multiplication.

Four different cases of convolution are described here : continuous-time, standard ; continuous-time, circular ; discrete-time, standard ; and discrete-time, circular. Whether discrete or continuous, standard convolution between g and h may be denoted $g * h$, and circular convolution, $g \circledast h$.

B.1 Continuous-time

$$(g * h)(t) = \int_{-\infty}^{\infty} g(u)h(t-u)du \tag{B.1}$$

$$(g \circledast h)(t) = \int_{0}^{t} g(u)h(t-u)du + \int_{t}^{T} g(u)h(t+T-u)du \tag{B.2}$$

B.2 Discrete-time

The following cases of convolution are the direct adaptation of the previous two to discrete-time. The standard convolution of discrete signals $g[n]$ and $h[n]$, $n \in \mathbb{Z}$, is

$$g[n] * h[n] = \sum_{m=-\infty}^{\infty} g[m]h[n-m]. \tag{B.3}$$

The circular convolution of $g[n]$ and $h[n]$ over the interval $\mathbb{Z} \cap [0, N-1]$ is

$$g[n] \circledast h[n] = \sum_{m=0}^{n} g[m]h[n-m] + \sum_{m=n+1}^{N-1} g[m]h[n+N-m]. \tag{B.4}$$

The reader may have found an alternative way of writing discrete circular convolution, as in $\sum_{m=0}^{N-1} g[m]h[n-m]$. B.4 is preferred given that, in the case where g and h are digital arrays of length N, it avoids out-of-bound referencing.

B.3 Algebraic Properties

Three algebraic properties of convolution which are not necessarily obvious are those of commutativity, distributivity and associativity. Next we state those properties in turn, and for each give a proof. The proofs are all given for continuous-time, standard convolution, but hold for all cases of convolution seen above.

B.3.1 Commutativity

To prove that

$$f * g = g * f, \tag{B.5}$$

we proceed as follows :

$$\begin{aligned}(f * g)(t) &= \int_{-\infty}^{\infty} f(u)g(t-u)du \\ = (g * f)(t) &= \int_{-\infty}^{\infty} g(u)f(t-u)du\end{aligned}$$

Let $v = t - u$, and therefore, $u = t - v$, $dv = -du$, $u \to -\infty \Rightarrow v \to \infty$. The substitution makes it obvious that the equality holds.

B.3.2 Distributivity

To prove that

$$f * (g + h) = f * g + f * h, \tag{B.6}$$

let $x(t) = g(t) + h(t)$. Then,

$$\begin{aligned}(f * x)(t) &= \int_{-\infty}^{\infty} f(u)x(t-u)du \\ &= \int_{-\infty}^{\infty} f(u)(g(t-u) + h(t-u))du \\ &= \int_{-\infty}^{\infty} f(u)g(t-u)du + \int_{-\infty}^{\infty} f(u)h(t-u)du \\ &= f * g + f * h,\end{aligned}$$

consistently with B.6.

B.3.3 Associativity

$$(f * g) * h = f * (g * h) \tag{B.7}$$

$$\begin{aligned}(f * g) * h &= \int_{-\infty}^{\infty} (f * g)(u) h(t-u) du \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(v) g(u-v) dv \right) h(t-u) du \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(v) g(u-v) h(t-u) dv du\end{aligned}$$

Looking at $f(v)g(u-v)h(t-u) = x(u,v)$, we use Fubini's theorem, which states that $\int_a^b \int_c^d x(u,v) du dv = \int_c^d \int_a^b x(u,v) dv du$, and write

$$(f * g) * h = \int_{-\infty}^{\infty} f(v) \int_{-\infty}^{\infty} g(u-v) h(t-u) du dv$$

On the other hand, we have

$$f * (f * h) = \int_{-\infty}^{\infty} f(v)(g * h)(t-v) dv,$$

which, equated with $(f * g) * h$, implies that

$$
\begin{aligned}
(g * h)(t - v) &= \int_{-\infty}^{\infty} g(w)h(t - v - w)dw \\
&= \int_{-\infty}^{\infty} g(u - v)h(t - u)du.
\end{aligned}
$$

Let $w = u - v$, and it it becomes obvious that the identity holds.

Appendix C

Time-Frequency Domain Operation Equivalences

C.1 Time and Frequency Shifts

In this section we examine the effect of time-shifting a signal onto its spectrum, and, conversely, the effect of frequency-shifting signal's spectrum onto its time-domain representation. The mathematical development is given in detail in the first case (that of the Fourier transform), but is spared to the reader for the other transforms, as the method is identical.

C.1.1 On the Fourier Transform

Let

$$G(\omega) = \int_{-\infty}^{\infty} f(t+\tau)e^{-j\omega t}dt$$

be the Fourier Transform of the signal $f(t)$ time-shifted by τ seconds. We want to express $G(\omega)$ in function of $\mathrm{FT}\{f(t)\} = F(\omega)$, to see what effect time-shifting has on the spectrum.

Let $u = t + \tau$, and equivalently $du = dt$, to get

$$\begin{aligned} G(\omega) &= \int_{-\infty}^{\infty} f(u) e^{-j\omega(u-\tau)} du \\ &= e^{j\tau\omega} \int_{-\infty}^{\infty} f(u) e^{-j\omega u} du \\ &= e^{j\tau\omega} F(\omega) \end{aligned}$$

A time shift τ on the signal is hereby recognised to carry a frequency-dependent phase shift $\tau\omega$ in its frequency-domain representation. More concisely, we write

$$\mathrm{FT}\{f(t+\tau)\} = e^{j\tau\omega} F(\omega). \tag{C.1}$$

We proceed similarly to observe the effect of frequency-shifting the spectrum on the time-domain signal. Let

$$g(t) = \int_{-\infty}^{\infty} F(\omega + \theta) e^{j\omega t} d\omega,$$

the time-domain representation of the spectrum $F(\omega)$ frequency-shifted by θ radians per second. We want to express $g(t)$ in terms of $\mathrm{FT}^{-1}\{F(\omega)\} = f(t)$, to see what effect shifting in the frequency domain has on the time-domain.

Let $\phi = \omega + \theta$. This substitution yields

$$\begin{aligned} g(t) &= \int_{-\infty}^{\infty} F(\phi) e^{j(\phi-\theta)t} d\phi \\ &= e^{-j\theta t} \int_{-\infty}^{\infty} F(\phi) e^{j\phi t} d\phi \\ &= e^{-j\theta t} f(t). \end{aligned}$$

Frequency-shifting a spectrum by θ radians per second has the effect of phase-modulating the time-domain signal by $-\theta t$ radians. More concisely, we write that

$$\mathrm{FT}^{-1}\{F(\omega+\theta)\} = e^{-j\theta t} f(t). \tag{C.2}$$

When $f(t)$ is a real-valued signal, $g(t)$ becomes complex-valued. This can be understood when looking at real signals like at complex signals whose phase constantly equals zero.

C.1.2 On the other transforms

Applying the same technique of substitution on the Fourier series, Z-transform and DFT, we get the following results :

$$\mathrm{FS}\{f(t+\tau)\} = e^{jk2\pi\tau/T} F[k], \tag{C.3}$$

$$\mathrm{FS}^{-1}\{F[k+l]\} = e^{-jl2\pi t/T} f(t), \tag{C.4}$$

$$\mathrm{Z}\{f[n+m]\} = e^{j\omega m}F(\omega), \tag{C.5}$$

$$\mathrm{Z}^{-1}\{F(\omega+\theta)\} = e^{-j\theta n}f[n], \tag{C.6}$$

$$\mathrm{DFT}\{f[n+m]\} = e^{jk2\pi m/N}F[k] \tag{C.7}$$

and

$$\mathrm{DFT}^{-1}\{F[k+l]\} = e^{-jl2\pi n/N}f[n]. \tag{C.8}$$

C.2 Convolution Theorem

The convolution theorem states that the frequency transform of the convolution of two signals equals the product of the frequency transforms of each signal. Conversely, it also states that the frequency transforms of the product of two signals equals a constant times the convolution between the frequency transforms of each signal. In the latter case, the constant stems from the scaling present in the inverse frequency transform.

More concisely, if $F = \mathrm{F}\{f\}$, and $G = \mathrm{F}\{g\}$, (where $\mathrm{F}\{.\}$ is any of the transforms described in A,) we write that

$$\mathrm{F}\{f * g\} = FG, \tag{C.9}$$

and that

$$\mathrm{F}\{fg\} = \frac{1}{c} F * G. \tag{C.10}$$

A proof will be given explicitly here for the Fourier Transform. Following the same steps, the convolution theorem can easily be proven for the other transforms as well, and to avoid redundancy, the equivalences will be then given directly.

C.2.1 Proofs

In the following proofs, we use the short-hand writing $\int_{-\infty}^{\infty} dt \doteq \int dt$.

Consider two continuous signals $f(t)$ and $g(t)$ integrable over $\mathbb{R}$. The Fourier Transform of their convolution is introduced and developed as follows :

$$\begin{aligned}
\mathrm{FT}\{f(t) * g(t)\} &= \int \int f(u) g(t-u) du e^{-j\omega t} dt \\
&= \int \int f(u) g(t-u) e^{-j\omega t} du dt \\
&= \int f(u) \int g(t-u) e^{-j\omega t} dt du,
\end{aligned}$$

the last form obtained using Fubini's theorem. Now we proceed to the substitution $v = t - u$, yielding

$$\begin{aligned}
\mathrm{FT}\{f(t) * g(t)\} &= \int f(u) \int g(v) e^{-j\omega v} e^{-j\omega u} dv du \\
&= \int f(u) e^{-j\omega u} du \int g(v) e^{-j\omega v} dv \\
&= F(\omega) G(\omega),
\end{aligned}$$

which proves C.9, at least in the case of the Fourier transform.

Now we set to prove C.10 :

$$\begin{aligned}
\mathrm{FT}^{-1}\{F(\omega) * G(\omega)\} &= \frac{1}{2\pi} \int \int F(\vartheta) G(\omega - \vartheta) d\vartheta e^{j\omega t} d\omega \\
&= \frac{1}{2\pi} \int \int F(\vartheta) G(\omega - \vartheta) e^{j\omega t} d\vartheta d\omega \\
&= \frac{1}{2\pi} \int F(\vartheta) \int G(\omega - \vartheta) e^{j\omega t} d\omega d\vartheta,
\end{aligned}$$

Letting $\phi = \omega - \vartheta$ yields

$$\begin{aligned}
\mathrm{FT}^{-1}\{F(\omega) * G(\omega)\} &= \frac{1}{2\pi} \int F(\vartheta) \int G(\phi) e^{j\vartheta t} e^{j\phi t} d\phi d\vartheta \\
&= \frac{1}{2\pi} \int F(\vartheta) e^{j\vartheta t} d\vartheta \int G(\phi) e^{j\phi t} d\phi \\
&= 2\pi f(t) g(t),
\end{aligned}$$

which implies that

$$\mathrm{FT}\{f(t) g(t)\} = \frac{1}{2\pi} F(\omega) * G(\omega). \tag{C.11}$$

C.10 for the Fourier transform is proven, with $c = 2\pi$.

C.2.2 Transform-specific equivalences

$$\mathrm{FT}\{f(t) * g(t)\} = F(\omega)G(\omega), \tag{C.12}$$

$$\mathrm{FT}\{f(t)g(t)\} = \frac{1}{2\pi}F(\omega) * G(\omega), \tag{C.13}$$

$$\mathrm{FS}\{f(t) \circledast g(t)\} = F[k]G[k], \tag{C.14}$$

$$\mathrm{FS}\{f(t)g(t)\} = \frac{1}{T}F[k] * G[k], \tag{C.15}$$

$$\mathrm{Z}\{f[n] * g[n]\} = F(\omega)G(\omega), \tag{C.16}$$

$$\mathrm{Z}\{f[n]g[n]\} = \frac{1}{2\pi}F(\omega) \circledast G(\omega), \tag{C.17}$$

and

$$\mathrm{DFT}\{f[n] \circledast g[n]\} = F[k]G[k], \tag{C.18}$$

$$\mathrm{DFT}\{f[n]g[n]\} = \frac{1}{N}F[k] \circledast G[k]. \tag{C.19}$$

Bibliography

[AI04] Mototsugu Abe and Julius O. Smith III. Design criteria for the quadratically interpolated fft method (i): Bias due to interpolation. Technical Report STAN-M-114, CENTER FOR COMPUTER RESEARCH IN MUSIC AND ACOUSTICS DEPARTMENT OF MUSIC, STANFORD UNIVERSITY, October 2004.

[All77] Jont B. Allen. Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3), 1977.

[AS05] M. Abe and III Smith, J.O. Am/fm rate estimation for time-varying sinusoidal modeling. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 3, pages iii/201 – iii/204 Vol. 3, 2005.

[AW03] Jos Arrillaga and Neville R. Watson. *power system harmonics.* John Wiley & Soms Ltd, West Sussex, England, 2003.

[Ban09] Balázs Bank. Energy-based synthesis of tension modulation in strings. In *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09)*, Como, Italy, 2009.

[BDA$^+$05] Juan-Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 13(5), 2005.

[Bou00] Richard Boulanger, editor. *THE CSOUND BOOK*. Massachusetts Institute of Technology, Cambridge, Massachusetts, 2000.

[BS03] Balázs Bank and László Sujbert. Modeling the longitudinal vibration of piano strings. In *Proceedings of the Stockholm Music Acoustics Conference, August 6-9, 2003 (SMAC 03)*, Stockholm, Sweden, 2003.

[CA93] Antoine Chaigne and Anders Askenfelt. Numerical simulations of piano strings. i. a physical model for a struck string using finite difference methods. *Journal of the Acoustical Society of America*, 95(2), 1993.

[Cho73] John Chowning. The synthesis of complex audio spectra by means of frequency modulation. *Journal of the Audio Engineering Society*, 21(7), 1973.

[Cro80] R. E. Crochiere. A weighted overlap-add method of short-time fourier analysis/synthesis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(1), 1980.

[DBR06] Bertrand David, Roland Badeau, and Ga el Richard. Hrhatrac algorithm for spectral line tracking of musical signals. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 2006.

[dCK02] Alain de Cheveign and Hideki Kawahara. Yin, a fundamental frequency estimator for pitch and music. *Journal of the Acoustical Society of America*, 95(2), 2002.

[Ear03] John Eargle. *HANDBOOK OF RECORDING ENGINEERING.* Springer Science+Business Media, Inc., New York, USA, 2003.

[EKHV02] Cumhur Erkut, Matti Karjalainen, Patty Huang, and Vesa Välimäki. Acoustical analysis and model-based sound synthesis of the kantele. *Journal of the Acoustical Society of America*, 112(4), 2002.

[Fab10] Marco Fabiani. Frequency, phase and amplitude estimation of overlapping partials in monaural musical signals. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, 2010.

[FBS62] Harvey Fletcher, E. Donnell Blackham, and Richard Stratton. Quality of piano tones. *Journal of the Acoustical Society of America*, 111(4), 1962.

[FM33] Harvey Fletcher and W. A. Munson. Loudness, its definition, measurement and calculation. *Journal of the Acoustical Society of America*, 5:82–108, 1933.

[FR91] Neville H. Fletcher and Thomas D. Rossing. *The Physics of Musical Instruments*. Springer-Verlag New York Inc., New-York, USA, 1991.

[GA94] A. S. Galembo and A. Askenfelt. Measuring inharmonicity through pitch extraction. *STL-QPSR*, 35(1), 1994.

[GA99] A. S. Galembo and A. Askenfelt. Signal representation and estimation of spectral parameters by inharmonic comb filters with application to the piano. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 7(2), 1999.

[Ger03] David Gerhard. Pitch extraction and fundamental frequency: History and current techniques. Technical Report TR-CS 2003-06, University of Regina, CANADA, November 2003.

[GK96] N Giordano and A. J. Korty. Motion of a piano string: Longitudinal vibrations and the role of the bridge. *Journal of the Acoustical Society of America*, 100(6), 1996.

[GLT11] John Glover, Victor Lazzarini, and Joseph Timoney. Real-time detection of musical onsets with linear prediction and sinusoidal modeling. *EURASIP Journal on Advances in Signal Processing*, 2011(1):68, 2011.

[HAC97] Jr. Harold A. Conklin. Piano strings and phantom partials. *Journal of the Acoustical Society of America*, 102(1), 1997.

[HAC99] Jr Harold A. Conklin. Generation of partials due to nonlinear mixing in a stringed instrument. *Journal of the Acoustical Society of America*, 105(1), 1999.

[Hag03] S. J. Hagen. Exponential decay kinetics in "downhill" protein folding. *Proteins: Structure, Function and Bioinformatics*, 50(1), 2003.

[Har78] Fredric J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1), 1978.

[Har98] Willian M. Hartman. *Signals, Sound, and Sensation.* Springer Science+Business Media, LLC, New-York, USA, 1998.

[HM03] Stephen Hainsworth and Malcolm Macleod. On sinusoidal parameter estimation. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, London, United Kingdom, 2003.

[Hod11] Matthieu Hodgkinson. Exponential-plus-constant fitting based on fourier analysis. In *JIM2011 – 17èmes Journées d'Informatique Musicale*, Saint-Étienne, France, 2011.

[HR71a] Lejaren Hiller and Pierre Ruiz. Synthesizing musical sounds by solving the wave equation for vibrating objects: Part 1. *Journal of the Acoustical Engineering Society*, 19(6), 1971.

[HR71b] Lejaren Hiller and Pierre Ruiz. Synthesizing musical sounds by solving the wave equation for vibrating objects: Part 2. *Journal of the Acoustical Engineering Society*, 19(7), 1971.

[HTL10] Matthieu Hodgkinson, Joseph Timoney, and Victor Lazzarini. A model of partial tracks for tension-modulated steel-string guitar tones. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, 2010.

[HWS87] R. L. Hughson, K. H. Weisiger, and G. D. Swanson. Blood lactate concentration increases as a continuous function in progressive exercise. *Journal of Applied Physiology*, 62(5), 1987.

[HWTL09] Matthieu Hodgkinson, Jian Wang, Joseph Timoney, and Victor Lazzarini. Handling inharmonic series with median-adjustive trajectories. In *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09)*, Como, Italy, 2009.

[III92] Julius O. Smith III. Physical modeling using digital waveguides. *Computer Music Journal*, 16(4), 1992.

[IS87] Julius O. Smith III and Xavier Serra. Parshl: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation. In *Proceedings of the International Computer Music Conference (ICMC-87)*, Tokyo, Japan, 1987.

[KM02] Florian Keiler and Sylvain Marchand. Survey on extraction of sinusoids in stationary sounds. In *Proceedings of the 5th International Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, 2002.

[Kno44] Armand F. Knoblaugh. The clang tone of the pianoforte. *Journal of the Acoustical Society of America*, 16(1), 1944.

[KVJ93] Matti Karjalainen, Vesa Välimäki, and Zoltán Jánosi. Towards high-quality sound synthesis of the guitar and string instruments. In *International Computer Music Conference*, Tokyo, Japan, 1993.

[LD99] J. Laroche and M. Dolson. New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects. In *Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on*, pages 91 –94, 1999.

[LDS07] Nelson Lee, Zhiyao Duan, and Julius O. Smith. Excitation signal extraction for guitar tones. *International Computer Music Association*, 2007.

[Leh08] Heidi-Maria Lehtonen. Analysis of piano tones using an inharmonic inverse comb filter. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland, 2008.

[LEVK01] Mikael Laurson, Cumhur Erkut, Vesa Välimäki, and Mika Kuuskankare. Methods for modeling realistic playing in acoustic guitar synthesis. *Computer Music Journal*, 25(3), 2001.

[LF84] K. A. Legge and N. H. Fletcher. Nonlinear generation of missing modes on a vibrating string. *Journal of the Acoustical Society of America*, 76(1), 1984.

[Mar98] Sylvain Marchand. Improving spectral analysis precision with an enhanced phase vocoder using signal derivatives. In *In Proc. DAFX98 Digital Audio Effects Workshop*, pages 114–118. MIT Press, 1998.

[MD08] Sylvain Marchand and Philippe Depalle. Generalization of the derivative analysis method to non-stationary sinusoidal modeling. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland, 2008.

[MI86] Philip M. Morse and K. Uno Ingard. *Theoretical Acoustics.* Princeton University Press, New Jersey, USA, 1986.

[Moo04] Brian C. J. Moore. *An Introduction to the Psychology of Hearing.* Elsevier Ltd., San Diego, USA, 2004.

[MQ86] Robert J. McAulay and Thomas F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4), 1986.

[MSW83] M. E. McIntyre, R. T. Schumacher, and J. Woodhouse. On the oscillations of musical instruments. *Journal of the Acoustical Society of America*, 74(5), 1983.

[NKS04] J. J. Nichols and P. E. King-Smith. The impact of hydrogel lens settling on the thickness of the tears and contact lens. *Investigative Ophthalmology and Visual Science*, 45(8), 2004.

[Nut81] Albert H. Nuttall. Some windows with very good sidelobe behavior. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29(1), 1981.

[OSB99] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck. *DISCRETE-TIME SIGNAL PROCESSING, 2^{nd} Ed.* Prentice Hall International, Inc., New Jersey, USA, 1999.

[PG79] Martin Piszczalski and Bernard A. Galler. Predicting musical pitch from component frequency ratios. *Journal of the Acoustical Society of America*, 66(3), 1979.

[PL87] Michael Podlesak and Anthony R. Lee. Longitudinal vibrations in piano strings. *Journal of the Acoustical Society of America*, 81(S1), 1987.

[Por81] Michael R. Portnoff. Time-scale modification of speech based on short-time fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3), 1981.

[PT04] Trevor Pinch and Frank Trocco. *Analog Days: The Invention and Impact of the Moog Synthesizer*. Harvard University Press, Massachusetts, USA, 2004.

[Rab12] Rudolf Rabenstein. private communication, 2012.

[Rai00] Daniel R. Raichel. *The Science and Applications of Acoustics*. Springer-Verlag New York Inc., New-York, USA, 2000.

[RLV07] Jukka Rauhala, Heidi-Maria Lehtonen, and Vesa Välimäki. Fast automatic inharmonicity estimation algorithm. *Journal of the Acoustical Society of America*, 121(5), 2007.

[Roa96] Curtis Roads. *the computer music tutorial*. Massachusetts Institute of Technology, Massachusetts, USA, 1996.

[RPK86] R. Roy, A. Paulraj, and T. Kailath. Esprit – a subspace rotation approach to estimation of parameters of cisoids in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(5), 1986.

[Ser89] Xavier Serra. *A SYSTEM FOR SOUND ANALYSIS/TRANSFORMATION/SYNTHESIS BASED ON A DETERMINISTIC PLUS STOCHASTIC DECOMPOSITION*. PhD thesis, Stanford University, 1989.

[SG06] Kevin M. Short and Ricardo A. Garcia. Signal analysis using the complex spectral phase evolution (cspe) method. In *AES 120th Convention*, Paris, France, 2006.

[Sha49] Claude E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1), 1949.

[Sip06] Michael Sipser. *Introduction to the Theory of Computation, 2^{nd} Ed.* Thompson Course Technology, Massachusets, USA, 2006.

[Smi93] Julius O. Smith. Efficient synthesis of stringed musical instruments. In *International Computer Music Conference*, Tokyo, Japan, 1993.

[Smi11] Julius O. Smith. *Physical Audio Signal Processing*. http://-ccrma.stanford.edu/~jos/pasp/, 2011. online book.

[Ste96] Ken Steiglitz. *A Digital Signal Processing Primer*. Addison-Wesley Publishing Company, Inc., Menlo Park, California, USA, 1996.

[TI01] Caroline Traube and Julius O. Smith III. Extracting the fingering and the plucking points on a guitar string from a recording. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics 2001*, 2001.

[TR03] Lutz Trautman and Rudolf Rabenstein. *Digital Sound Synthesis by Physical Modeling Using the Functional Transformation Method*. Kluwer Academic/Plenum Publishers, New York, 2003.

[VHKJ96] Vesa Välimäki, Jyri Huopaniemi, Matti Karjalainen, and Zoltán Jánosi. Physical modeling of plucked string instruments with application to real time sound synthesis. *Journal of the Acoustical Society of America*, 44(5), 1996.

[ZÖ2] Udo Zölzer, editor. *DAFX*. John Wiley and Sons, Ltd, Chichester, England, 2002.