TECHNISCHE UNIVERSITÄT TUHH HAMBURG-HARBURG

Verteilte Systeme

An Architecture for Robust Partial Tracking and Onset Localization in Single Channel Audio Signal Mixes

Ludger Solbach
Distributed Systems Department
Technical University of Hamburg–Harburg
Germany

1998

An Architecture for Robust Partial Tracking and Onset Localization in Single Channel Audio Signal Mixes

Vom Promotionsausschuß der
Technischen Universität Hamburg–Harburg
zur Erlangung des akademischen Grades
Doktor-Ingenieur
genehmigte Dissertation

von Ludger Solbach

aus Hagen/Westfalen

1998

1. Gutachter: Prof. Dr. Friedrich Mayer-Lindenberg

2. Gutachter: Prof. Dr.-Ing. Norbert Fliege

3. Gutachter: Dr.-Ing. habil. Udo Zölzer

Tag der mündlichen Prüfung: 30. Oktober 1998

Danksagung

Herrn Prof. Dr. Friedrich Mayer-Lindenberg gilt mein Dank für Initiierung und unterstützende Begleitung dieser Arbeit. Herrn Prof. Dr.-Ing. Norbert Fliege und Herrn Dr.-Ing. habil. Udo Zölzer danke ich für die Übernahme der Korreferate sowie Herrn Prof. Dr. rer. nat. Ulrich Killat für den Vorsitz des Prüfungsausschusses. Zu großem Dank verpflichtet bin ich Herrn Rolf Wöhrmann für seinen weit über das gewöhnliche Maß hinausgehenden Einsatz bei der programmtechnischen Umsetzung. Weitere Unterstützung wurde mir zuteil durch Herrn Dirk Bächle und Herrn Jörg Kliewer. Ihnen und allen weiteren ständigen, wissenschaftlichen und studentischen Mitarbeitern des Arbeitsbereichs *Verteilte Systeme* der Technischen Universität Hamburg-Harburg, insbesondere Frau Angela Bojarski und Herrn Henry Koplien, danke ich für die freundschaftliche und kollegiale Arbeitsathmosphäre. Herzlichster Dank an Bernd Hage, Andreas Popp und Otto Wohlmuth und für treue Freundschaft!

Nicht zuletzt und ganz besonders danke ich meinen Eltern für die liebevoll unterstützende Begleitung meines bisherigen Lebenswegs.

Contents

Chapter 1

Introduction

The beginnings of research in the field of acoustic signal theory date back to the times of Helmholtz and Ohm in the first half of the 19th century. This research area can be roughly divided into the matters of sound generation, transmission, reception and perception. While the first three categories are mainly covered by physics, sound perception is also of interest in physiology, psychoacoustics, information theory and artificial intelligence.

A commonly accepted term for the analysis of complex auditory sound mixtures is *auditory scene analysis* (ASA) [Bregman, 1990], a term drawing an analogy between visual and auditory scenes. The most important but also most challenging problem for its artificial counterpart, *computational auditory scene analysis* (CASA), is source separation. It is very rare that environmental sounds appear isolated. In most cases they will be superimposed by signals originating from different sources. The human hearing system exhibits an astounding ability to perform source separation, like when following the melody line of a Cello in a string quartet or isolating one speaker out of many at a cocktail party. Of course, this ability also has its limitations. Professional musicians can reliably separate a maximum of three concurrent melodies ([Baumann, 1995], p.102, referring to a work by Huron). Research in the field of CASA is concerned with the objective of meeting or possibly surpassing human performance. The driving force behind this development is the increasing need for improved man–machine and environment–machine interfaces for multimedia applications, machine control and artificial intelligence.

Sound can sometimes be felt through vibrations of the human body and frequently appears in coherence with the visibility of some movement. It triggers subconscious reactions such as feelings and emotions. Sound perception is a process of understanding, categorization and learning. Previously acquired knowledge, listening experiences, visual impressions and other factors play an important role. Apparently, human ASA is a very complex process that cannot be explained by mere bottom–up processing from low to high levels of abstraction. Consequently it is nowadays accepted among experts that straight bottom–up processing is not sufficient to achieve a comparable performance with artificial analysis systems [Slaney, 1998]. Such a system would also have to include a top–down information flow from higher to lower hierarchies of abstraction. Building on works focusing on the representation and acquisition of higher level knowledge in certain domains (e.g. [Tanguiane, 1993] for music), this insight has prompted some researchers to investigate possible benefits of incorporating top–down expectations into CASA (e.g. [Scheirer, 1998]).

Despite the relevance of top–down information flow, there is still room for improvement at the low–level processing stages. There are mostly two factors that have been neglected in the past: the relevance of robustness against background noise and the importance of appropriately dealing with the time–frequency resolution trade–off. Another issue that has been treated like a stepchild is resynthesis. If performed, it was usually for mere a–posteriori validation of the analysis process (e.g. [Cooke, 1993]), without making use of the beneficial potential of online adaptive feedback cancellation. The thesis presented here is mainly concerned with these issues. Its primary goal is to develop and evaluate an automated system for partial parameter extraction from single channel recordings in nonstationary, noisy environment. The time–frequency trade–off and stochastic disturbances are taken into account by the introduction of three different modules, one for signal components concentrated in time, another for those concentrated in frequency and a third one for stochastic components. These modules do not operate independently from each other. Instead they cooperate, each one taking advantage of the insight acquired by its collaborators. In this sense, the system is not strictly bottom–up. It is shown, that feedback loops can already provide a considerable advantage when applied within a low abstraction level of audio signal analysis. Through this design, the dependance of post–processing heuristics, e.g. those used in [Cooke, 1993] for partial track continuation, is circumvented. Although signal–theoretic considerations rather than physiological or psychoacoustic findings were followed as guidelines in the development of the architecture, this approach leads to a system bearing some similarities with properties of the human auditory system, most notably temporal and spectral masking effects.

The structure of this thesis is as follows: underlying concepts are introduced in Chapter 2. This comprises relevant aspects of time–frequency distribution theory, the notions of group delay, time spread, instantaneous amplitude, frequency and bandwidth and the properties of the analytic gammatone filter, which is a key building block of the system architecture. A detailed description of the system follows in Chapter 3. Moreover, comparisons with previous approaches and considerations for multiprocessor implementations are given in this chapter. Finally, results illustrating the capabilities of the proposed architecture are presented in Chapter 4. A list of the symbols and acronyms used throughout this thesis is given in Appendix A. Appendix B describes the AnnaLisa software, by use of which the results in Chapter 4 have been obtained. Further appendices contain brief overviews over further fundamental concepts closely related to the work presented.

Chapter 2

Prerequisites

This chapter introduces fundamental concepts on which the following description of the system architecture will build. Most notably, relevant aspects of time–frequency distribution theory and the notions of group delay, time spread, instantaneous amplitude, frequency and bandwidth are introduced. Moreover, an extensive analysis of the properties of the analytic gammatone filter, a key building block of the architecture, is given.

2.1 Time–Frequency Distributions

The notion of *frequency* is central to our understanding of the physical world. We all have some intuition about the nature of this quantity, as when we look at the colors of a rainbow or listen to a singing voice. Mathematically, the transform linking a time signal to its frequency representation is the *Fourier transform*, which, starting from the days of its discovery[1], has become a powerful and indispensable tool in many fields of natural and engineering sciences.

Definition 2.1 (Fourier Transform) *The Fourier transform $\mathcal{F}[s(t)]$ of a time signal $s(t)$ satisfying*

$$\int_{-\infty}^{\infty} |s(t)| < \infty \tag{2.1}$$

is given by

$$\mathcal{F}[s(t)] = S(f) = \int_{-\infty}^{\infty} s(t) e^{-j2\pi f t} dt. \tag{2.2}$$

and its inverse by

$$\mathcal{F}^{-1}[S(f)] = s(t) = \int_{-\infty}^{\infty} S(f) e^{j2\pi f t} df. \tag{2.3}$$

[1]1907 by Jean Baptiste Joseph Fourier [Cohen, 1992; Lüke, 1985].

The condition (2.1) is sufficient but not necessary for the existence of $\mathcal{F}[s(t)]$ and the validity of (2.3). See [Papoulis, 1962] for details.

The Fourier transform maps a signal from the time domain into the frequency domain. However, if we need information about the frequency content of *time–varying* signals, the Fourier transform only tells us *what* components can be found at which amplitudes and phases but not *when* they actually occur. In order to answer the latter question we need some kind of joint time–frequency representation.

2.1.1 Linear Time–Frequency Distributions

Let $L_2(\mathbb{R})$ denote the space of square integrable functions with a single real–valued argument. Given a time signal $s(t) \in L_2(\mathbb{R})$, a time–frequency distribution (TFD) is a representation of $s(t)$ with time and frequency as its parameters. A TFD tells us, which frequencies occur at which times in the input signal. There are many different possibilities for obtaining a TFD for a given signal $s(t)$. For a long time the standard method has been calculating the short–time Fourier transform (STFT) of a signal $s(t)$:

$$F_s(t,f) = \int_{-\infty}^{\infty} s(\tau) \cdot g^*(\tau - t) \cdot e^{-j2\pi f\tau} \; d\tau, \quad \text{with } s(t), g(t) \in L_2(\mathbb{R}), \tag{2.4}$$

where the asterisk denotes complex conjugation. This is a windowed version of the Fourier transform with a single window of constant shape sliding along the time axis.

In the past few years wavelet transforms have become an important tool in signal processing [Rioul and Vetterli, 1991]. The continuous wavelet transform (CWT) of a signal $s(t) \in L_2(\mathbb{R})$ is given by

$$W_s(t,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} s(\tau) \cdot g^* \left(\frac{\tau - t}{a} \right) \; d\tau, \qquad a > 0, \tag{2.5}$$

or equivalently in the frequency domain using Parseval's identity

$$W_s(t,a) = \sqrt{a} \int_{-\infty}^{\infty} S(f) \cdot G^*(af) e^{j2\pi f t} df, \qquad a > 0, \tag{2.6}$$

where $G(f)$ denotes the Fourier transform of $g(t)$. As the parameter a is scaling the *mother–wavelet* $g(t)$, it is called *scale parameter*. Strictly speaking, the wavelet transform is a *time–scale* rather than a *time–frequency* distribution. However, if $G(f)$ exhibits a unique maximum with monotonically decreasing slopes, a one–to–one correspondence of the form $f \sim a^{-1}$ between scale parameter a and frequency parameter f can be established.

Given the wavelet transform $W_s(t,a)$, the signal $s(t)$ can be reconstructed by [Daubechies, 1990]

$$s(t) = c_g^{-1} \cdot \int_{-\infty}^{\infty} \int_{0}^{\infty} W_s(\tau, a) \cdot \frac{1}{\sqrt{a}} \, g\left(\frac{t - \tau}{a} \right) \frac{da \, d\tau}{a^2}, \tag{2.7}$$

where c_g is a constant satisfying

$$c_g = \int_{-\infty}^{\infty} \frac{|G(f)|^2}{|f|} df < \infty. \tag{2.8}$$

The latter equation is a necessary and sufficient condition for the admissibility of a time function $g(t)$ as a mother–wavelet. Equation (2.8) implies that

$$\int_{-\infty}^{\infty} |g(t)|^2 \, dt < \infty, \tag{2.9}$$

i.e. $g(t) \in L_2(\mathbb{R})$ and

$$\int_{-\infty}^{\infty} g(t) \, dt = G(0) = 0. \tag{2.10}$$

Thus, admissible functions must have finite energy and their transfer functions must have at least one zero at $f = 0$. Functions satisfying these conditions look like short waves, which has been the reason for naming them *wavelets*[2]. In terms of moments, (2.10) translates to the demand for a vanishing zeroth moment of $g(t)$. In general, a wavelet $g(t)$ is said to have n vanishing moments, if and only if for all non-negative integers $k \leq n$

$$\int_{-\infty}^{\infty} t^k \cdot g(t) \, dt = 0. \tag{2.11}$$

The relevance of the number of vanishing moments will be addressed in Section 2.3.1.

An important common property of both the CWT and the STFT is their linearity, which makes them more suitable for the analysis of multicomponent signals than bilinear TFDs, such as the so–called *spectrogram* resulting from squaring the magnitude of the STFT or the *scalogram* which is the squared magnitude of the CWT. Both spectrogram and scalogram can be considered as smoothed versions of the *Wigner–Ville–distribution*, which in turn belongs to *Cohen's class of distributions* [Cohen, 1995; Hlawatsch and Boudreaux-Bartels, 1992]. From their bilinearity follows that spectrograms and scalograms of multicomponent signals contain cross–term artifacts.

The main difference between STFT and CWT lies in the variation of the analysis windows along the frequency axis. While the STFT window remains unaltered, the CWT window changes its scale due to the scaling factor a. According to the scaling property of the Fourier transform

$$\mathcal{F}\left[s(at)\right] = \frac{1}{|a|} \cdot S\left(a^{-1} f\right), \tag{2.12}$$

the frequency window width is stretched by the same factor a^{-1} by which the time window width is narrowed, while the area (i.e. the product of the window length in both

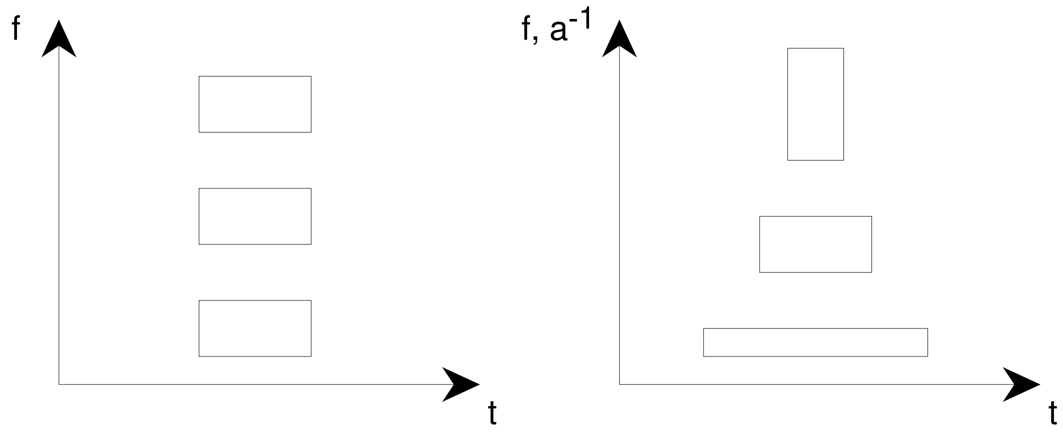

Figure 2.1: Windowing in the time-frequency plane, left STFT, right CWT.

domains) remains constant (see Fig. 2.1). If $\log(f)$ is displayed on the vertical axis instead of f, constant quotients with respect to f become constant differences with respect to $\log(f)$, so the $\log(f)$ frequency window width of the CWT is a constant. This is in close agreement with musical notation where pitch intervals are linearly spaced along a $\log(f)$ axis. Constant quotients of frequency values are also observed between the harmonics produced by weakly nonlinear systems excited by a fundamental sine. On a $\log(f)$–axis the harmonics will always run parallel to the fundamental, whereas on a linear f–axis the frequency trajectories of the harmonics diverge from each other (rising sweep) or converge to each other (falling sweep). The parallelity of the trajectories of the harmonics is an important advantage of the wavelet transform in pattern recognition applications and might also be the reason why the pitch perception in the human hearing mechanism exhibits logarithmic frequency resolution over a wide frequency range [Zwicker and Fastl, 1990]. In short, there are several good arguments for the CWT being a more appropriate TFD for acoustic signals than the STFT. It is for this reason, that the wavelet transform has found an increasing number of applications in sound analysis, see e.g. [Kronland-Martinet, 1988; Kronland-Martinet *et al.*, 1987].

The CWT given by (2.5) represents a constant–Q filter bank with an infinite number of bands having infinitely small distances in frequency from the upper and lower neighbors. Since this idealized filter bank cannot be realized in practice, one has to modify (2.5) by sampling the CWT at certain values for a and t. In the extreme case of orthogonal wavelets the wavelet representation is redundancy–free. These transforms are well–suited for numerical and data compression applications and can be realized efficiently by perfectly reconstructing multirate filter banks [Fliege, 1994]. However,

[2] Originally coined in French as *ondelettes*.

their use in signal analysis applications is limited, because they exhibit aliasing in the subbands corresponding to a lack of shift–invariance [Simoncelli *et al.*, 1992]. As a consequence, if a signal is shifted in time, the wavelet coefficients might change drastically across scale instead of just being shifted in time as well. A possibility to achieve shift–invariance with reduced data rate is giving up the regularity of sampling by employing amplitude dependent schemes like in [Mallat and Zhong, 1992]. This, however, leads to significant complications for signal reconstruction. In order to achieve alias–free subband processing on a regular sampling grid, oversampling becomes necessary [Fliege and Zölzer, 1990; Zölzer, 1997], inevitably leading to an increase of redundancy. For optimum time–shift invariance the continuous wavelet transform given by (2.5) should be sampled on a fine grid, yielding a *quasi–continuous* wavelet transform. The price to pay is an immense redundancy of this representation. Due to this redundancy, there is no unique way for signal reconstruction from the quasi–continuous wavelet transform. Based on the theory of frames [Daubechies, 1990] an iterative algorithm, known as *frame algorithm* [Gröchenig, 1993], can be applied.

Although the computational burden for calculating a quasi–continuous wavelet transform can be significantly reduced compared to straight–forward finite impulse response (FIR) filter implementations by the use of multirate techniques [Shensa, 1992], the computing costs are still high. In this respect infinite impulse response (IIR) filters offer a considerable advantage, which is exploited in the architecture presented in this thesis.

2.1.2 Time–Frequency Spread

It is a well–known fact, that a function cannot be arbitrarily well concentrated in both time and frequency. This follows immediately from (2.12), the scaling property of the Fourier transform. Obviously, the narrower the time function (i.e. $a \to \infty$), the wider is its Fourier transform. Center and spread of a signal in either domain can be conveniently defined analogously to mean and standard deviation in statistics [Papoulis, 1987]. Given the signal energy

$$E = \int_{-\infty}^{+\infty} |s(t)|^2 \, dt = \int_{-\infty}^{+\infty} |S(f)|^2 \, df, \tag{2.13}$$

we define the *window center* in the time domain as

$$t_0 = \frac{1}{E} \int_{-\infty}^{+\infty} t \cdot |s(t)|^2 \, dt. \tag{2.14}$$

In the frequency domain we have for the frequency window center[3]

$$f_0 = \frac{1}{E} \int_{-\infty}^{+\infty} f \cdot |S(f)|^2 \, df. \tag{2.15}$$

[3]For real–valued time signals we always arrive at $f_0 = 0$ because of their symmetric power spectral density. This is why the center frequency of a real–valued bandpass signal must be obtained by considering its analytic counterpart (see Definition 2.2 in Section 2.2).

The *duration* or *time window width* are defined as

$$\Delta t = \sqrt{\frac{1}{E}\int_{-\infty}^{+\infty}(t-t_0)^2\cdot|s(t)|^2\,dt} \tag{2.16}$$

and the *bandwidth* or *frequency window width* as

$$\Delta f = \sqrt{\frac{1}{E}\int_{-\infty}^{+\infty}(f-f_0)^2\cdot|S(f)|^2\,df}. \tag{2.17}$$

For the case of converging integrals in (2.14), (2.15), (2.16) and (2.17), one can derive for the area of the window in the time–frequency plane [Papoulis, 1987]

$$\Delta t\cdot\Delta f \geq \frac{1}{4\pi}, \tag{2.18}$$

meaning the concentration of a signal in the time–frequency plane is limited by this lower bound. The minimum area of $\frac{1}{4\pi}$ only holds for the Gaussian function and its shifted variants in the time–frequency plane [Papoulis, 1987]. Analogously, lower bounds can be established for any two physical quantities being related via operators that do not commute [Cohen, 1995].

Equation (2.18) is often called the *uncertainty principle*, a term originating from quantum mechanics. This term, however, it is somewhat misleading in the given context, as there is nothing uncertain about the spread of a deterministic signal. The notion of uncertainty might have been responsible for several attempts to interpret (2.18) as a fundamental limit to the precision of our knowledge about the properties of a signal in time and frequency[4]. See e.g. [Yen, 1987] where the author conjectures that the negative terms in the Wigner–Ville distribution result from a "violation" of the uncertainty principle. While it is already difficult to see how (2.18) could be violated, it does not express any restrictions for resolving *local* signal properties in the time–frequency plane anyway [Loughlin *et al.*, 1992]. The key point is that the restriction to a single fixed window must be given up in favor of a signal–dependent windowing scheme.

The time–frequency tradeoff has inspired some researchers to devise various kinds of multiresolution analyses, such as the *wavelet package transform* [Coifman and Wickerhauser, 1992] and the *multiresolution Fourier transform* (MFT) [Pearson, 1991; Wilson *et al.*, 1992]. It is argued that the arbitrariness of choosing a single window can only be overcome by combining information gained from a large set of window families with different resolutions. Naturally, the question arises which and how many resolutions should be chosen and in which way this increased amount of data should

[4]This assertion would stand in clear contradiction to the widely accepted requirement of falsifiability in natural science [Popper, 1935]. A statement such as *a measurement of x can never be performed with a precision better than* Δx is a metaphysical dogma, which not only does not make any falsifiable predictions but even asserts the impossibility of such predictions.

be fused. Clearly, accumulating vast amounts of already over–complete data representations is a questionable if not intractable approach due to physical limitations. It is not only questionable but even pointless, since with a single highly redundant family of functions already, further local resolutions can be derived at any location in the time–frequency plane, just by combining a few adjacent coefficients. An example for this is the approach of collecting several narrowband signals in order to form a broadband signal as done in the onset detection approach described in this thesis. While the narrowband windows target signals localized in frequency, the broadband window resulting from their combination is responsible for the signals localized in time, thus resulting in a two–resolution approach. Further intermediate resolutions could be generated by adding more schemes of coefficient combination.

2.2 Signals Localized in Frequency

Many real world signals can be appropriately modeled by a sum of sinusoids with time–varying amplitude. In order to characterize such a signal, it is essential to define a set of parameters, by which any sinusoid can be referred to unambiguously. This is not as straightforward as it might seem at first sight, not even in the single signal case. For instance, consider the signal

$$s_r(t) = a(t) \cdot \sin(2\pi f_0 t), \qquad 0 \leq a(t) \leq a < \infty. \tag{2.19}$$

One might be tempted to argue that this is a signal with time dependent amplitude $a(t)$ and constant frequency f_0. However, without imposing further restrictions this is not the only possible interpretation. For example, we can find a phase distribution $\psi(t)$ such that $\sin[\psi(t)] = 1/a \cdot s_r(t)$ or equivalently

$$s_r(t) = a \cdot \sin[\psi(t)], \tag{2.20}$$

which would be a signal of constant amplitude and non–constant frequency. In fact, there are infinitely many pairs $\{a_i(t), \psi_i(t)\}$ that can equivalently characterize the signal given by (2.19) [Picinbono, 1997].

2.2.1 Definitions

The standard approach to resolve the ambiguity mentioned above is by introducing the *analytic signal* notion, see e.g. [Boashash, 1992; Picinbono, 1997].

Definition 2.2 (Analytic Signal) *A time signal $s(t)$ is called* analytic*, if*

$$\forall f < 0 : \mathcal{F}[s(t)] = 0.$$

For analytic signals, $\mathcal{F}[s(t)]$ is asymmetric with respect to the imaginary axis, so $s(t)$ must be complex–valued.

Definition 2.3 (Canonical Pair) *$\{a(t), \psi(t)\}$, with $a(t) \in \mathbb{R}_+$ and $\psi(t) \in \mathbb{R}$, is a* canonical pair *if and only if $a(t) \cdot e^{j\psi(t)}$ is an analytic signal.*

In order to relate a real valued signal to its associated canonical pair, the *Hilbert transform* is introduced.

Definition 2.4 (Hilbert Transform Operator) *Let $s(t)$ be a function for which the Fourier transform exists. Then, the* Hilbert transform operator *$\mathcal{H}$ is a linear operator, for which*

$$\mathcal{F}\left[\mathcal{H}[s(t)]\right] = -j \cdot \mathrm{sgn}(f) \cdot \mathcal{F}\left[s(t)\right].$$

The Hilbert transform maintains the even symmetry of the power spectral density modulus and the uneven symmetry of the phase. Thus, the Hilbert transform of a real–valued time signal is again a real–valued signal. The factor $-j$ translates to a phase shift of $-\frac{\pi}{2}$. Note, that $s_1(t) \neq s_2(t)$ does not guarantee $\mathcal{H}[s_1(t)] \neq \mathcal{H}[s_2(t)]$, since the $\mathrm{sgn}(f)$ function masks out any singularity that $\mathcal{F}\left[s(t)\right]$ might have at $f = 0$. However, with use of the Hilbert transform, a corresponding analytic signal can be found for any real–valued signal due to the following corollary:

Corollary 2.1 *For any real–valued signal $s_r(t)$ there is an associated analytic signal $s(t) \in \mathbb{C}$ given by*

$$s(t) = s_r(t) + j\mathcal{H}[s_r(t)]$$

and an associated canonical pair

$$\{a(t), \psi(t)\} = \{|s(t)|, \arg[s(t)]\}.$$

Proof *The Fourier transform of $s(t)$*

$$\begin{aligned} \mathcal{F}[s(t)] &= \mathcal{F}[s_r(t)] + j\mathcal{F}[\mathcal{H}[s_r(t)]] \\ &= \mathcal{F}[s_r(t)] + \mathrm{sgn}(f) \cdot \mathcal{F}[s_r(t)] \\ &= 2 \cdot \epsilon(f) \cdot \mathcal{F}[s_r(t)], \end{aligned}$$

with $\epsilon(f)$ denoting the unit step at $f = 0$, is zero for $f < 0$. Thus, due to Definitions 2.2 and 2.3, $s(t) = |s(t)| \cdot e^{j\arg[s(t)]}$ is an analytic signal with the associated canonical pair $\{|s(t)|, \arg[s(t)]\}$.
$\square$

The original signal $s_r(t)$ may be recovered from its corresponding analytic signal $s(t)$ due to the following corollary:

Corollary 2.2 *For any real–valued signal $s_r(t)$ and its associated analytic signal $s(t)$ holds*

$$s_r(t) = \mathrm{Re}[s(t)].$$

Proof *As the Hilbert transform maintains the even symmetry the power spectral density modulus and the uneven symmetry of the phase, the Hilbert transform of a real–valued signal is again a real–valued signal. As $s_r(t)$ is real–valued, we have*

$$\mathrm{Re}[s(t)] = \mathrm{Re}\left[s_r(t) + j\mathcal{H}[s_r(t)]\right] = s_r(t).$$

□

From Corollary 2.1 and Corollary 2.2 follows that for any real–valued signal $s_r(t)$ there is indeed exactly one corresponding analytic signal $s(t)$.

Definition 2.5 *For any real–valued, bandlimited signal $s_r(t)$ and its associated canonical pair $\{a(t), \psi(t)\}$ we define*

- *instantaneous amplitude*

$$a(t),$$

- *instantaneous phase*

$$\psi(t).$$

At this point one might be tempted to undeliberately define *instantaneous frequency* as $f(t) = \frac{1}{2\pi} \cdot \frac{d}{dt}\psi(t)$, but there are several problems associated with this definition that should be considered. First of all, even if $s(t) = a(t) \cdot e^{j\psi(t)}$ is an analytic signal due to Definition 2.2, this does not necessarily mean that it is also an *analytic function* in the function theoretic sense, because the derivative of $s(t)$ might not exist everywhere. However, it can be shown that signals being bandlimited according to the following definition are indeed analytic functions [Papoulis, 1987]:

Definition 2.6 (Bandlimited Signal) *A time signal $s(t)$ is called* bandlimited, *if there exist finite $f_l, f_r \in \mathbb{R}$ such that*

$$\forall f \notin [f_l, f_r] : \mathcal{F}[s(t)] = 0.$$

Thus, if an *analytic signal* is also right–hand bandlimited, it is indeed an *analytic function* and its derivative exists everywhere.[5] However, even if the signal itself is bandlimited, this does not imply that its instantaneous phase is also an analytic function. An example for a bandlimited, analytic signal with non–differentiable instantaneous phase is

$$s(t) = a \cdot e^{j2\pi f_1 t} + a \cdot e^{j2\pi f_2 t}, \tag{2.21}$$

[5]The distinction between *analytic signal* and *analytic function* has become necessary, because common use of the term *analyticity* in the signal processing literature does not strictly adhere to the mathematical definition.

which can be rewritten as

$$s(t) = a_0(t) \cdot e^{j2\pi f_0 t} \tag{2.22}$$

with

$$f_0 = \frac{f_1 + f_2}{2}, \qquad a_0(t) = a \cdot \cos\left[2\pi \cdot \frac{f_1 - f_2}{2} \cdot t\right].$$

This signal's instantaneous amplitude $|a_0(t)|$ is zero periodically and the zero crossings of $a_0(t)$ lead to phase discontinuities with phase jumps of π. Thus, in spite of this signal being bandlimited due to Definition 2.6, its instantaneous frequency is periodically infinite. On the other hand, the impossibility of sensefully defining instantaneous frequency for the given example should not be surprising, as the signal was constructed by summing up *two* signals with *two different* instantaneous frequencies. We conclude that for defining instantaneous frequency, some *monochromaticity* restriction is needed, inevitably leading to the notion of *instantaneous bandwidth.* Due to (2.17), the overall frequency spread of the analytic signal

$$s(t) = a(t) \cdot e^{j\psi(t)} \tag{2.23}$$

can be calculated as [Cohen, 1995]

$$\begin{aligned}
\Delta^2 f &= \frac{1}{E} \cdot \int_{-\infty}^{+\infty} (f - f_0)^2 \cdot |S(f)|^2 \, df \\
&= \frac{1}{E} \cdot \int_{-\infty}^{+\infty} s^*(t) \cdot \left(\frac{1}{j2\pi}\frac{d}{dt} - f_0\right)^2 s(t) \, dt \\
&= \frac{1}{E} \cdot \int_{-\infty}^{+\infty} \left|\left(\frac{1}{j2\pi}\frac{d}{dt} - f_0\right) s(t)\right|^2 dt \\
&= \frac{1}{E} \cdot \left[\int_{-\infty}^{+\infty} \left(\frac{a'(t)}{2\pi \cdot a(t)}\right)^2 \cdot |s(t)|^2 \, dt \right. \\
&\qquad \left. + \int_{-\infty}^{+\infty} \left(\frac{1}{2\pi} \cdot \psi'(t) - f_0\right)^2 \cdot |s(t)|^2 \, dt\right],
\end{aligned} \tag{2.24}$$

where f_0 is the mean frequency due to (2.15). Thus, the frequency spread can be considered as made up of two basic contributions. The second term in (2.24), governed by the derivative of the instantaneous phase, is the *frequency modulated* (FM) component of the overall bandwidth [Cohen, 1995]. If it equals the mean frequency for all times, its contribution to the overall frequency spread is zero. Thus, it seems justified to interpret the derivative of the instantaneous phase as the instantaneous frequency. Furthermore, as the overall bandwidth might still be greater than zero, even if the instantaneous frequency equals the mean frequency f_0 for all times, it seems appropriate to interpret the first term in (2.24), the *amplitude modulated* (AM) component, as governed by what we call the *instantaneous bandwidth*

$$\Delta f(t) = \frac{1}{2\pi} \cdot \left| \frac{a'(t)}{a(t)} \right| = \frac{1}{2\pi} \cdot \left| \frac{d}{dt} \log a(t) \right|. \tag{2.25}$$

Intuitively we would require a monochromatic signal to have a good localization around its instantaneous frequency. Thus, the additional restriction we impose is

$$\frac{d}{dt}\psi(t) \gg \left| \frac{d}{dt} \log a(t) \right|. \tag{2.26}$$

The signal (2.21) considered above is analytic but (2.26) is not satisfied everywhere. Taking these considerations into account, we have the following definition:

Definition 2.7 *For any real–valued, positive, bandlimited signal $s_r(t)$ and its associated canonical pair $\{a(t), \psi(t)\}$ we define*

- *instantaneous frequency*

$$f(t) = \frac{1}{2\pi} \cdot \frac{d}{dt}\psi(t),$$

- *instantaneous bandwidth*

$$\Delta f(t) = \frac{1}{2\pi} \cdot \left| \frac{d}{dt} \log a(t) \right|,$$

provided that $f(t) \gg \Delta f(t)$.

The notion of instantaneous bandwidth is important for the understanding of the shortcomings of analysis concepts based on partial modeling only. It becomes apparent that signals with locally non–differentiable amplitude cannot be sensefully characterized by instantaneous frequency and bandwidth. In order to cope with the presence of such signals the notion of singularities will be introduced as complementary to the analytic signal concept in Section 2.3.1.

It is interesting to note, that for a stable one–pole linear system the instantaneous bandwidth due to Definition 2.7 equals the 3 dB–bandwidth, because for an amplitude of the form $a(t) = e^{-\lambda t}, \lambda \in \mathbb{R}$ we have

$$\Delta f(t) = \frac{1}{2\pi} \cdot \left| \frac{d}{dt} \log e^{-\lambda t} \right| = \frac{\lambda}{2\pi}, \tag{2.27}$$

Thus, for the impulse response of a one pole system, $\Delta f(t)$ is constant and determined by the real part of the pole location, coinciding with its 3dB–bandwidth.

The notion of a signal *partial* is extensively used throughout the sound analysis literature but it is rarely ever defined what it actually means. In this thesis we use the following definition:

Definition 2.8 (Partial) *A signal is called* partial *if its associated analytic signal satisfies $f(t) \gg \Delta f(t)$,*

i.e. if the signal satisfies the condition for Definition 2.7. Given this definition, the model assumed for a single partial is the following:

$$s(t) = a(t) \cdot e^{j\left(2\pi \int_0^t f(\tau)d\tau + \psi_0\right)}, \quad \text{with } a(t) > 0, f(t) \gg \Delta f(t). \tag{2.28}$$

At this point the question how we can deal with a superposition of partials naturally arises. But what makes a signal multicomponent? Cohen [1992] suggests that

$$\frac{\Delta f_1(t) + \Delta f_2(t)}{2} < |f_1(t) - f_2(t)| \tag{2.29}$$

should hold in order to consider two partials as separated at t, i.e. the bands associated with the partials must not overlap each other. This is the condition adopted in this thesis. However, it is important to note, that this mathematical definition does not necessarily correspond to psychoacoustic findings. For example, with the mathematical definition, two partials could merge into a single one at an intersection and subsequently fall apart again, while a human listener would never notice a change in partial number (see also Section 3.2.5.2). Nevertheless, the intersection is an interesting point, because this is where the human auditory system must make a decision of how to assign belonging partial pieces from both sides of the intersection. In this respect the change in partial number due to the mathematical definition can be considered as a reflection of this significance.

2.2.2 Partial Parameter Estimation in Gaussian White Noise

Consider the model

$$s(t) = a(t) \cdot e^{j\left(2\pi \int_0^t f(\tau)d\tau + \psi_0\right)} + n(t), \quad a(t) > 0, f(t), \psi_0 \in \mathbb{R}. \tag{2.30}$$

where $n(t)$ is complex Gaussian white noise. Both amplitude and frequency estimation are largely affected by the presence of noise. First of all, Definition 2.7 cannot be applied immediately, since $s(t)$ is nowhere differentiable. This can be overcome by considering the sampled equivalent of (2.30):

$$s(kT_s) = a(kT_s) \cdot e^{j\left(2\pi T_s \sum_0^{kT_s} f(kT_s) + \psi_0\right)} + n_s(kT_s), \tag{2.31}$$

provided that the sampling theorem is satisfied for $a(t) \cdot e^{j2\pi \int_0^t f(\tau)d\tau}$ at the sampling rate $f_s = T_s^{-1}$ and that $s(t)$ is passed through an ideal lowpass filter with bounding frequency $\frac{f_s}{2}$ before sampling. Now, the instantaneous frequency can be estimated as

$$\hat{f}(kT_s) = \frac{\arg[s(kT_s)] - \arg[s((k-1)T_s)]}{2\pi T_s}. \tag{2.32}$$

The estimate thus obtained can be improved by weighted averaging of adjacent phase differences if local stationarity is assumed, that is if $a(t) \approx const$ and $f(t) \approx const$. The situation for the amplitude is not less difficult, since averaging moduli as done in [Wang, 1994] yields a biased estimate of the amplitude in the presence of noise. In the extreme case of zero amplitude, the expectation value of $|s(kT_s)|$ is directly proportional to the standard deviation of the noise. See Appendix C for a discussion of this issue.

One possible method to achieve maximum likelihood parameter estimation for a single partial in complex Gaussian white noise is performing a one–dimensional search for the maximum of the signal's Fourier transform [Rife and Boorstyn, 1974]. However, as pointed out in Section 2.1.1, the constant time–frequency window shape equally affecting the whole frequency range under consideration is a significant flaw of this method, especially when nonstationary signals are considered.

Maximum likelihood parameter estimation in the multiple partial signal case is much more complicated. The estimation of the partial number alone is already a nontrivial problem, for which many different criteria have been proposed [Kay, 1988]. The problem gets even more difficult for small signal–to–noise ratios. In these cases principal component decomposition methods have been applied with some success [Marple, 1987]. They all share the property of high computation load and difficulties in nonstationary environments.

An alternative to estimating the partial parameters directly is determining the parameters of a linear transfer function model driven by Gaussian white noise. Parametric modeling is generally considered superior to non–parametric methods, if the model is chosen properly [Kay, 1988]. The most general linear model is the *autoregressive moving average* (ARMA) model consisting of a *moving average* (MA) part represented by the numerator of the transfer function and an *autoregressive* (AR) part represented by its denominator. For the estimation of partial parameters, AR models are the appropriate choice, because it is the poles of a transfer function that are responsible for the sharp peaks in its power spectral density. The fact that AR parameter estimation leads to solving a linear matrix equation, as opposed to involved nonlinear computations in the MA or ARMA case is a pleasant bonus. For a short overview of AR modeling, see Appendix F.

2.2.3 Error Bounds

Given the signal model (2.31), we now collect some error bound results for estimating partial parameters in noise. The frequency is assumed to be constant over the observation interval and the noise is of variance $2\sigma_n^2$, with the real and imaginary part being statistically independent of each other.[6] In [Rife and Boorstyn, 1974] the

[6]This problem formulation is not equivalent to estimating the parameters of a signal composed of a real valued sinusoid and its Hilbert transform, because the resulting complex noise is not white [James *et al.*, 1994].

Cramér–Rao–Bounds (CRB) (see Appendix E) for amplitude A and frequency f of a single partial are derived as

$$\sigma_A^2 \geq \frac{\sigma_n^2}{N} \tag{2.33}$$

and

$$\sigma_f^2 \geq \frac{12\sigma_n^2 f_s^2}{(2\pi A)^2 N(N^2-1)}, \tag{2.34}$$

where σ_n^2 is the variance of both, real and imaginary part, of the noise and N is the number of complex–valued samples considered. As the estimator in the complex case can be considered as consisting of two separate independent estimators, one for the real and one for the imaginary part, both in Gaussian white noise of variance σ_n^2, the CRBs (2.33) and (2.34) are increased by a factor of two if only $\mathrm{Re}[s(t)]$ is measured. Stoica and Nehorai [1988] consider the asymptotic (i.e. large–sample) CRB for the colored noise case, obtaining

$$\lim_{N\to\infty} \sigma_f^2 \geq \frac{12\sigma_n^2 f_s^2}{(2\pi A)^2 N^3} \cdot |H(f_0)|^2 \tag{2.35}$$

which is basically the same as (2.34) with an additional factor of $|H(f_0)|^2$, where f_0 is the frequency of the signal and $H(f)$ is the transfer function of the coloring filter. If both, partial and noise, are passed through the same coloring filter, there is a factor $|H(f_0)|^2$ in both the numerator and the denominator, so we get back to (2.34).

In the multicomponent case, the *Fisher information matrix* is no longer diagonal [Friedlander and Francos, 1995]. This leads to a practical impossibility of exact maximum likelihood parameter estimation, except for only a few selected cases [Kay, 1988]. If the partials are harmonically related, the asymptotic error bound for estimating the frequency of the fundamental is given by [James *et al.*, 1994]

$$\sigma_f^2 \geq \frac{12\sigma_n^2 f_s^2}{(2\pi)^2 N(N^2-1)\sum_{k=0}^{K-1} k^2 a_k^2}, \tag{2.36}$$

if the number K of amplitudes a_k is known.

With the lower error bounds given above, it is obvious that downsampling should be avoided for maximum estimator performance, since the decrease in the number of observations N would lead to a deterioration of the estimates.

2.3 Signals Localized in Time

In the previous section, signals localized in frequency were discussed. Ideal sine waves are ideally concentrated in frequency. This is no longer the case if the instantaneous amplitude varies over time. At points where the instantaneous amplitude is discontinuous, the analytic signal concept breaks down as the instantaneous bandwidth becomes unbounded. The time–frequency trade–off explained in Section 2.1.2 suggests that this might be due to the presence of a signal component mainly localized in time.

Consider the Fourier transform of a signal $s(t)$:

$$S(f) = A(f) \cdot e^{j\Psi(f)}, \quad A(f) \in \mathbb{R}_+. \tag{2.37}$$

Analogously to (2.24) we have for the time spread due to (2.16)

$$\begin{aligned}
\Delta^2 t &= \frac{1}{E} \cdot \int_{-\infty}^{+\infty} (t - t_0)^2 \cdot |s(t)|^2 \, dt \\
&= \frac{1}{E} \cdot \int_{-\infty}^{+\infty} S^*(f) \cdot \left(-\frac{1}{j2\pi} \frac{d}{df} - t_0 \right)^2 S(f) \, df \\
&= \frac{1}{E} \cdot \int_{-\infty}^{+\infty} \left| \left(-\frac{1}{j2\pi} \frac{d}{df} - t_0 \right) S(f) \right|^2 df \\
&= \frac{1}{E} \cdot \left[\int_{-\infty}^{+\infty} \left(-\frac{A'(f)}{2\pi \cdot A(f)} \right)^2 \cdot |S(f)|^2 \, df \right. \\
&\quad \left. + \int_{-\infty}^{+\infty} \left(\frac{1}{2\pi} \cdot \Psi'(f) + t_0 \right)^2 \cdot |S(f)|^2 \, df \right],
\end{aligned} \tag{2.38}$$

where t_0 is the mean time defined by (2.14). Thus, the time spread can be considered as made up of two basic contributions: The second term in (2.38) is governed by the *group delay*

$$d(f) = -\frac{1}{2\pi} \cdot \Psi'(f), \tag{2.39}$$

which is the time shift at frequency f. If it equals the mean time t_0 for all frequencies, its contribution to the overall time spread is zero and each frequency component of the signal is localized around the same time t_0. The first term is governed by the time spread at a given frequency

$$\Delta t(f) = \frac{1}{2\pi} \cdot \left| \frac{d}{df} \log A(f) \right|. \tag{2.40}$$

Like in the case of signals concentrated in the frequency domain, we need to impose some conditions for $d(f)$ and $\Delta t(f)$ to be well defined. The analyticity condition according to Definition 2.2 is paralleled by the requirement of causality:

Definition 2.9 (Causal Signal) *A time signal $s(t)$ is* causal, *if*

$$\forall\, t < 0 : s(t) = 0.$$

Analogously to Definition 2.6, the signal is required to be time–limited:

Definition 2.10 (Time–limited Signal) *A time signal $s(t)$ is called* time–limited, *if there exist finite $t_l, t_r \in \mathbb{R}$ such that*

$$\forall\, t \notin [t_l, t_r] : s(t) = 0.$$

Analogously to the requirement of concentration around the instantaneous frequency, $\Delta t(f)$ must satisfy

$$\Delta t(f) \ll d(f). \tag{2.41}$$

Finally, two signals are considered separate in the time domain at frequency f, if [Cohen, 1995]

$$\frac{\Delta t_1(f) + \Delta t_2(f)}{2} < |d_1(f) - d_2(f)|\,, \tag{2.42}$$

which is analogous to (2.29). If $\Delta f(t)$ and $\Delta t(f)$ are both finite and $\Delta f(t) \cdot \Delta t(f) \gg 1$ holds, s(t) is called *asymptotic*. If the instantaneous frequency of an asymptotic signal is also monotonic, then $f(t)$ is approximately the inverse of $d(f)$ [Boashash, 1992].

2.3.1 Isolated Singularities

In terms of time localization the most extreme case is the Dirac impulse, which is the time–localization counterpart of the infinitely extended phasor $e^{j2\pi f t}$ being ideally concentrated in the frequency domain. The Dirac impulse belongs to the class of distributions with a single isolated *singularity*.

Definition 2.11 (Singularity) *We call a real–valued signal $s(t), t \in \mathbb{R}$* singular *in t_0, if it is not differentiable in t_0, but every neighborhood of t_0 contains points in which it is differentiable. The* singularity *in t_0 is called* isolated, *if there is a neighborhood of t_0 without further singularities.*

An example for an isolated singularity is $t = 0$ for the Dirac impulse $\delta(t)$, the unit step $\epsilon(t)$ and the linear ramp $\epsilon(t) \cdot t$. Without loss of generality, we restrict our considerations to singularities located at $t = 0$.

Definition 2.12 (Homogeneity) *A function $s(t)$ is homogeneous of degree μ, if and only if there is a $\mu \in \mathbb{R}$ such that for every $a \in \mathbb{R}_+$*

$$s\left(\frac{t}{a}\right) = a^{-\mu} s(t).$$

The examples for signals with isolated singular points given above are also homogeneous, the Dirac impulse being homogeneous[7] of degree -1, the unit step of degree 0 and the linear ramp of degree 1.

Corollary 2.3 *If $s(t)$ is homogeneous and the wavelet transform $W_s(t,a)$ converges for all $t \in \mathbb{R}$, the following condition holds:*

$$\frac{t}{a} = const \quad \Longrightarrow \quad \arg\left[W_s(t,a)\right] = const.$$

Proof *From the definition of $W_s(t,a)$ in (2.5) we have*

$$W_s(t,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} s(\tau)\cdot g^*\left(\frac{\tau - t}{a}\right)\, d\tau, \qquad a > 0.$$

From the homogeneity of $s(t)$ follows

$$\begin{aligned} W_s(t,a) &= a^{\mu-\frac{1}{2}} \int_{-\infty}^{\infty} s\left(\frac{\tau}{a}\right)\cdot g^*\left(\frac{\tau - t}{a}\right)\, d\tau \\ &= a^{\mu+\frac{1}{2}} \int_{-\infty}^{\infty} s(\tau)\cdot g^*\left(\tau - \frac{t}{a}\right)\, d\tau. \end{aligned} \tag{2.43}$$

That is, if $\frac{t}{a}$ is constant, the wavelet transform $W_s(t,a)$ is constant up to a factor $a^{\mu+\frac{1}{2}}$ for the modulus. In particular, the phase of $W_s(t,a)$ is constant, if $\frac{t}{a}$ is constant.
□

From Corollary 2.3 follows, that under the given conditions one can observe a characteristic local pattern in time–scale space formed by lines of constant phase spreading across all scales. Due to the constance of $\frac{t}{a}$ along the phase lines, they all converge to $t \to 0$ for $a \to 0$, thus forming a characteristic cone–shaped pattern. As the wavelet transform is linear, this argument can be extended to sums of homogeneous signals with isolated singularities at $t = 0$, provided that the wavelet transform converges for each of them separately. Each line l_i of the pattern satisfies $t = t_0 + c_i \cdot a$, $c_i \in \mathbb{R}$. On a logarithmic scale axis we have $a' = \log a$ instead, from which follows $t = t_0 + c_i \cdot e^{a'}$. Thus, the lines of constant phase caused by a singularity are exponentials in the time–scale plane. See Fig. 2.2 reproduced from [Kliewer, 1993] for an example showing the Dirac impulse function analyzed by an analytic Gaussian wavelet. The zero phase is colored black, fading to white towards 2π.

Corollary 2.3 requires the integral in (2.43) to converge for all $t \in \mathbb{R}$. In order to ensure this, the wavelet $g(t)$ must have a sufficient number of vanishing moments. For the unit step and the Dirac impulse, Equation (2.10), resulting from the wavelet

[7] See e.g. [Lüke, 1985; Fliege, 1991] for a derivation of $\delta\left(a^{-1}t\right) = |a| \cdot \delta(t)$.

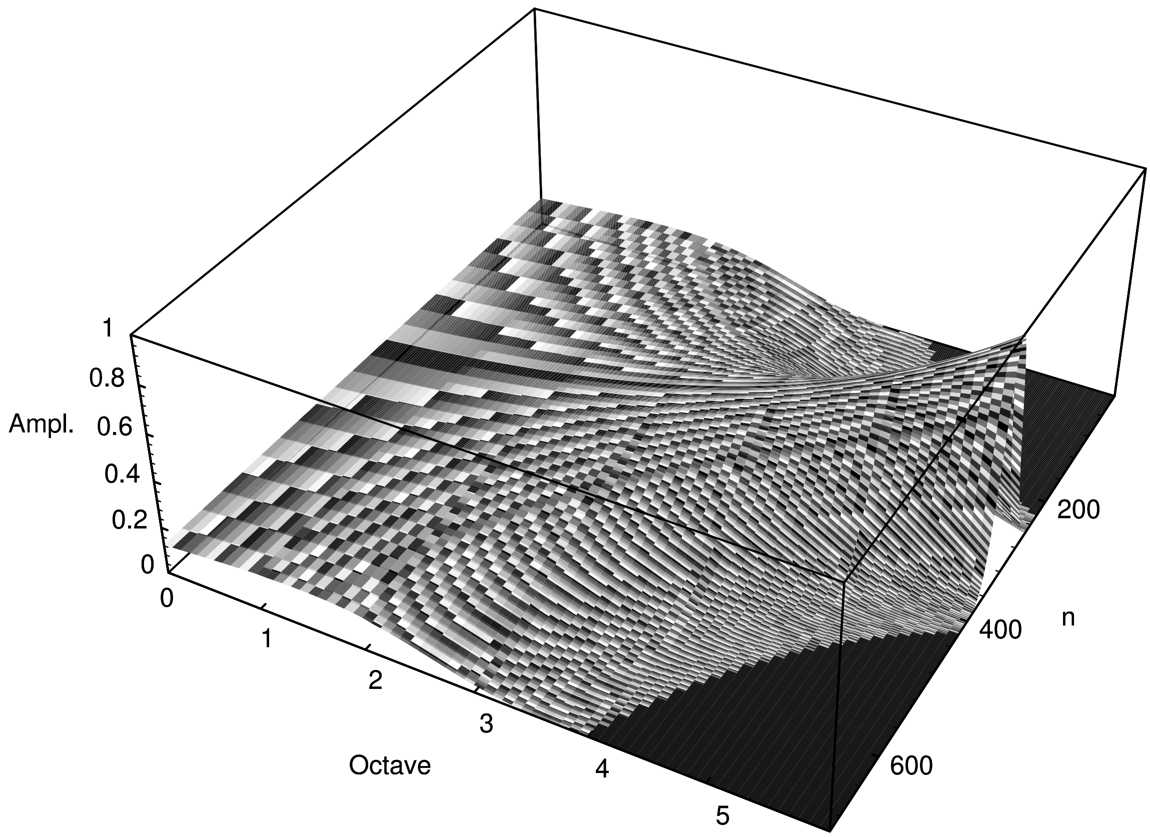

Figure 2.2: Dirac impulse analyzed by an analytic Gaussian wavelet.

admissibility condition (2.8), is sufficient for convergence. For a causal homogeneous signal with a degree of homogeneity $\mu > 0$, the following requirement must be satisfied:

$$\exists c \in \mathbb{R}\, \forall\, a \in \mathbb{R}_+,\, t \in \mathbb{R}\ :\ \left| \int_0^\infty \tau^\mu \cdot g^* \left(\tau - \frac{t}{a} \right) d\tau \right| < c, \tag{2.44}$$

for which

$$\forall\, a \in \mathbb{R}_+\ :\ \lim_{t \to \infty} \int_0^\infty \tau^{\mu - 1} \cdot g^* \left(\tau - \frac{t}{a} \right) d\tau = 0, \tag{2.45}$$

is a necessary condition, i.e. the wavelet must have at least $\mu - 1$ vanishing moments. The estimation of the degree of homogeneity μ by evaluating the modulus of the wavelet transform across scales can be used to identify the nature of the singularity [Mallat and Zhong, 1992]. However, as we are dealing with superpositions of several signals, the result would be difficult to interpret in most nontrivial cases. At least we can hope to find a method for the localization of the characteristic phase pattern in order to localize onsets of signals without using too much specific knowledge of the signal characteristics.

2.3.2 Estimation of Arrival Times

In the following, the maximum likelihood estimator for the arrival time τ of a known real–valued, causal target signal $s(t)$ is derived. For a brief overview of the maximum likelihood estimation concept, see Appendix E. It is assumed that $s(t)$ has zero group delay and $s(t - \tau)$ is causal and concentrated around its group delay $d(f) = \tau$, such that it lies completely within a finite observation interval $[0, T]$, i.e.

$$T \gg \tau + \Delta t \quad \text{and} \quad \tau \gg \Delta t, \tag{2.46}$$

with Δt denoting the root mean square duration of $s(t)$ due to (2.16). Consider the signal:

$$v(t) = s(t - \tau) + n(t), \tag{2.47}$$

where $n(t) \in L_2(0, T)$ is a Gaussian noise process satisfying

$$\frac{1}{T} \cdot \int_0^T n(t) \cdot x(t)\, dt = \begin{cases} N_0, & \text{for } x(t) = n(t) \\ 0, & \text{for } x(t) = \frac{\partial^i}{\partial \tau^i} s(t - \tau), \end{cases} \tag{2.48}$$

for any $\tau \in [0, T]$ and $i \in \{0, 1, 2\}$. As $n(t)$ is Gaussian distributed, the likelihood is

$$p_s(v(t)|\tau) = \frac{1}{\sqrt{2\pi\, N_0}} \cdot \exp\left[- \frac{\frac{1}{T} \cdot \int_0^T (v(t) - s(t - \tau))^2\, dt}{2\, N_0} \right]. \tag{2.49}$$

For maximizing the likelihood with respect to τ we determine

$$\max_{\tau} \log p_s(v(t)|\tau). \tag{2.50}$$

Inserting (2.49) into (2.50) and dropping all terms that are invariant with respect to τ yields

$$\max_{\tau} \int_0^T 2v(t) \cdot s(t-\tau) - s^2(t-\tau)\, dt. \tag{2.51}$$

With (2.46) we have

$$\frac{\partial}{\partial \tau} \int_0^T s^2(t-\tau)\, dt \approx 0, \tag{2.52}$$

so it suffices to find

$$\max_{\tau} \int_0^T v(t) \cdot s(t-\tau)\, dt. \tag{2.53}$$

Due to (2.53), the maximum likelihood estimate is the maximum of the signal resulting from the convolution of $v(t)$ with the time–reversed target signal $s(-t)$. The filter $s(-t)$ is the so–called *matched filter* of $s(t)$ in the Gaussian white noise case [Helstrom, 1995; Mertins, 1996].

In the following we derive an approximate *Cramér–Rao bound* for the error variance of the arrival time estimate $\sigma^2_{\hat{\tau}}$. The target $s(t)$ is assumed to be an energy–normalized lowpass signal. The generalization to bandpass signals is straight–forward by considering their lowpass prototypes. From (E.9) we have

$$\begin{aligned}
\sigma^2_{\hat{\tau}} &\geq -E\left\{\left(\frac{\partial^2 \ln p_s(v(t)|\tau)}{\partial \tau^2}\right)\right\}^{-1} \\
&= 2\, N_0\, T \cdot E\left\{\left(\frac{\partial^2}{\partial \tau^2} \int_0^T 2v(t)s(t-\tau) - s^2(t-\tau)\, dt\right)\right\}^{-1}, \quad \text{with (2.49)} \\
&\approx N_0\, T \cdot E\left\{\left(\frac{\partial^2}{\partial \tau^2} \int_0^T v(t) \cdot s(t-\tau)\, dt\right)\right\}^{-1}, \quad \text{with (2.52)} \\
&= N_0\, T \cdot E\left\{\left(\int_0^T (s(t-\tau) + n(t)) \cdot \left(\frac{\partial^2}{\partial \tau^2} s(t-\tau)\right)\, dt\right)\right\}^{-1}, \quad \text{with (2.47)} \\
&= N_0\, T \cdot \left(\int_{-\infty}^{\infty} s(t) \cdot \left(\frac{\partial^2}{\partial t^2} s(t)\right)\, dt\right)^{-1}, \quad \text{with (2.46) and (2.48)} \\
&\approx N_0\, T \cdot \left(\int_{-\infty}^{\infty} (2\pi f)^2 \cdot |S(f)|^2\, df\right)^{-1},
\end{aligned} \tag{2.54}$$

using Parseval's identity. With $s(t)$ being energy–normalized and Δf denoting its root mean square bandwidth given by (2.17), we finally arrive at

$$\sigma_{\hat{\tau}}^2 \geq \frac{N_0\, T}{(2\pi\Delta f)^2}. \tag{2.55}$$

We conclude that it is advantageous for arrival time estimation to have a broadband target signal in a narrow observation time interval. This is why partial onset localization based on the application of thresholds to narrowband bandpass filter outputs (as in [Baumann, 1995; Cooke, 1993; Moorer, 1975; Serra, 1989]) is inappropriate. Consequently, the onset detection algorithm incorporated in the architecture presented in Chapter 3 is a wideband approach.

2.4 The Gammatone Filter

The cochlea with the basilar membrane is the central part of the inner ear, where acoustic pressure waves are converted to neural impulse trains. The basilar membrane responds in a frequency selective manner, different frequencies resonating in different regions [Allen, 1985]. A good approximation to the frequency selective behavior of the basilar membrane is the so–called gammatone filter [Lyon, 1996; Slaney, 1993; Patterson *et al.*, 1992; Cooke, 1993], defined by the impulse response

$$g_r(t) = \gamma(n,\lambda)\cdot\epsilon(t)\, t^{n-1}\cdot e^{-\lambda t}\cdot\cos(2\pi f_0 t), \quad n\geq 1, \lambda>0, \tag{2.56}$$

where $\epsilon(t)$ is the unit step function, n the filter order, $\lambda > 0$ the damping factor, $\gamma(n, \lambda)$ some normalization constant and f_0 the center frequency of the filter. See Fig. 2.3 for an example of a gammatone impulse response.

The physiological justification for the gammatone filter is not the main reason why it was chosen as a basic building block for the architecture proposed in this thesis. More importantly, this filter provides a good concentration in the time–frequency plane and a low group delay at low computational costs, if the filter parameters are chosen properly. As extensive use of the gammatone filter is made in this thesis, its most important properties are derived in the following, some of which, like the analytic expressions for time–frequency spread (2.73) [Solbach *et al.*, 1998], equivalent rectangular bandwidth (2.76) and the autocorrelation function (2.99), do not seem to have been published previously for arbitrary filter orders. Under certain conditions that will be discussed in Section 2.4.9.1, the continuous lowpass prototype of the gammatone filter, given by

$$g_{n,\lambda}(t) = \gamma(n,\lambda)\cdot\epsilon(t)\cdot t^{n-1}e^{-\lambda t}. \tag{2.57}$$

can be considered instead of (2.56) without loss of generality.

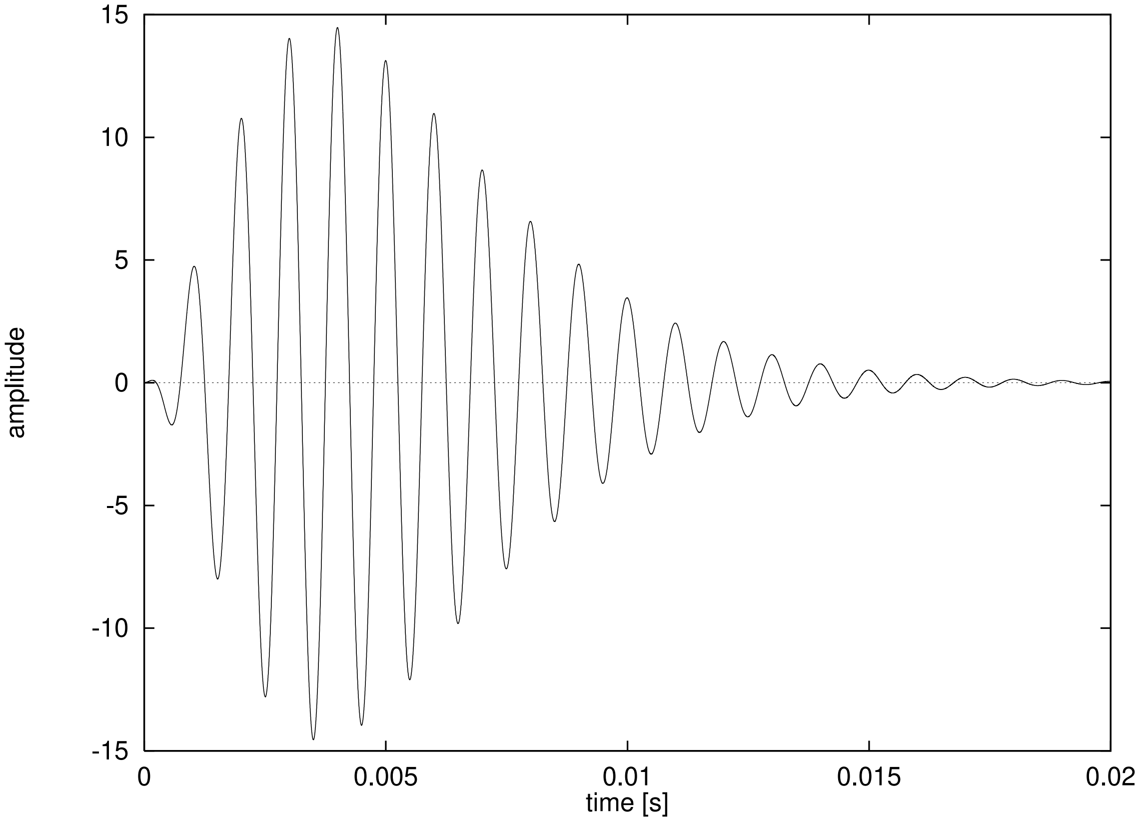

Figure 2.3: Gammatone filter of order $n = 3$, center frequency $f_0 = 1$ kHz, bandwidth $\Delta f = 50$ Hz.

2.4.1 Relation with the Gamma Distribution

If the normalization constant $\gamma(n, \lambda)$ is chosen as

$$\gamma_a(n, \lambda) = \frac{\lambda^n}{\Gamma(n)} \tag{2.58}$$

with the gamma function[8]

$$\Gamma(n) = \int_0^\infty t^{n-1} e^{-t} dt, \quad n > 0, \tag{2.59}$$

such that

$$\int_0^\infty g_{n,\lambda}(t)\, dt = \int_0^\infty \frac{\lambda^n}{\Gamma(n)} \cdot t^{n-1} e^{-\lambda t}\, dt = 1, \qquad \text{for any } n \text{ and } \lambda, \tag{2.60}$$

the gammatone lowpass prototype given by (2.57) is identical to the so–called *gamma distribution* (or *Erlang distribution*) known in statistics as the probability density of

[8]From $\Gamma(1) = 1$ and $\Gamma(n+1) = n\,\Gamma(n)$ follows $\Gamma(n+1) = n!$ for all $n \in \mathbb{N}_0$. This is why the gamma function can be interpreted as a generalization of the factorial function.

the n–th arrival time of a Poisson counting process with arrival rate λ [Davenport, 1970]. It is this identity where the name of the gammatone filter stems from. As we are not dealing with a distribution but with the impulse response of a linear system, we are not restricted to choosing $\gamma(n, \lambda)$ such that (2.60) is satisfied.

2.4.2 Energy

For the energy of the impulse response $g_{n,\lambda}(t)$ we have

$$
\begin{aligned}
E(n, \lambda) &= \int_0^\infty \left(\gamma(n, \lambda) \cdot t^{n-1} e^{-\lambda t}\right)^2 \, dt \\
&= \gamma^2(n, \lambda) \cdot \frac{\Gamma(2n-1)}{(2\lambda)^{2n-1}} \cdot \int_0^\infty \frac{(2\lambda)^{2n-1}}{\Gamma(2n-1)} \cdot t^{(2n-1)-1} e^{-(2\lambda)t} \, dt \\
&= \gamma^2(n, \lambda) \cdot \frac{\Gamma(2n-1)}{(2\lambda)^{2n-1}}, \qquad \text{with (2.60).}
\end{aligned}
\tag{2.61}
$$

Thus, for energy normalization the normalization constant must be

$$
\gamma_e(n, \lambda) = \sqrt{\frac{(2\lambda)^{2n-1}}{\Gamma(2n-1)}}. \tag{2.62}
$$

2.4.3 Frequency Response

The Fourier transform of (2.57) is

$$
\begin{aligned}
G_{n,\lambda}(f) &= \gamma(n, \lambda) \cdot \int_0^\infty t^{n-1} e^{-\lambda t} e^{-j2\pi f t} \, dt \\
&= \gamma(n, \lambda) \cdot \left. \frac{\Gamma(n)}{s^n} \right|_{s=\lambda+j2\pi f} \\
&= \gamma(n, \lambda) \cdot \frac{\Gamma(n)}{(\lambda + j2\pi f)^n}.
\end{aligned}
\tag{2.63}
$$

For the frequency response magnitude we have

$$
|G_{n,\lambda}(f)| = \gamma(n, \lambda) \cdot \frac{\Gamma(n)}{(\lambda^2 + 4\pi^2 f^2)^{\frac{n}{2}}}, \tag{2.64}
$$

With $\gamma(n, \lambda) = \gamma_a(n, \lambda)$ according to (2.58) we have $|G_{n,\lambda}(0)| = 1$, i.e. the case of amplitude normalization.

2.4.4 Unit Step Response

With '$*$' denoting the convolution operator and $\epsilon(t)$ the unit step function, we obtain for $\lambda > 0$, i.e. for a stable filter, by repeated partial integration:

$$\begin{aligned}
\epsilon(t) * g_{n,\lambda}(t) &= \epsilon(t) \cdot \int_0^t \gamma(n,\lambda) \cdot t^{n-1} e^{-\lambda t}\, dt \\
&= -\epsilon(t) \cdot \gamma(n,\lambda) \cdot \frac{\Gamma(n)}{\lambda^n} \cdot \sum_{i=1}^{n} \frac{(\lambda t)^{n-i}}{(n-i)!} \cdot e^{-\lambda t} \Bigg|_0^t \\
&= \epsilon(t) \cdot \gamma(n,\lambda) \cdot \frac{\Gamma(n)}{\lambda^n} \cdot \left(1 - e^{-\lambda t} \cdot \sum_{i=1}^{n} \frac{(\lambda t)^{n-i}}{(n-i)!}\right). \qquad (2.65)
\end{aligned}$$

In terms of statistics theory, (2.65) with $\gamma(n,\lambda) = \gamma_a(n,\lambda)$ according to (2.58) evaluated at t_0 is the probability that the n–th arrival of a Poisson process with arrival rate λ has occurred for $t < t_0$. As expected for the case of amplitude normalization, we have $\lim_{t\to\infty} (\epsilon(t) * g_{n,\lambda}(t)) = 1$.

2.4.5 Impulse Response Amplitude Peak Time

The time t_p where the gammatone filter impulse response reaches its maximum is at the zero of the first derivative of (2.57). With $t > 0$ we get

$$\begin{aligned}
& \frac{d}{dt}\left(t^{n-1} e^{-\lambda t}\right) = 0 \\
\iff \quad & e^{-\lambda t} t^{n-2} \cdot [n - 1 - \lambda t] = 0 \\
\implies \quad & t_p(n,\lambda) = \frac{n-1}{\lambda}. \qquad (2.66)
\end{aligned}$$

2.4.6 Time–Frequency Spread

Without loss of generality we assume energy normalization, that is we set $\gamma(n,\lambda)$ as in (2.62). From (2.14) and (2.15) follows

$$f_0 = 0, \qquad (2.67)$$

and

$$\begin{aligned}
t_0(n,\lambda) &= \int_{-\infty}^{+\infty} t \cdot |g_{n,\lambda}(t)|^2\, dt \\
&= \gamma^2(n,\lambda) \cdot \frac{\Gamma(2n)}{(2\lambda)^{2n}} \cdot \int_0^\infty \frac{(2\lambda)^{2n}}{\Gamma(2n)} \cdot t^{2n-1} e^{-2\lambda t}\, dt \\
&= \gamma^2(n,\lambda) \cdot \frac{\Gamma(2n)}{(2\lambda)^{2n}}, \qquad \text{with (2.60)} \\
&= \frac{2n-1}{2\lambda}, \qquad \text{with (2.62).} \qquad (2.68)
\end{aligned}$$

For the time window width we get using (2.16):

$$\begin{aligned}\Delta t(n,\lambda) &= \sqrt{\int_{-\infty}^{+\infty} (t-t_0)^2 \cdot |g_{n,\lambda}(t)|^2\, dt} \\ &= \sqrt{\int_{-\infty}^{+\infty} |t \cdot g_{n,\lambda}(t)|^2\, dt - 2t_0 \int_{-\infty}^{+\infty} t \cdot |g_{n,\lambda}(t)|^2\, dt + t_0^2 \int_{-\infty}^{+\infty} |g_{n,\lambda}(t)|^2\, dt} \\ &= \sqrt{\frac{\gamma^2(n,\lambda)}{\gamma^2(n+1,\lambda)} - 2t_0^2 + t_0^2} \\ &= \frac{1}{2\lambda} \cdot \sqrt{2n-1}. \end{aligned} \tag{2.69}$$

Using (2.17) we get with $f_0 = 0$ for the frequency window width

$$\begin{aligned}\Delta f(n,\lambda) &= \sqrt{\int_{-\infty}^{+\infty} f^2 \cdot |G_{n,\lambda}(f)|^2\, df} \qquad (2.70) \\ &= \sqrt{\frac{(2\lambda)^{2n-1} \cdot \Gamma^2(n)}{\Gamma(2n-1)}} \cdot \sqrt{\int_{-\infty}^{+\infty} \frac{f^2}{(\lambda^2 + 4\pi^2 f^2)^n}\, df}\end{aligned}$$

From this expression it is apparent that the integral in (2.70) does not converge for $n \le \frac{3}{2}$. It shows that the solution can be found more easily in the time domain, where we get from (2.70) using Parseval's identity:

$$\begin{aligned}\Delta f(n,\lambda) &= \sqrt{\int_{-\infty}^{+\infty} \left| \frac{1}{j2\pi} \frac{d}{dt} g_{n,\lambda}(t) \right|^2 dt} \\ &= \frac{1}{2\pi} \cdot \sqrt{\int_{-\infty}^{+\infty} (n-1-\lambda t)^2 \cdot g_{n-1,\lambda}^2(t)\, dt}, \quad n > \frac{3}{2}. \end{aligned} \tag{2.71}$$

After some straight–forward calculations we arrive at

$$\Delta f(n,\lambda) = \frac{\lambda}{2\pi} \cdot \sqrt{\frac{1}{2n-3}}, \quad n > \frac{3}{2}. \tag{2.72}$$

With (2.69) we finally find for the window area in the time–frequency plane

$$(\Delta f \Delta t)(n) = \frac{1}{4\pi} \sqrt{\frac{2n-1}{2n-3}}, \quad n > \frac{3}{2}, \tag{2.73}$$

visualized in Fig. 2.4.

As expected, the window area depends on the filter order n only. After fixing n to a certain value, the form of the window is determined by the parameter λ. For $n \to \infty$, the window size approaches $\frac{1}{4\pi}$ which is the Gaussian window size (see Section 2.1.2).

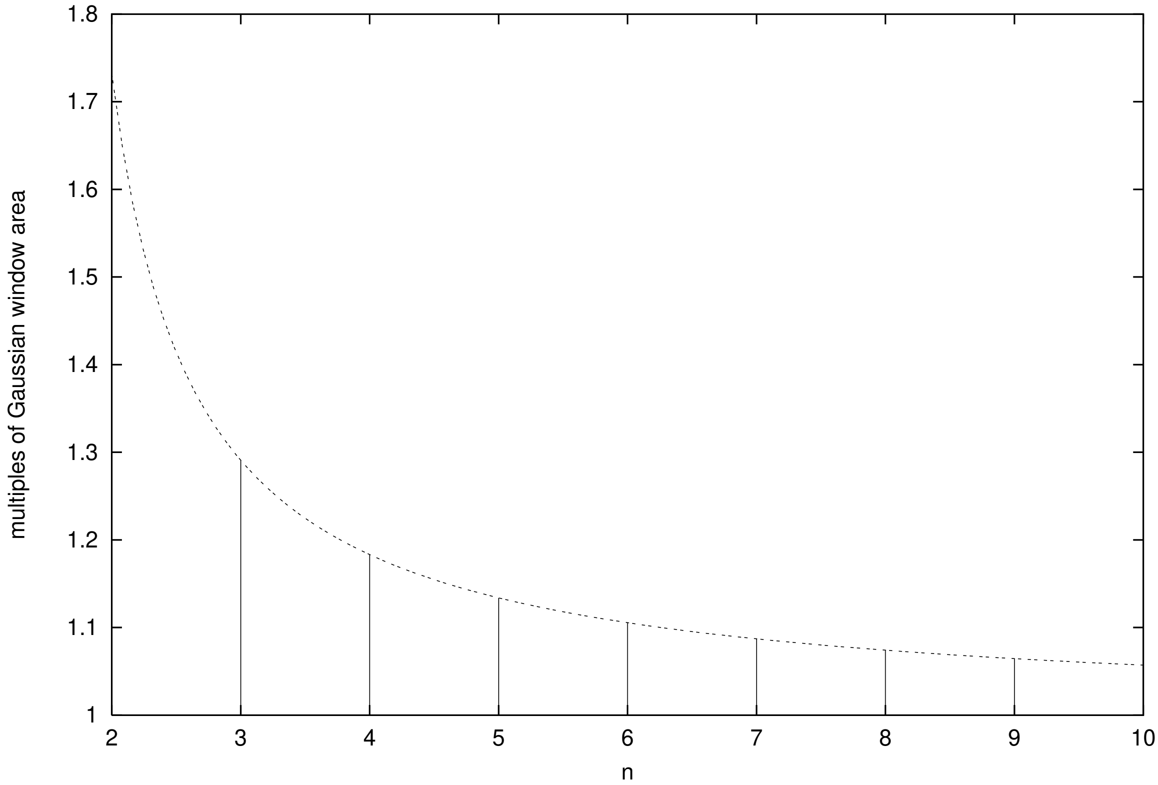

Figure 2.4: Window area of the gammatone filter of order n as given by (2.72). The unit of the y-axis is the area of the Gaussian function.

This result could be expected from the fact that an increase of n in (2.63) is basically (up to a scaling factor) equivalent to the repeated convolution with the first order gammatone lowpass prototype, which, due to the central limit theorem, is known to converge against the Gaussian function.

Time–frequency localization optimality is not the only localization property of interest for analyzing functions. Another interesting property is *time–scale* localization for which a chirp–like variant of the gammatone filter in the form of

$$g_r(t) = \gamma(n, \lambda) \cdot \epsilon(t) t^{n-1} \cdot e^{-\lambda t + j c_0 \log\left(\frac{t}{t_0}\right)}, \tag{2.74}$$

is the optimum solution [Cohen, 1993]. This led Irino [Irino, 1996] to the development of the gammachirp filter, which reportedly exhibits a slightly better approximation to basilar membrane filtering in the human cochlea [Irino and Patterson, 1997], as it provides an asymmetric frequency response. The issue of basilar membrane filter asymmetry is also addressed in Section 2.4.10.

2.4.7 Equivalent Rectangular Bandwidth

Throughout CASA literature the term *equivalent rectangular bandwidth* (ERB) has been used, e.g. in [Lyon, 1996; Patterson *et al.*, 1992; Cooke, 1993; Slaney, 1988]. The results obtained thusfar enable us to calculate a simple analytic expression for the ERB for arbitrary n and λ, a result that does not seem to be available in literature yet. The ERB of a bandpass filter is defined as the width of a rectangular filter with the same peak gain and impulse response energy. From this definition follows

$$ERB_{n,\lambda} \cdot |G_{n,\lambda}(0)|^2 = E(n, \lambda). \tag{2.75}$$

Inserting from (2.61), (2.62) and (2.64) yields

$$\begin{aligned} ERB_{n,\lambda} &= \frac{\Gamma(2n-1)}{2^{2n-1} \cdot \lambda^{2n-1}} \cdot \frac{\lambda^{2n}}{\Gamma^2(n)} \\ &= \frac{\Gamma(2n-1)}{2^{2n-1}\Gamma^2(n)} \cdot \lambda. \end{aligned} \tag{2.76}$$

Now that we have found an analytic expression for the ERB, it is interesting to compare it to the root mean square bandwidth Δf. Inserting for λ from (2.72) yields

$$\frac{ERB_{n,\lambda}}{\Delta f(n, \lambda)} = \frac{\pi\, \Gamma(2n-1)\, \sqrt{2n-3}}{2^{2(n-1)}\, \Gamma^2(n)}, \tag{2.77}$$

visualized in Fig. 2.5. We see that for $n = 3$ we have $ERB_{n,\lambda} \approx 2 \cdot \Delta f(n, \lambda)$.

2.4.8 Group Delay and Phase

One reason why the gammatone filter was chosen as a basic building block for the architecture presented is the possibility to have both short group delay and a good concentration in the time–frequency plane with the filter order n chosen properly. As will be seen in Section 3.3.2, the group delay is a crucial parameter for the distance two onsets must have to be distinguishable. Moreover, as will be shown in Section 3.2.3, the group delay of an adaptive filter has an immediate influence on tracking stability. For the phase of the n–th order gammatone lowpass prototype we have

$$\begin{aligned} \arg[G_{n,\lambda}(f)] &= -\arg\left([\lambda + j2\pi f]^n\right) \\ &= -n \cdot \arctan\left[\frac{2\pi f}{\lambda}\right] \\ &= -n \cdot \arctan\left[\frac{f}{\Delta f \cdot \sqrt{2n-3}}\right], \end{aligned} \tag{2.78}$$

with use of (2.72). The phase is displayed in Fig. 2.6. The group delay $d_{n,\lambda}(f)$ is

$$\begin{aligned} d_{n,\lambda}(f) &= -\frac{1}{2\pi} \cdot \frac{d}{df} \arg[G_{n,\lambda}(f)] \\ &= \frac{n \cdot \lambda}{\lambda^2 + 4\pi^2 f^2}. \end{aligned} \tag{2.79}$$

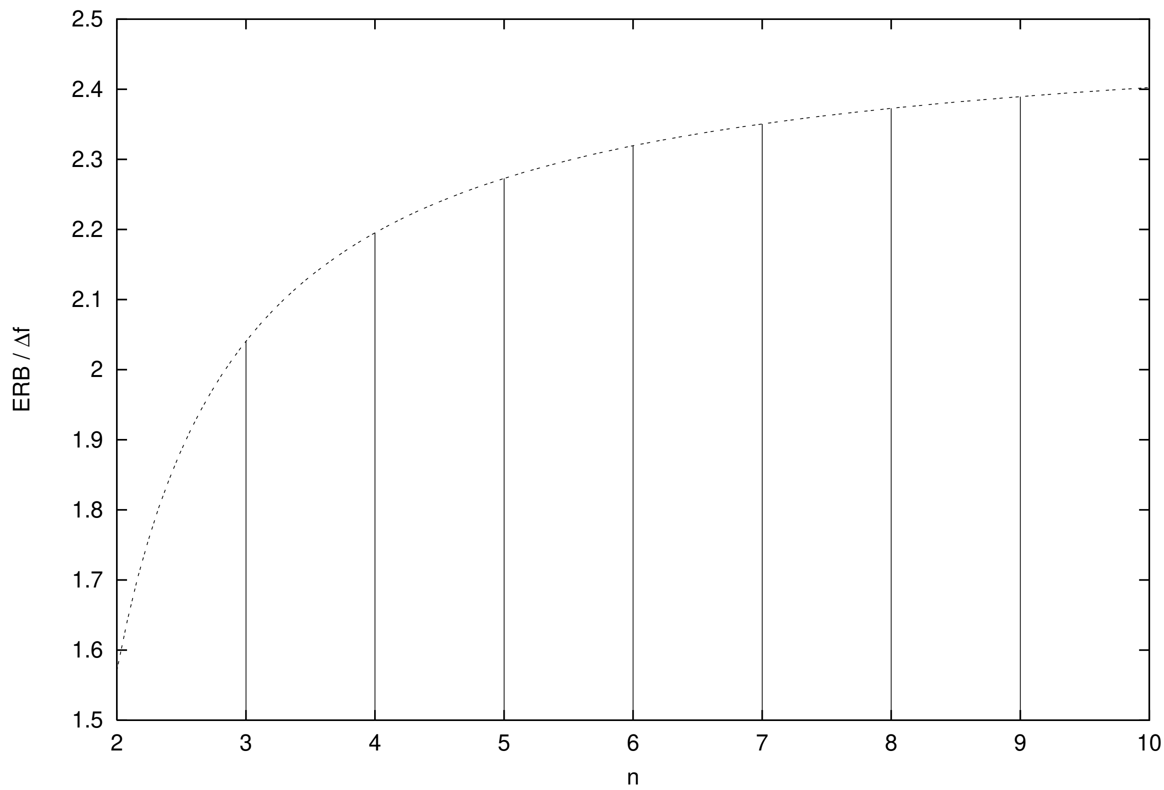

Figure 2.5: $\frac{ERB}{\Delta f}$ depending on the gammatone filter order n.

For the group delay at $f = 0$ we arrive at

$$\begin{aligned} d_{n,\lambda}(0) &= \frac{n}{\lambda} \\ &= \frac{n}{2 \cdot \pi \cdot \Delta f(n,\lambda) \cdot \sqrt{2n-3}}, \text{ with (2.72).} \end{aligned} \tag{2.80}$$

The dependence of the group delay on the filter order for a fixed bandwidth is exemplarily displayed in Fig. 2.7 for $\Delta f = 10$ Hz. For Δf fixed, $d_{n,\lambda}(f)$ reaches its minimum at $n = 3$, since

$$\frac{d}{dn} \cdot \frac{n}{\sqrt{2n-3}} = \frac{n-3}{(2n-3)^{\frac{3}{2}}} = 0 \Rightarrow n = 3 \tag{2.81}$$

and

$$\frac{d^2}{dn^2} \cdot \frac{n}{\sqrt{2n-3}} = \frac{6-n}{(2n-3)^{\frac{5}{2}}} > 0 \text{ for } n = 3. \tag{2.82}$$

As the group delay is minimum for $n = 3$ at a given bandwidth, this filter order is used throughout the architecture presented in Chapter 3 and in the examples given in Chapter 4.

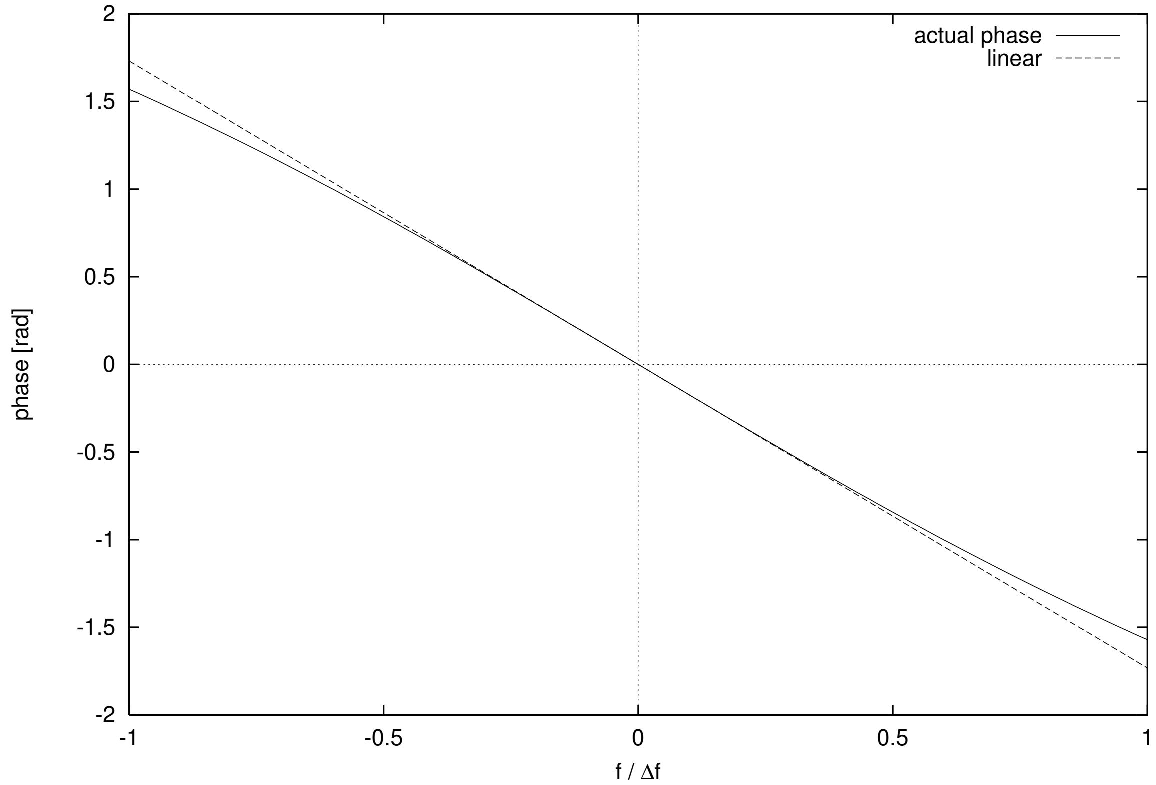

Figure 2.6: Phase for $n = 3$.

2.4.9 Implementation of a Gammatone Wavelet Filter Bank

For realizing a quasi–continuous wavelet transform on a digital computer, the time–scale plane must be sampled on a fine grid. This is most easily achieved by dividing the plane into adjacent bands, each of them represented by a discrete–time equivalent of a continuous–time wavelet filter with a fixed scale parameter.

2.4.9.1 The Gammatone Filter as a Wavelet

Most gammatone implementations employ filters with real–valued impulse responses [Slaney, 1993; Patterson *et al.*, 1992]. Instead, an analytic variant of the gammatone filter is used throughout this thesis. This choice offers a simplified interpretability of the filter outputs in terms of instantaneous amplitude, phase and frequency. A quasi–analytic variant of the gammatone filter is constructed by multiplication of the lowpass prototype impulse response (2.57) with $2 \cdot e^{j2\pi f_0 t}$, yielding

$$g_{\mathbf{k}}(t) = 2 \cdot \gamma(n, \lambda) \cdot \epsilon(t) \cdot t^{n-1} e^{-(\lambda - j2\pi f_0)t}, \tag{2.83}$$

where the parameter vector

$$\mathbf{k} = (n, \lambda, f_0) \tag{2.84}$$

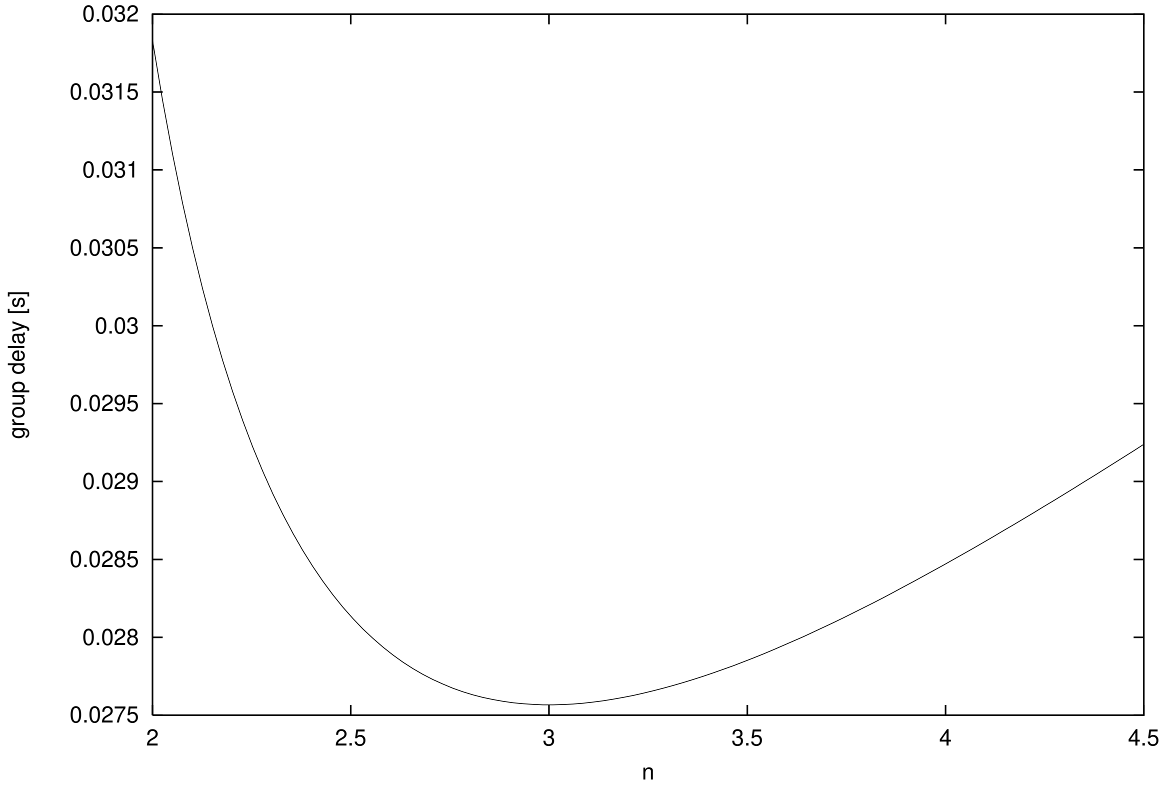

Figure 2.7: Group delay for $\Delta f = 10$ Hz.

is introduced for the sake of compact notation. The factor 2 serves for maintaining normalization in the case of real–valued input signals, since analytic filters cancel out the mirror components residing in the negative frequency halfplane. The Laplace transform of (2.83) is

$$G_{\mathbf{k}}(s) \quad = \quad \gamma(n, \lambda) \cdot \frac{2\,\Gamma(n)}{(s + (\lambda - j2\pi f_0))^n}. \tag{2.85}$$

This filter has an n–fold pole at $-\lambda + j2\pi f_0$. Like the Gaussian wavelet, the resulting filter does not strictly satisfy the conditions for wavelet admissibility (2.8) and analyticity (Def. 2.2), since $\forall f \leq 0 : G_{\mathbf{k}}(f) \neq 0$. However, this effect is negligible for sufficiently small bandwidths $\Delta f(n, \lambda)$. In order to quantify approximate admissibility we compute the quotient of the amplifications at $f = 0$ and $f = f_0$:

$$\begin{aligned} \frac{|G_{\mathbf{k}}(0)|}{|G_{\mathbf{k}}(j2\pi f_0)|} \quad &= \quad \frac{\lambda^n}{(\lambda^2 + 4\pi^2 f_0^2)^{\frac{n}{2}}} \\ &= \quad \frac{1}{\left(1 + \frac{1}{2n-3} \cdot Q^2\right)^{\frac{n}{2}}}, \end{aligned} \tag{2.86}$$

with (2.72) and the filter quality $Q = \frac{f_0}{\Delta f}$. Thus, for gammatone filters with $n = 3$ and relative bandwidths of $Q^{-1} < 0.1$ we have

$$\frac{|G_{\mathbf{k}}(0)|}{|G_{\mathbf{k}}(j2\pi f_0)|} < 0.005, \tag{2.87}$$

a value for which it is justified to consider the residual amplification at $f = 0$ as negligible. For these parameters the gammatone filter is approximately admissible as a wavelet due to (2.8). Furthermore, if $|G_{\mathbf{k}}(0)|$ is already close to zero, the gammatone wavelet is also approximately analytic, since the filter response magnitude decreases monotonically from $f = 0$ to $f \to -\infty$. As a consequence, the analytic gammatone filter according to (2.83) is linked with its real–valued counterpart defined by (2.56) via the Hilbert transform and Corollary 2.1, i.e.

$$g_{\mathbf{k}}(t) \approx 2 \cdot (g_r(t) + j\,\mathcal{H}[g_r(t)])\,. \tag{2.88}$$

2.4.9.2 Realization of a Gammatone Filter

For the realization of a wavelet band centered around a certain frequency f_0, the transfer function of the respective continuous–time gammatone filter must be transformed into a discrete–time equivalent (see Appendix D). In Slaney's [1993] filter bank implementation, the impulse–invariant transform is immediately applied to gammatone bandpass filters. The problem with this method is the occurrence of aliasing components. They increase with the filter center frequency approaching the Nyquist rate. By contrast, Cooke [1993] realizes his gammatone filters as base–band lowpass prototypes. The input signal is shifted in frequency by multiplication with $e^{-j2\pi f_0 kT}$, where f_0 is the bandpass center frequency, and then passed through a gammatone low–pass prototype with the desired bandwidth. Finally, the resulting signal is shifted back to its original frequency location by multiplication with $e^{j2\pi f_0 kT}$. At first sight it might appear as if the aliasing problem had been alleviated, since the impulse–invariant transform is applied to the lowpass prototypes instead of the original bandpasses. However, for resynthesis the signal would have to be shifted to its original frequency location, so we get back to the same aliasing problem, which could only be reduced by introducing an extra upsampling step. These extra computational costs, can be avoided by using the bilinear transform given by

$$s = 2f_s \cdot \frac{z-1}{z+1}, \tag{2.89}$$

as a method for obtaining a discrete–time approximation of the gammatone filter. This approach completely avoids aliasing (see Appendix D). Setting

$$s_0 = \lambda - j2\pi f_0 \tag{2.90}$$

and inserting (2.89) into (2.85) we get

$$G_{\mathbf{k}}(z) = \gamma(n,\lambda) \cdot \frac{2\,\cdot\Gamma(n)}{(2f_s \cdot \frac{z-1}{z+1} + s_0)^n} = \frac{2\,\gamma(n,\lambda)\cdot\Gamma(n)\cdot\alpha^n(z+1)^n}{(z+\beta)^n} \tag{2.91}$$

with

$$\alpha = \frac{1}{s_0 + 2f_s}, \quad \text{and} \quad \beta = \frac{s_0 - 2f_s}{s_0 + 2f_s}. \tag{2.92}$$

For the coefficients a_i and b_i of the n–th order discrete–time gammatone filter

$$G_{\mathbf{k}}(z) = \frac{\sum_{i=0}^{n} a_i \cdot z^{-i}}{\sum_{i=0}^{n} b_i \cdot z^{-i}} \tag{2.93}$$

we find by simple calculation

$$a_i = 2\,\gamma(n,\lambda) \cdot \Gamma(n) \cdot \alpha^n \cdot \binom{n}{i} \quad \text{and} \quad b_i = \beta^i \cdot \binom{n}{i}, \tag{2.94}$$

with $\gamma(n,\lambda)$ given by either $\gamma_e(n,\lambda)$ due to (2.62) for energy normalization or by $\gamma_a(n,\lambda)$ according to (2.58) for amplitude normalization.

If a gammatone filter with a center frequency close to half the sampling rate is realized, the actual peak of the transfer function will not be found at the center frequency but shifted towards a lower value. This is due to the fact, that the bilinear transform causes a distortion of the frequency axis, which is barely noticeable for small frequencies but getting more and more pronounced towards the Nyquist rate. This may be considered as a flaw but it is less critical than the aliasing problems experienced when applying the impulse–invariant transform to high–frequency bandpass filters. From a more physiological oriented viewpoint the arising filter asymmetry might even be desirable (see Section 2.4.10).

The resulting distortion of the frequency axis can be calculated by inserting $z = e^{j2\pi f' T_s}$ and $s = j2\pi f$ into (2.89), yielding

$$f' = \frac{f_s}{\pi} \cdot \arctan\left(\frac{\pi \cdot f}{f_s}\right). \tag{2.95}$$

The top panel of Fig. 2.8 illustrates the distortion of the transfer function for two different sampling rates. In order to compensate for the center frequency shift, we compute f_0 for which f_0' equals the desired frequency and use the result for the calculation of s_0, thus yielding

$$s_0 = \lambda - j2f_s \cdot \tan\left(\frac{\pi \cdot f_0}{f_s}\right) \tag{2.96}$$

instead of (2.90). With this center frequency compensation, the resulting frequency response magnitudes for the given example are shown in the bottom panel of Fig. 2.8. Note, that without additional modification of λ, the narrowing of the filter bandwidth caused by the bilinear transform is even more pronounced after center frequency compensation. It is at least difficult to find analytic correction terms for λ, so in order

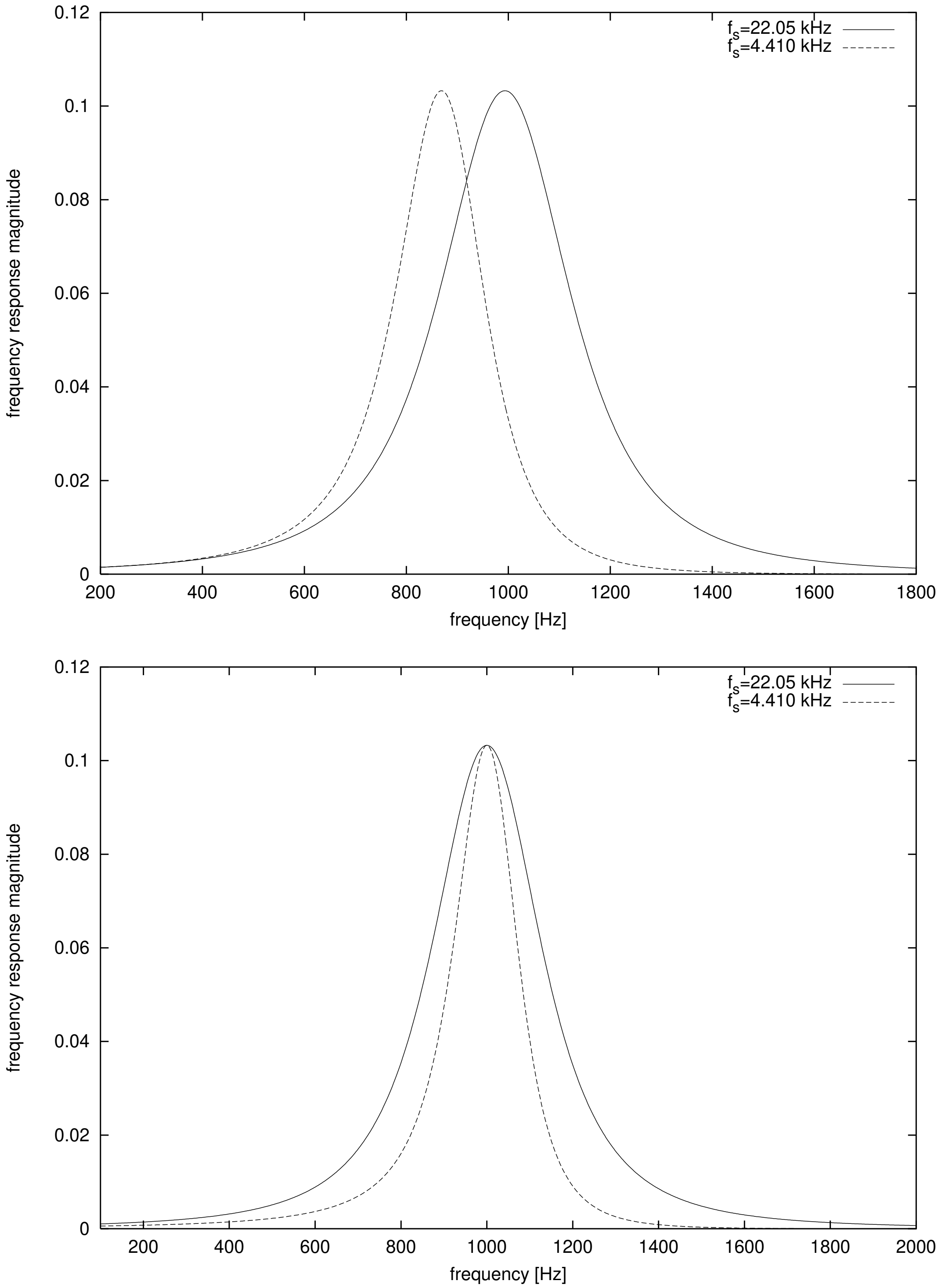

Figure 2.8: Top panel: Influence of the sampling rate on the effective frequency response magnitude for $n = 3$, $\lambda = 500\ \frac{1}{\mathrm{s}}$ and $f_0 = 1$ kHz. Bottom panel: Center frequency compensated frequency response magnitude.

to keep the degenerative effects on relative bandwidth invariability and energy normalization low, we require the sampling rate f_s to exceed f_0 by at least one order of magnitude.

To summarize, the following calculations must be performed after fixing filter order n, center frequency f_0 and bandwidth Δf:

- λ by (2.72),
- $\gamma(n, \lambda)$ by either (2.62) or (2.58),
- s_0 by (2.96),
- α and β by (2.92),
- a_i and b_i by (2.94).

With the increase of the number of poles, digital IIR filters tend to get sensitive to coefficient rounding errors [Chirlian, 1994]. No such problems were experienced with double floating point precision for $n = 3$. However, depending on the number of bits used for coefficient representation it might be necessary to realize an n–th order gammatone filter as a cascade of first and second order filters.

2.4.9.3 Parametrization of the Gammatone Wavelet Filter Bank

Now that we know how to realize a single wavelet band, the question remains how to segregate the time–scale plane into different bands. As the gammatone filter bandwidth Δf is proportional to the damping constant λ, the gammatone filter $g_{\mathbf{k}}(t)$ is scaled by varying λ and f_0, while keeping their quotient constant. Each band can be imagined as covering a strip with a width of $\Delta f_i = f_{0_i} \cdot Q^{-1}$. Thus, for even coverage of the time–scale plane, the distances between adjacent band center frequencies should also be proportional to the center frequencies. This results in a logarithmic filter arrangement (see Fig. 2.9), similar to a musical scale. After fixing the filter order n, the missing parameters necessary to completely characterize the filter bank are specified through

- the relative bandwidth Q^{-1},
- the number of octaves,
- the number of filters per octave N_v,
- the lowermost center frequency $f_{0_{min}}$.

As a rule of thumb we set $Q^{-1} \approx 2^{\frac{1}{N_v}} - 1$ to have the bands just touch each other. According to psychoacoustic findings, the ERB of basilar membrane filters in the human cochlea can be approximated as [Patterson *et al.*, 1992]

$$ERB = 24.7 + 0.10794 \cdot f_0. \tag{2.97}$$

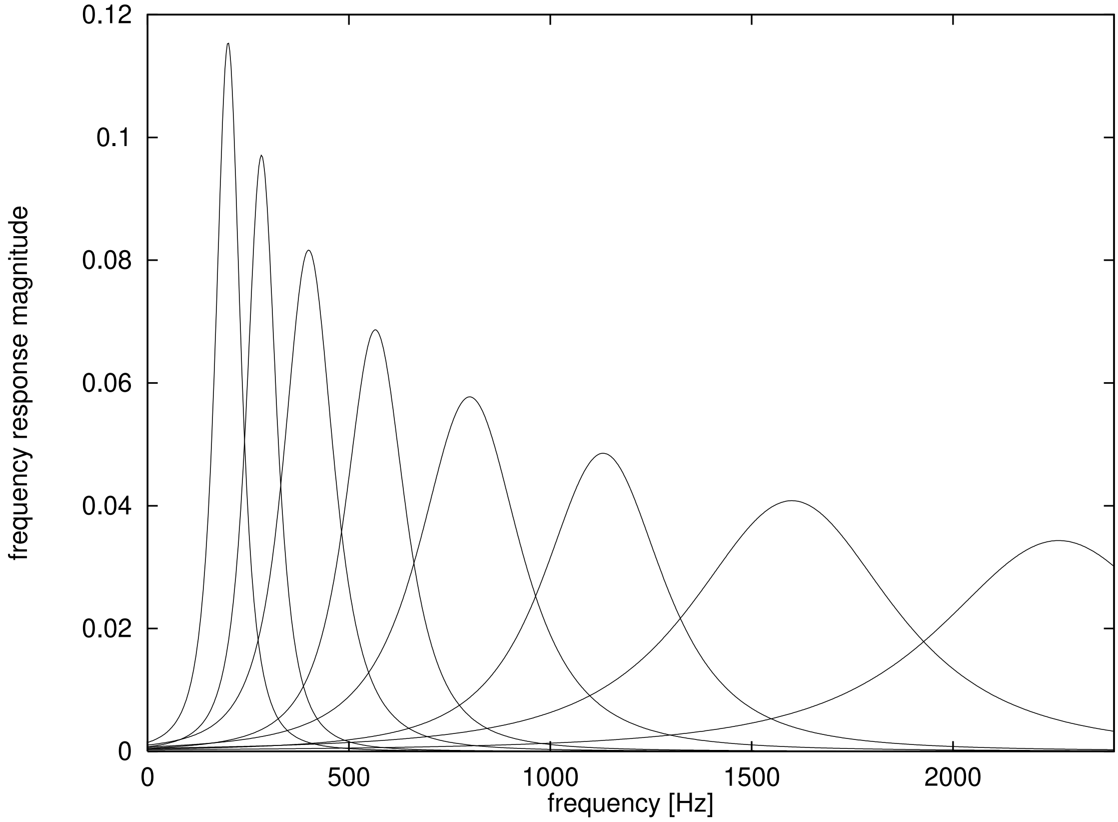

Figure 2.9: Example of an energy–normalized gammatone wavelet filter bank with $n = 3$, frequency response magnitudes.

If we take this equation as a guideline for the construction of our gammatone filterbank, we get with (2.77) for $f_0 \gg 24.7/0.10794$:

$$Q^{-1} = \Delta f / f_0 \approx 0.054. \tag{2.98}$$

Thus, for high center frequencies f_0, the relative bandwidth $\frac{\Delta f}{f_0}$ is roughly one semitone, since $2^{1/12} - 1 \approx 0.0595$.

2.4.10 Asymmetry in Auditory Filtering

The distortion of the frequency axis caused by the bilinear transform (see Section 2.4.9.2) also affects the symmetry property of the gammatone filter. This might be considered a flaw. On the other hand, there has been growing interest in the implementation of asymmetric auditory filters, because the basilar membrane filter slopes tend to be shallow on the low frequency side and steep on the high frequency side of the gain peak location [Zwicker and Fastl, 1990]. This led Lyon [1996] to the development of the all–pole gammatone filter and Irino/Patterson [1997] to the gammachirp filter. These approaches were motivated by mere physiological evidence without trying to answer the question for possible advantages of this asymmetry.

In this respect the asymmetry occurring quite naturally through alias–free sampling at low sampling rates could possibly fill the gap. It is known that the hair cells in the cochlea perform some kind of stochastic sampling [Ghitza, 1992] which is as such subject to the Nyquist criterion. For making optimum use of the neural transmission channels, it is advantageous to perform the sampling at a rate being as low as possible. Thus, in order to avoid aliasing while maintaining a reasonably localized time domain response, the filters must be as steep as possible towards half the sampling rate, trading for a more shallow frequency response on the low frequency side. Interestingly, the frequency response resulting from applying the bilinear transform results in a filter asymmetry bearing a striking similarity to the asymmetric models proposed in literature [Lyon, 1996; Irino and Patterson, 1997], if the center frequency of the original continuous–time gammatone filter surpasses half the sampling rate. An example is shown in Fig. 2.10 (note the double logarithmic scale). There is one more property

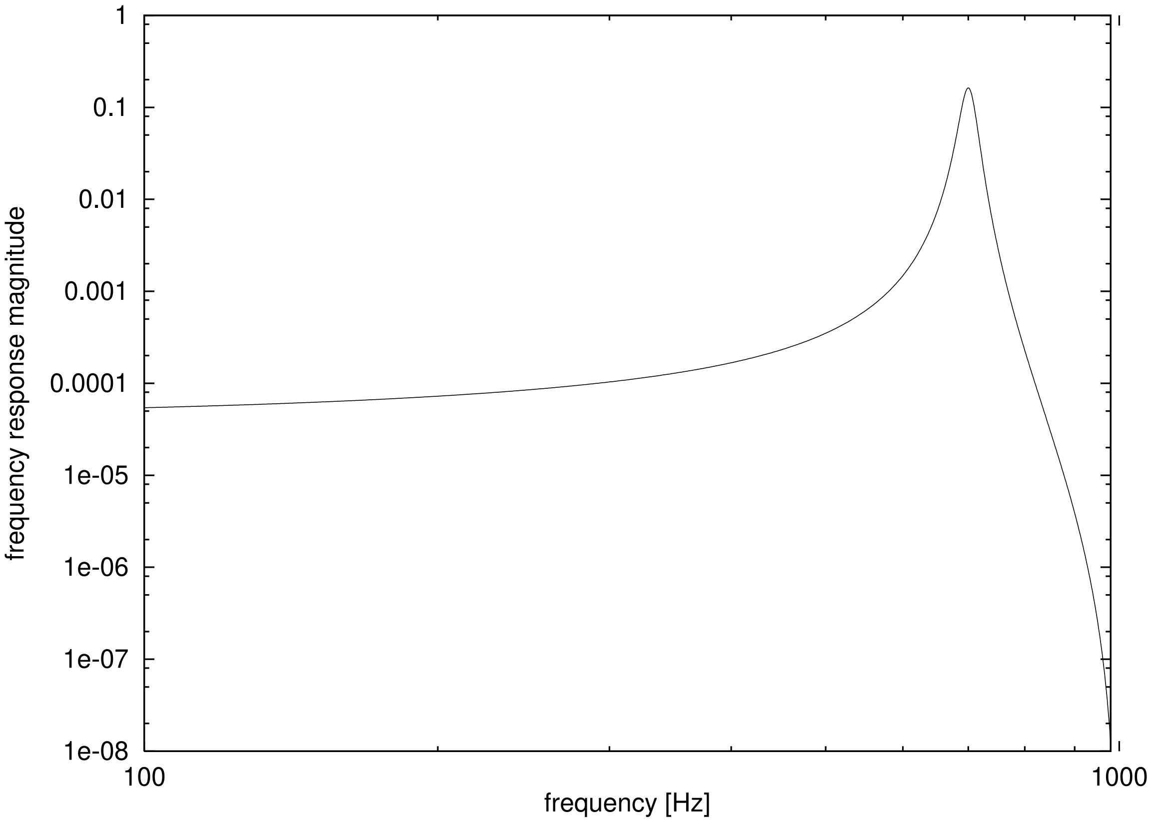

Figure 2.10: Asymmetric gammatone frequency response for $n = 3$, $\lambda = 200.0\ \mathrm{s}^{-1}$, $f_s = 2$ kHz, $f_0 = 1.25$ kHz.

of basilar membrane filtering that could be attributed to sampling effects. While a constant relative bandwidth can be observed over a wide frequency range, this is not true for low frequencies. In this range the passbands are wider than expected. As constant–Q filtering is considered advantageous, one is led to asking for a reason why this property is not maintained for low frequencies. In fact, it is difficult to realize stable, discrete–time IIR filters with low center frequency and narrow bandwidth at high sampling rates due to the influence of rounding errors (see Appendix D). While the basilar membrane is not a discrete–time system itself, it is but part of a dynamic feedback loop involving hair cells as well as afferent and efferent nerve fibers. In this respect the basilar membrane is a part of a system involving stochastic sampling of continuous quantities, so similar stability problems as those occurring in digital computer simulations might arise if the filters get too narrow.

2.4.11 Autocorrelation Function

Be $s_0 = \lambda - j\omega_0$, $m = n - 1$ and $\epsilon(t)$ the unit step at $t = 0$. For the autocorrelation function $\phi_{gg}(\tau)$ of the gammatone wavelet $g_{\mathbf{k}}(t)$ we have

$$
\begin{aligned}
\frac{\phi_{gg}(\tau)}{\gamma^2(n,\lambda)} &= \int_{-\infty}^{\infty} \left[\epsilon(t) t^m e^{-s_0 t}\right]^* \cdot \epsilon(t+\tau)(t+\tau)^m e^{-s_0(t+\tau)}\, dt \\
&= \int_{\frac{|\tau|}{2}}^{\infty} \left(t^2 - \left(\frac{\tau}{2}\right)^2\right)^m e^{-2\lambda t + j\omega_0\tau}\, dt \\
&= e^{j\omega_0\tau} \int_{\frac{|\tau|}{2}}^{\infty} e^{-2\lambda t} \sum_{i=0}^{m} \binom{m}{i} t^{2i} (-1)^{m-i} \left(\frac{|\tau|}{2}\right)^{2m-2i} dt \\
&= e^{j\omega_0\tau} \sum_{i=0}^{m} \left[\binom{m}{i} (-1)^{m-i} \left(\frac{|\tau|}{2}\right)^{2m-2i} \int_{\frac{|\tau|}{2}}^{\infty} t^{2i} e^{-2\lambda t}\, dt\right] \\
&= e^{j\omega_0\tau} \sum_{i=0}^{m} \left[\binom{m}{i} (-1)^{m-i} \left(\frac{|\tau|}{2}\right)^{2m-2i} \int_{0}^{\infty} \left(t + \frac{|\tau|}{2}\right)^{2i} e^{-2\lambda\left(t+\frac{|\tau|}{2}\right)}\, dt\right] \\
&= e^{-\lambda|\tau| + j\omega_0\tau} \sum_{i=0}^{m} \left[\binom{m}{i} (-1)^{m-i} \left(\frac{|\tau|}{2}\right)^{2m-2i} \cdot \right. \\
&\qquad \left. \cdot \sum_{l=0}^{2i} \binom{2i}{l} \left(\frac{|\tau|}{2}\right)^{2i-l} \int_{0}^{\infty} t^l e^{-2\lambda t}\, dt\right] \\
&= e^{-\lambda|\tau| + j\omega_0\tau} \sum_{i=0}^{m} \left[\binom{m}{i} (-1)^{m-i} \sum_{l=0}^{2i} \binom{2i}{l} \left(\frac{|\tau|}{2}\right)^{2m-l} \frac{l!}{(2\lambda)^{l+1}}\right].
\end{aligned}
\tag{2.99}
$$

Chapter 3

System Architecture

The class of asymptotic signals, i.e. the class of signals with a large bandwidth–duration product (see Section 2.3), is of major importance in speech and music. If the signal to be analyzed consists of more than one such component, the separation of contributions originating from different sources is a desirable achievement which is – as human performance indicates – possible in many situations, in which existing artificial systems still fail. The quality of both, time and frequency resolution, is essential for signal separation. As the discussion of time–frequency spread in Section 2.1.2 revealed, satisfactory results can only be achieved by multiresolution approaches. In the architecture described in the following, two resolutions are employed, a narrowband resolution for localizing signals concentrated in frequency (cf. Section 2.2) and a wideband one for those concentrated in time (cf. Section 2.3). Both resolutions are realized using analytic gammatone filters of order $n = 3$ (see Section 2.4). This type of filter was chosen because of its low group delay, high time–frequency concentration, low computational complexity and physiological justification. An important feature of the architecture is adaptive feedback cancellation. As partials are continuously removed from the system input, onset localization and noise floor estimation are facilitated. All system thresholds are automatically updated during operation without the need of manual interception.

3.1 Architecture Overview

The architecture presented in the following is based on the signal model of partials with a pronounced onset. In the ideal case, the onset is maximally concentrated in time while the partial itself is maximally concentrated in frequency. If the onset is not concentrated in time, the partial is still detected by the system and is assigned a certain onset time, but for obvious reasons this time cannot correspond well with physical reality in this case. There are two types of modules in the system. While *partial trackers* (PTs) combined in *tracker groups* (TGs) take care of the partials, the *master module* is responsible for estimating the noise floor, localizing onsets and controlling the initialization and removal of PTs. There is exactly one master module in the system, while the number of PTs changes dynamically. A detailed diagram of the system for the case of one tracker group is shown in Fig. 3.1.

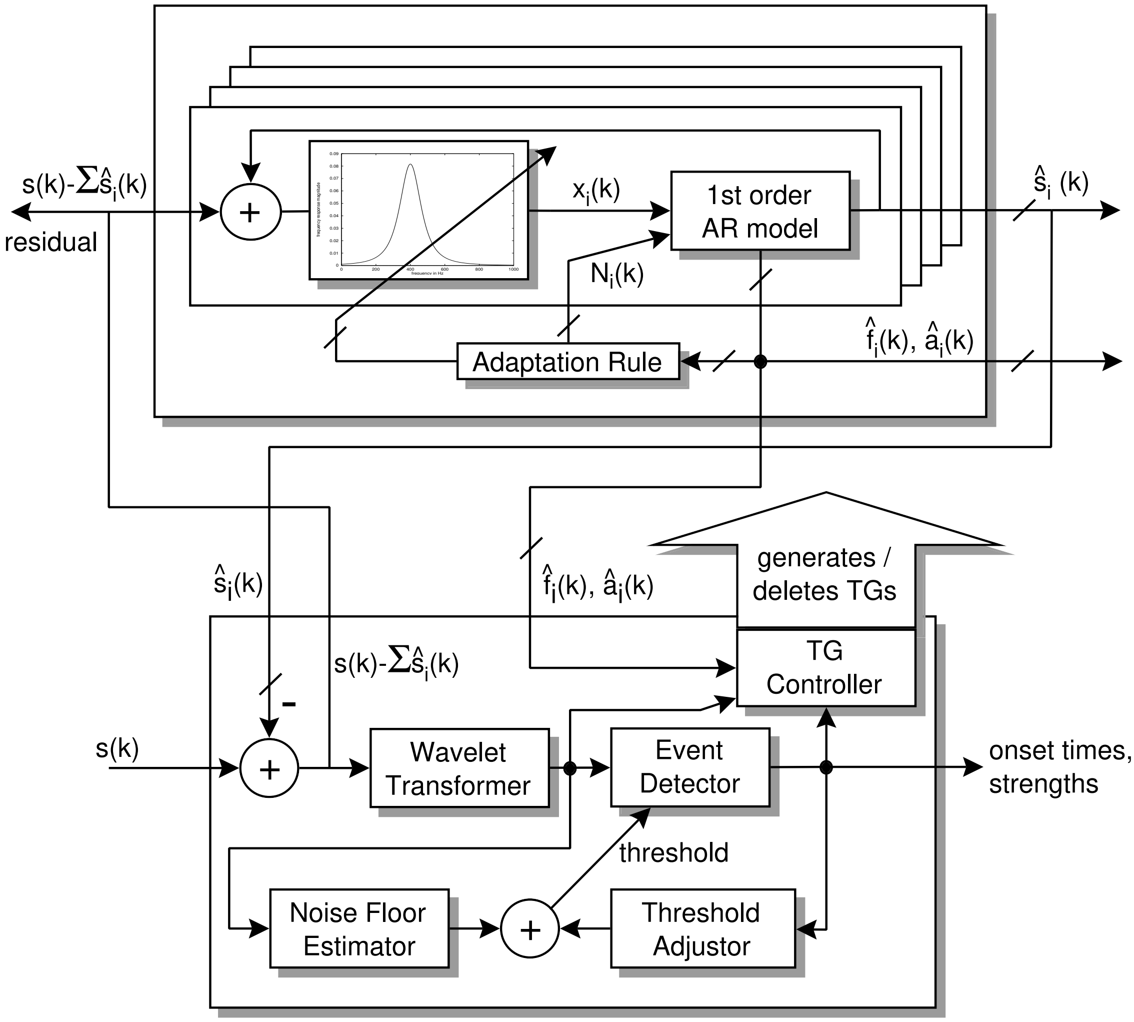

Figure 3.1: Diagram of the overall system, tracker group (top) and master module (bottom).

At sample index $k = 0$, the system only consists of the *master* receiving the overall input signal $s(k)$. In the following, the master generates TGs triggered by the detection of onsets. Shortly after an onset has been found, a partial localization algorithm, which is to be explained in Section 3.3.7, calculates rough estimates of partial locations. With these rough estimates a TG is instantiated after a backtracking step. The PTs of the TG contain tracking filters that are instantiated as one–to–one copies of the wavelet filters residing closest to the estimated frequency locations. Subsequently, each PT is responsible for a single partial, returning the current partial parameters and the predictions for the next partial sample $\hat{s}_i(k+1)$ to the master. The master collects all

$\hat{s}_i(k+1)$ and forms a residual $s(k) - \sum_i \hat{s}_i(k)$ which is then broadcast to the PTs, each PT_i reconstructing $s_i(k)$ at its input. Based on the results of the previous chapter, the functioning of the different blocks and the mechanisms of their interaction will be explained in the following subsections.

3.2 The Tracker Group

The structure of a tracker group (TG) is shown in Fig. 3.2. TGs consist of PTs, whose initial number and parameters are determined by the master module. The tracking

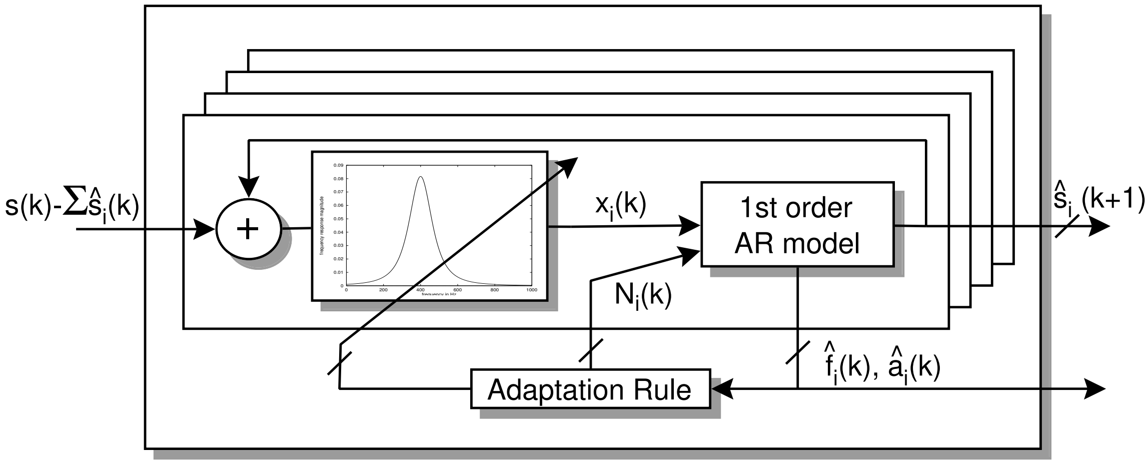

Figure 3.2: Tracker group (TG) with partial trackers (PTs).

filters are constant–Q analytic gammatone filters of order $n = 3$. The filter coefficients are updated on a sample–wise basis depending on the partial parameters extracted by a one–pole AR model estimator. The adaptation rule is a gradient descent scheme aiming at keeping the filter center frequency at the instantaneous frequency of the partial being tracked. The number $N_i(k)$ of samples employed for AR model estimation is continuously updated such that it is kept proportional to the time window width of the tracking filter. As opposed to the widely used approaches involving the STFT or a bank of filters with fixed coefficients, this scheme allows for the continuous tracking of partials with varying frequency without the necessity of interpolation between adjacent time frames and frequency bins.

3.2.1 Partial Parameter Estimation

In the single partial case, the signal $s(k)$ can be assumed to be of the form

$$s_i(k) = a_i(k) \cdot \cos\left(2\pi k T_s f_i(k) + \psi_0\right) + n(k), \tag{3.1}$$

where $n(k)$ is Gaussian white noise of variance σ_n^2. The tracking filter shown in Fig. 3.2 at the entrance of each PT is an amplitude–normalized analytic gammatone filter. If the signal satisfies (2.26) and if $f_i(k) = f_{0_i}(k)$, with $f_{0_i}(k)$ being the current tracking filter center frequency, we have for the signal at the entrance of the AR model estimator

$$x_i(k) = a_i(k) \cdot e^{j(2\pi k T_s f_i(k) + \psi_0)} + n_i(k), \tag{3.2}$$

where $n_i(k)$ is noise which is now autocorrelated according to equation (2.99) evaluated at kT_s (provided that $f_s \gg f_0$). In the ideal case of perfect tracking, the gammatone filter center frequency will be located at exactly the instantaneous partial frequency. In this case the poles of the tracking filter and the pole of the partial's z-transform are located at the same angle $2\pi f_{0_i}(k)/f_s$ in the z–plane. A first order complex–valued AR model is used to estimate the pole location from which $f_{0_i}(k)$ is estimated and a prediction $\hat{s}_i(k+1)$ for the next input sample is derived.

The first order complex–valued AR model is given by the transfer function

$$H_i(z) = \frac{1}{1 - h_i \cdot z^{-1}}. \tag{3.3}$$

with the impulse response

$$h_i(k) = \epsilon(k) \cdot h_i^k. \tag{3.4}$$

The problem formulation in (3.1) is inherently nonstationary, which is why the so–called *autocorrelation method* (see Appendix F) is an inappropriate choice for the estimation of $h_i(k)$. Instead, we use the forward prediction part of the *autocovariance method*. Considering the $N > 2$ samples from $x_i(k-N-1)$ to $x_i(k)$, (F.5) becomes

$$\boxed{\hat{h}_i(k) = \frac{\sum_{l=k-N+2}^{k} x_i(l) x_i^*(l-1)}{\sum_{l=k-N+2}^{k} x_i(l-1) x_i^*(l-1)}}. \tag{3.5}$$

The number of samples N_i used for the estimation is continuously adapted to the current time window size of the tracking filter Δt on a sample–wise basis. The expression in (3.5) can be efficiently evaluated by only two complex–valued multiplications, two additions and one division per sample. Thus, in this case the computational burden for the *least mean squares (LMS)* algorithm [Clarkson, 1993] is not significantly lower, since for the error calculation $e(k) = x_i(k) - h_i(k) \cdot x_i(k-1)$ and the coefficient update $h_i(k+1) = h_i(k) + c \cdot e_i(k) \cdot x_i^*(k)$, with some adaptation constant $c \in \mathbb{R}$, two complex–valued multiplications, two complex–valued additions and one real/complex multiplication would have to be evaluated.

As the poles modelling the signal are transformed into the z–plane via $z = e^{sT_s}$, the pole location estimate $\hat{s}_i(k) = -\hat{\lambda}_i(k) + j2\pi \hat{f}_i(k)$ gets mapped to

$$\hat{h}_i(k) = e^{(-\hat{\lambda}_i(k) + j2\pi \hat{f}_i(k))T_s}, \tag{3.6}$$

so the instantaneous frequency estimate is

$$\boxed{\hat{f}_i(k) = \frac{f_s}{2\pi} \cdot \arg\left[\hat{h}_i(k)\right]}. \tag{3.7}$$

As stated previously, there are two kinds of poles involved in the constitution of $x_i(k)$, the poles of the signal $s_i(k)$ and the poles of the gammatone filter. As the adaptation rule aims at keeping the center frequency of the filter where the dominant signal frequency resides, the center frequency of the tracking filter is ideally identical to the signal frequency, resulting in a phase shift of zero. Thus, we have $s_i(k) = x_i(k)$ for an amplitude–normalized tracking filter and the prediction delivered back to the master is

$$\boxed{\hat{s}_i(k+1) = \mathrm{Re}\left[\hat{h}_i(k) \cdot x_i(k)\right]}. \tag{3.8}$$

As shown previously (see (2.27)), the instantaneous bandwidth of a first order continuous–time linear system is identical to the modulus of the real part of its pole location, which allows for an estimation of the instantaneous bandwidth via

$$\boxed{\Delta\hat{f}_i(k) = \frac{1}{2\pi} \cdot \left|-\hat{\lambda}_i(k)\right| = \frac{f_s}{2\pi} \cdot \left|\log|\hat{h}_i(k)|\right|}. \tag{3.9}$$

Finally, the instantaneous amplitude is estimated as

$$\boxed{\hat{a}_i(k) = \left|\hat{h}_i(k) \cdot x_i(k-1)\right|}. \tag{3.10}$$

If $x_i(k)$ consisted of white noise only, $E\{\hat{a}_i(k)\}$ would be zero, since the expectation value of the numerator in (3.5) would be zero. Thus, contrary to the method of modulus averaging used in [Wang, 1994], there is no bias induced if the noise is white.

3.2.2 The Effect of Feedback Cancellation

The choice of a first order AR model imposes the signal model (3.3) for each partial. With two PTs present, both of them subtract their prediction $\hat{h}_i \cdot s_i(k-1)$ from the input signal $s(k)$. With $S_i(z)$ denoting the z–transform of $s_i(k)$, the resulting input signal of PT_1 has a z–transform of

$$S_1(z) = S(z) \cdot \left(1 - \hat{h}_2 \cdot z^{-1}\right) \tag{3.11}$$

and the one of PT_2 is

$$S_2(z) = S(z) \cdot \left(1 - \hat{h}_1 \cdot z^{-1}\right). \tag{3.12}$$

We see that PT_1 produces a zero of $z = \hat{h}_1$ at the input of PT_2 and vice versa. For an arbitrary number of PTs, the z–transform of the input at the i–th PT is

$$S_i(z) = S(z) \cdot \prod_{j \neq i} (1 - \hat{h}_j \cdot z^{-1}). \tag{3.13}$$

If $s(k)$ is indeed an AR process, $S(z)$ takes the form of (F.1). By factorizing the denominator, $S(z)$ can be written as

$$S(z) = \frac{G}{\prod_j (1 - h_j \cdot z^{-1})}, \tag{3.14}$$

with some $G \in \mathbb{R}$. Inserting (3.14) into (3.13) and assuming that $\forall i : \hat{h}_i = h_i$, the resulting input at tracker i is

$$S_i(z) = \frac{G}{1 - h_i \cdot z^{-1}}, \tag{3.15}$$

i.e. each PT only sees an input signal with a single pole, namely the pole of the partial it is currently tracking.

As a consequence of feedback cancellation, each PT strives for suppressing its partial components at the input of all competitors. Interestingly, there is also evidence for an adaptive feedback mechanism in the human auditory system. Several physiologically motivated models contain such a loop [Meddis *et al.*, 1990]. Inhibitive coupling mechanisms have been found to play a role not only in human hearing [Zwicker and Fastl, 1990; Beckenbauer, 1989] but also in vision [Shamma, 1993]. Two examples for artificial systems not being physiologically motivated but making use of inhibitive coupling are described in [Martin and Padmanabhan, 1993] and [Ramalingam and Kumaresan, 1994].

3.2.3 Adaptation Rule

In the following, a condition for adaptation stability in the single–partial case is derived. These considerations are inspired by the results given by Wang [1994] for the frequency locked loop, but the somewhat lengthy calculations there are considerably simplified. It is important to note that if there is more than one PT present at a time and two of them get too close to each other the following considerations are invalid. In this case the system can become unstable even though the stability condition is satisfied for each PT separately. In order to prevent this happening, a spectral masking mechanism ensures one of the PTs approaching one another to be deleted (see Sections 3.3.6–3.3.8).

The center frequency $f_0(k)$ of the tracking filter is adapted according to the rule

$$\boxed{f_0(k+1) = f_0(k) + g \cdot \left[\hat{f}(k-d) - f_0(k-d)\right], \quad g > 0}, \tag{3.16}$$

where $\hat{f}(k)$ is the estimated signal frequency, g is the adaptation constant and $d \in \mathbb{N}$ is the group delay of the PT in samples, which in turn depends on the current tracking filter bandwidth but is for now assumed to be constant. Demanding $g > 0$ is a necessary condition for f_0 to approach the frequency estimate $\hat{f}$, but it is not sufficient, as the stability of the adaptation process is not guaranteed for large values of g. With the z–transform of (3.16) we have

$$\frac{F_0(z)}{\hat{F}(z)} = \frac{g}{z^{d+1} - z^d + g}, \tag{3.17}$$

equivalently represented by the diagram visualized in Fig. 3.3. In the case of $d = 0$ we

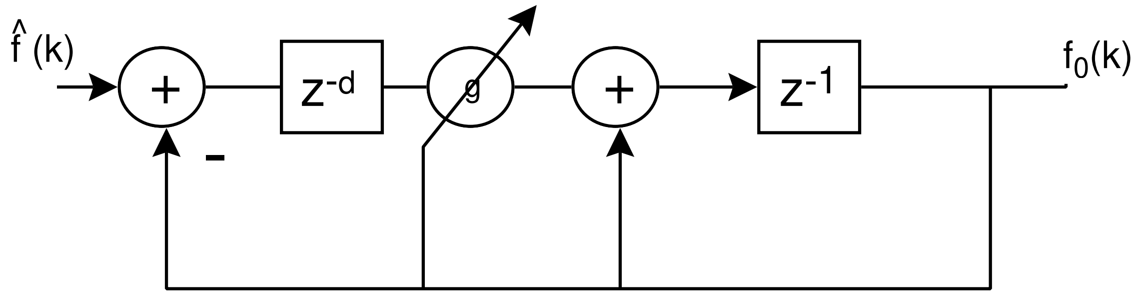

Figure 3.3: Adaptation rule.

have a system with a single pole at $1-g$. As stability requires all poles to lie within the unit circle, this system would be stable for $0 < g < 2$. However, as nontrivial filters with zero group delay are not realizable, we have to live with a smaller admissible range for g. For the general case of nonzero group delay we have to find g such that

$$z^{d+1} - z^d + g = 0 \tag{3.18}$$

has solutions for $|z| < 1$ only. For $g = 0$, one pole is on the unit circle. For very small $g > 0$, all poles lie within the unit circle, but as g grows, there are poles approaching the unit circle again. We look for the value of g, where the unit circle is hit. In this case we may set $z = e^{j\omega T}$ in (3.18), yielding

$$\begin{aligned} & e^{j\omega Td}\left(e^{j\omega T} - 1\right) + g = 0 \\ \Leftrightarrow\quad & e^{j\omega Td} e^{\frac{j\omega T}{2}} \cdot \left(e^{\frac{j\omega T}{2}} - e^{\frac{-j\omega T}{2}}\right) + g = 0 \\ \Leftrightarrow\quad & e^{j\omega T\left(d+\frac{1}{2}\right)} \cdot 2j \sin\left(\frac{\omega T}{2}\right) + g = 0 \\ \Leftrightarrow\quad & 2 \cdot e^{j\left(\omega T\left(d+\frac{1}{2}\right)+\frac{\pi}{2}\right)} \cdot \sin\left(\frac{\omega T}{2}\right) + g = 0. \end{aligned} \tag{3.19}$$

As $g \in \mathbb{R}$ and $g > 0$, this equation can only be satisfied, if

$$\begin{aligned} & \omega T\left(d + \frac{1}{2}\right) + \frac{\pi}{2} = k\pi, \quad k \in \mathbb{Z} \\ \Leftrightarrow\quad & \frac{\omega T}{2} = \frac{k\pi - \frac{\pi}{2}}{1 + 2d}, \quad k \in \mathbb{Z}. \end{aligned} \tag{3.20}$$

Inserting (3.20) into (3.19) we have

$$(-1)^k \cdot 2\sin\left(\frac{k\pi - \frac{\pi}{2}}{1+2d}\right) + g = 0. \tag{3.21}$$

As we have $d \in \mathbb{N}$, $\sin\left(\frac{k\pi-\frac{\pi}{2}}{1+2d}\right)$ is cyclic with period length $2+4d$. One of the infinitely many k, for which we arrive at the smallest positive g satisfying (3.21) is $k=0$. This leads to

$$g = 2\sin\left(\frac{\frac{\pi}{2}}{1+2d}\right), \tag{3.22}$$

which is the upper bound for g, ensuring the stability condition (3.18) is satisfied. Thus, finally we have for the admissible range of g:

$$0 < g < 2\sin\left(\frac{\frac{\pi}{2}}{1+2d}\right). \tag{3.23}$$

The upper bound for g is visualized in Fig. 3.4. Setting $d=0$ in (3.23), we arrive at

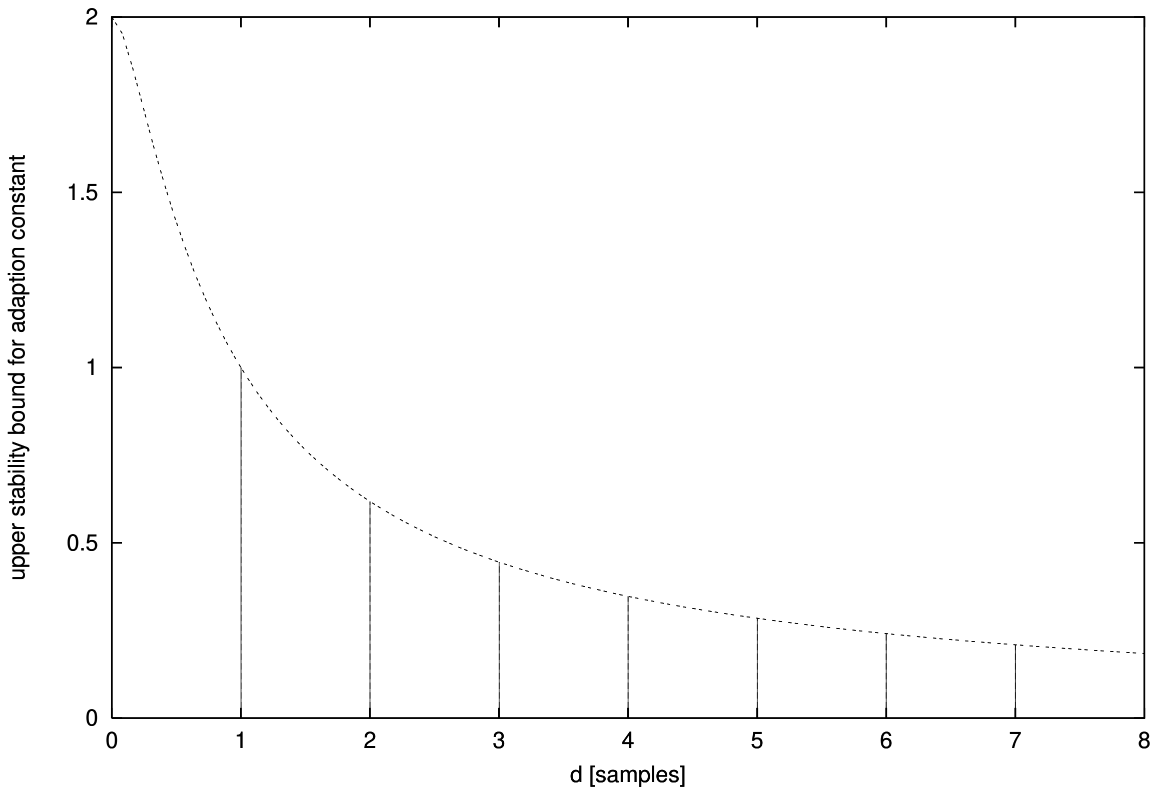

Figure 3.4: Upper bound for g depending on the PT group delay d.

$0 < g < 2$, a result already found above for zero group delay. If the group delay is increased, the upper bound for the tracking gain g is diminished. For $d \to \infty$, stable tracking becomes impossible. Obviously, minimum group delay of the PT is desirable in order to implement maximum adaptation gain. This is the reason why, once the decision for the gammatone filter is made, an order of $n = 3$ should be chosen, because this is where the gammatone filter group delay is minimum for a given bandwidth (see Section 2.4.8).

As the adaptation is carried out on a sample–wise basis, it is preferable not to evaluate a transcendental function for adaptation gain calculation. With

$$\frac{2}{\pi} \cdot x < \sin(x), \text{ for } \quad 0 < x < \frac{\pi}{2} \tag{3.24}$$

it is safe to set

$$\boxed{g = \frac{2}{1 + 2d}}. \tag{3.25}$$

Once the new center frequency $f_0(k+1)$ is set according to (3.16), the new bandwidth $\Delta f(k+1)$ is determined by the system parameter $Q^{-1} = \frac{\Delta f}{f_0}$. However, modifying the bandwidth implies also changing the filter group delay, so consequently the adaptation constant must be adjusted in order to maintain stability. This results in a time–varying $g(k)$ instead of a fixed g. Another consequence of bandwidth adaptivity is that the new time window width of the tracking filter $\Delta t(k+1)$ will generally differ from $\Delta t(k)$. In accordance with the time–frequency trade–off (see Section 2.1.2) we gain time resolution as frequency resolution is lost and vice versa. The benefit of the bandwidth adaptation mechanism would be rendered useless if the window size of the AR model estimator was not also adaptive. Thus, the number of most recent samples considered by the estimator is updated according to

$$\boxed{N(k) = \text{ceil}\left(\kappa \cdot f_s \cdot \Delta t(k)\right) = \text{ceil}\left(\sqrt{\frac{2n-1}{2n-3}} \cdot \frac{\kappa \cdot f_s}{4\pi \cdot Q^{-1} \cdot f_0(k)}\right)}, \tag{3.26}$$

where $\Delta t(k)$ is rewritten using (2.73), ceil(x) is a function rounding x upwards to the nearest integer and κ is a system constant, which is generally set to $\kappa = 1$ in the remainder of this thesis if not stated otherwise.

The contribution of the rectangular estimator window to the PT group delay is $N(k)/2$. The group delay of the gammatone tracking filter is given by (2.80). With (2.72) we arrive at a total PT group delay of

$$\boxed{d(k) = \text{ceil}\left(\frac{Q \cdot f_s \cdot \left(\kappa \cdot \sqrt{2n-1} + 2n\right)}{4\pi \cdot f_0(k) \cdot \sqrt{2n-3}}\right)}. \tag{3.27}$$

With these results we have the following calculations are performed for each PT at each adaptation step:

1. center frequency $f_0(k)$ by (3.16),
2. tracking filter bandwidth $\Delta f(k) = f_0(k) \cdot Q^{-1}$,
3. filter coefficients of the gammatone tracking filter as described in Section 2.4.9.2,
4. AR model estimator window size $N(k)$ by (3.26),
5. PT group delay $d(k)$ by (3.27),
6. adaptation gain $g(k)$ by (3.25).

3.2.4 Stationary Performance

In the following, the stationary properties of the PT structure are evaluated. For now and in everything that follows in the remainder of this thesis, the reference level of 0 dB is the maximum power of a pure sine at 16 bit resolution.

3.2.4.1 Tracking of a Single Partial in Noise

The signal investigated in the following consists of a sine of 1 kHz and −40 dB amplitude in Gaussian white noise. A single PT is instantiated with the center frequency of its tracking filter close to the sine frequency. After settling, frequency and amplitude estimation errors are determined for 12000 samples. Figure 3.5 shows the amplitude error variance in dB and the frequency error standard deviation in Hz, both in dependence on the signal–to–noise ratio (SNR). The parameter is the relative bandwidth of the gammatone tracking filter, which is tightly linked with the sample number N used for AR model estimation due to (3.26). For $f_0 = 1$ kHz, Table 3.1 gives the associated values for N.

Q^{-1}	0.025	0.05	0.075	0.1
N	91	61	46	23

Table 3.1: Estimator sample number depending on relative bandwidth at $f_0 = 1$kHz for $n = 3$ and $f_s = 22.05$ kHz.

It shows that the frequency error variance depends linearly on the SNR, which is in accordance with the CRBs for amplitude (2.33) and frequency (2.34). Note that the noise considered for the SNR does not include quantization noise and rounding errors. This explains the saturation effects for high SNR clearly visible for the amplitude variance in Fig. 3.5. Comparing the estimator performance with the CRBs, we find that the error variances are lower than one might expect. For instance, for a relative bandwidth of $Q^{-1} = 0.025$ we have $N = 91$ from Table 3.1 and thus with (2.33) and

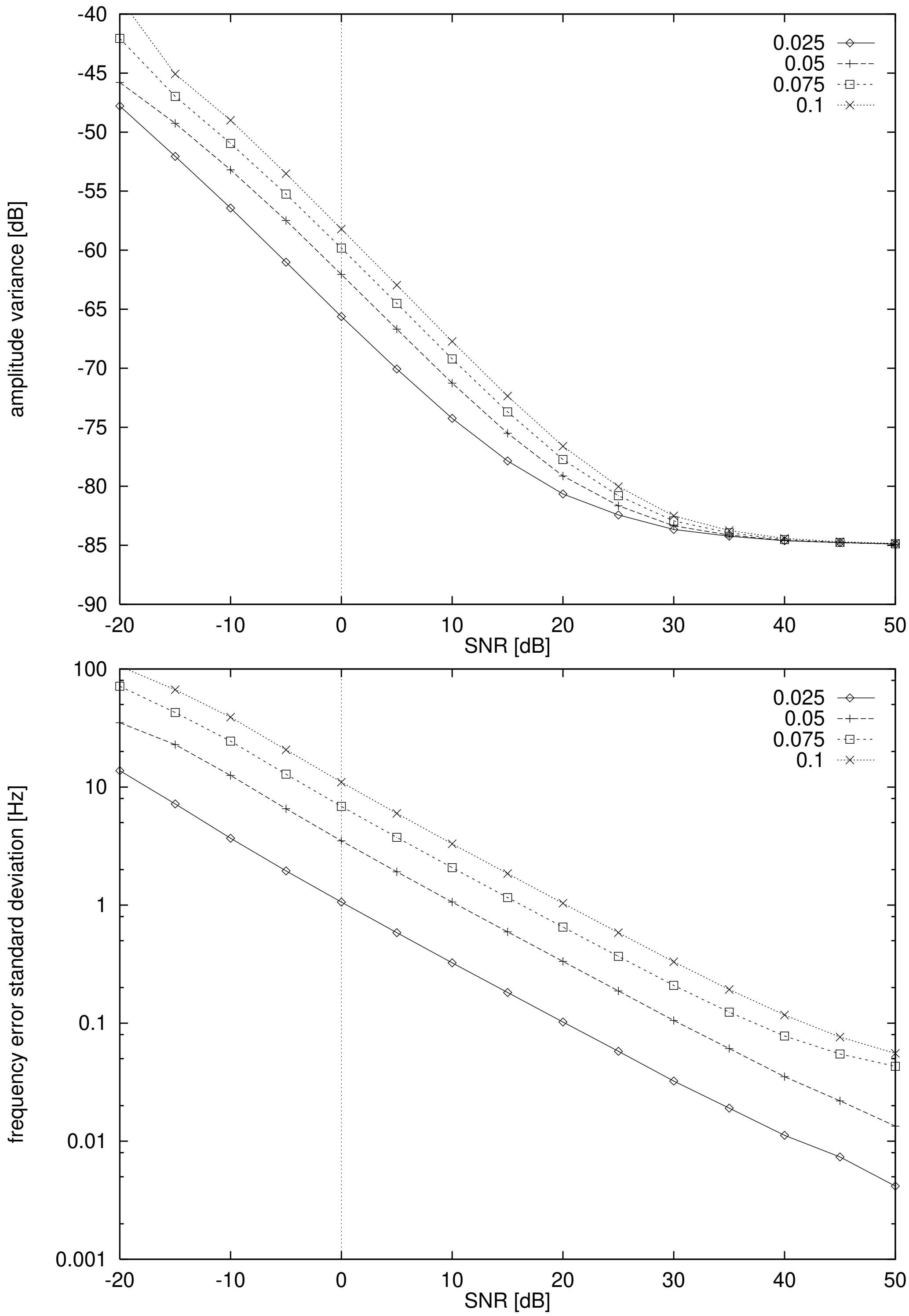

Figure 3.5: Amplitude error variance (top) and frequency error standard deviation (bottom) for a sine at 1 kHz and -40 dB depending on the SNR. Parameter is the PT's $Q^{-1} = \frac{\Delta f}{f_0}$, sampling rate is $f_s = 22.05$ kHz.

considering that the signal is real–valued:

$$10 \cdot \log_{10} \left(\frac{\sigma_A^2}{\sigma_n^2} \right) \geq 10 \cdot \log_{10} \frac{2}{91} = -16.58 \text{ dB}. \tag{3.28}$$

We would expect the amplitude variance for this relative bandwidth and −40 dB noise power to satisfy $\sigma_A \geq -40$ dB – 16.58 dB = −56.58 dB. In the top graph of Fig. 3.5, however, we find at 0 dB SNR an amplitude variance as low as −65.53 dB, which is considerably better. This can be explained by the fact that the estimator shown in Fig. 3.2 comprises both, AR model estimator and gammatone tracking filter. The tracking filter makes the attention of the AR model estimator focus on its passband, leading to a bias in both amplitude and frequency estimation. As the CRB is a lower limit for linear *unbiased* estimators only, this bound does not represent a lower bound for *biased* estimators such as the one described here.

From these considerations we would expect the bias to decrease as the bandwidth widens causing the CRB to be approached from below. In fact, with the same considerations as above, we find for $Q^{-1} = 0.1$:

$$10 \cdot \log_{10} \left(\frac{\sigma_A^2}{\sigma_n^2} \right) \geq 10 \cdot \log_{10} \frac{2}{23} = -10.61 \text{ dB}, \tag{3.29}$$

yielding a CRB of $\sigma_A \geq -40$ dB – 10.61 dB = −50.61 dB, while an actual variance of −58.21 dB was observed (see top graph in Fig. 3.5), a discrepancy which is 1.35 dB below the one for the case of $Q^{-1} = 0.025$ considered above. Similar effects arise for the frequency estimates: For 0 dB SNR and $Q^{-1} = 0.025$ we get $\sigma_f \geq 14.01$ Hz from the CRB in (2.34), but a standard deviation as low as 1.06 Hz is observed. For $Q^{-1} = 0.1$ we find $\sigma_f \geq 27.86$ Hz versus 10.99 Hz.

A bias in estimation is often considered as undesirable. Clearly, the parameters of a partial far beyond the passband of the tracking filter would be estimated very poorly, but from a more positive perspective, such a bias can also be regarded as an unavoidable byproduct of every selective attention mechanism. Advantageously, the estimator presented here is not only biased but also adaptive. Due to the center frequency adaptation mechanism, the selectivity is not constraint to a fixed passband but may adjust to partial frequency changes.

From the expressions for the CRBs it is apparent that the sampling rate f_s should have an immediate influence on the quality of the estimates if the time window size of the tracking filter in seconds is kept constant. In the case of Gaussian white noise, doubling the sampling rate means halving the error variance. However, due to the presence of the tracking filter, the additional samples gained with doubling the sampling rate are not uncorrelated to the ones already present at the lower rate. Nevertheless, with Fig. 3.6 it becomes apparent, that for moderate SNR doubling the sample rate yields an improvement of ~ 2 dB for the error variance.

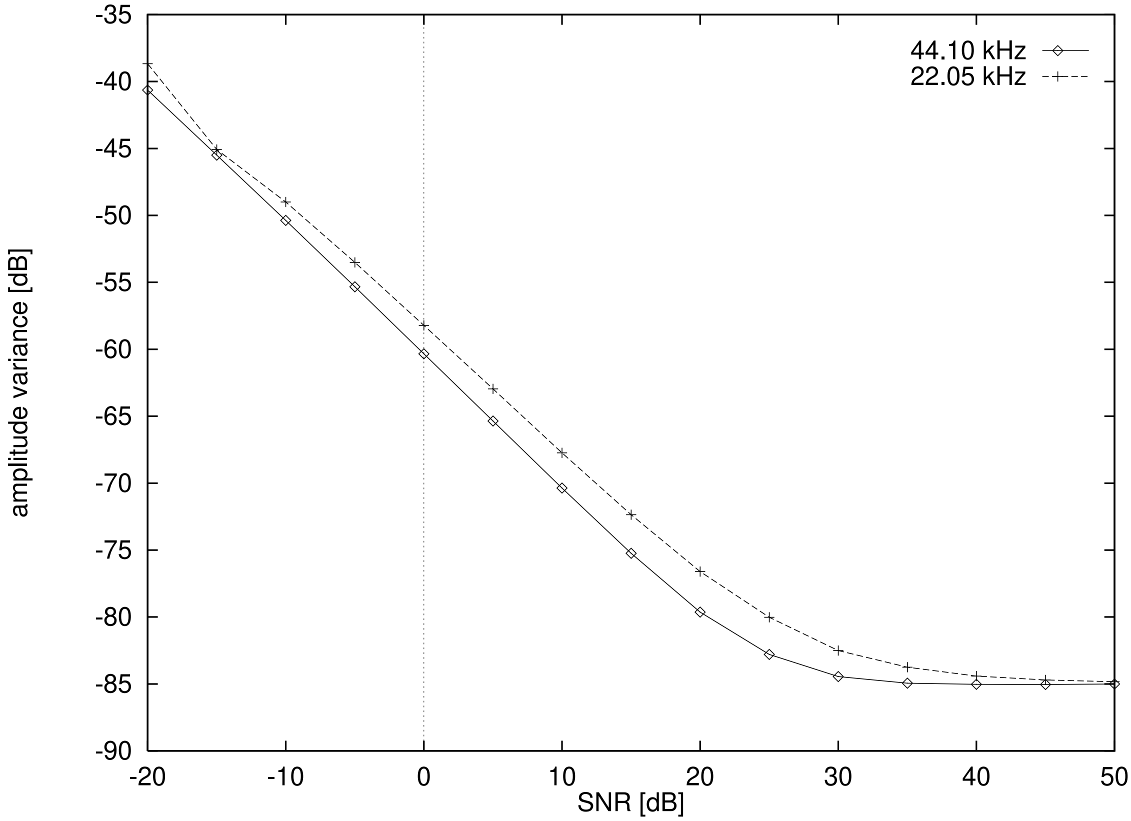

Figure 3.6: Amplitude error variance for two different sampling rates, $Q^{-1} = 0.025$.

3.2.4.2 Crosstalk

The issue of tracking stability in the presence of superimposed partials is of major importance. In this experiment, a sine of 1 kHz with an amplitude of −10 dB and a length of one second is created. A second sine with the same power is added with a delay of 1102 samples and a variable displacement in frequency. The distance is given in *cents*, a pseudo–unit commonly used in music. The frequency distance d_f in cent of a partial with frequency f_2 with respect to another partial with frequency f_1 satisfies the equation

$$\frac{f_2}{f_1} = 2^{\frac{d_f}{1200 \text{ cent}}} \tag{3.30}$$

Thus, a distance of 1200 cent means $f_2 = 2 \cdot f_1$, i.e. a distance of one octave, a distance of 100 cent corresponds to one semitone. Fig. 3.7 shows the error of the frequency estimate for the second sine depending on the displacement, averaged over 17000 samples. The relative bandwidth is $Q^{-1} = 0.05$, translating to 50 Hz or 84 cent for a filter centered around 1 kHz. It shows that if the sines are more than 115 cents apart, they do not influence each other. Experiments with different relative bandwidths yielded a proportional increase of this boundary distance. If the partials

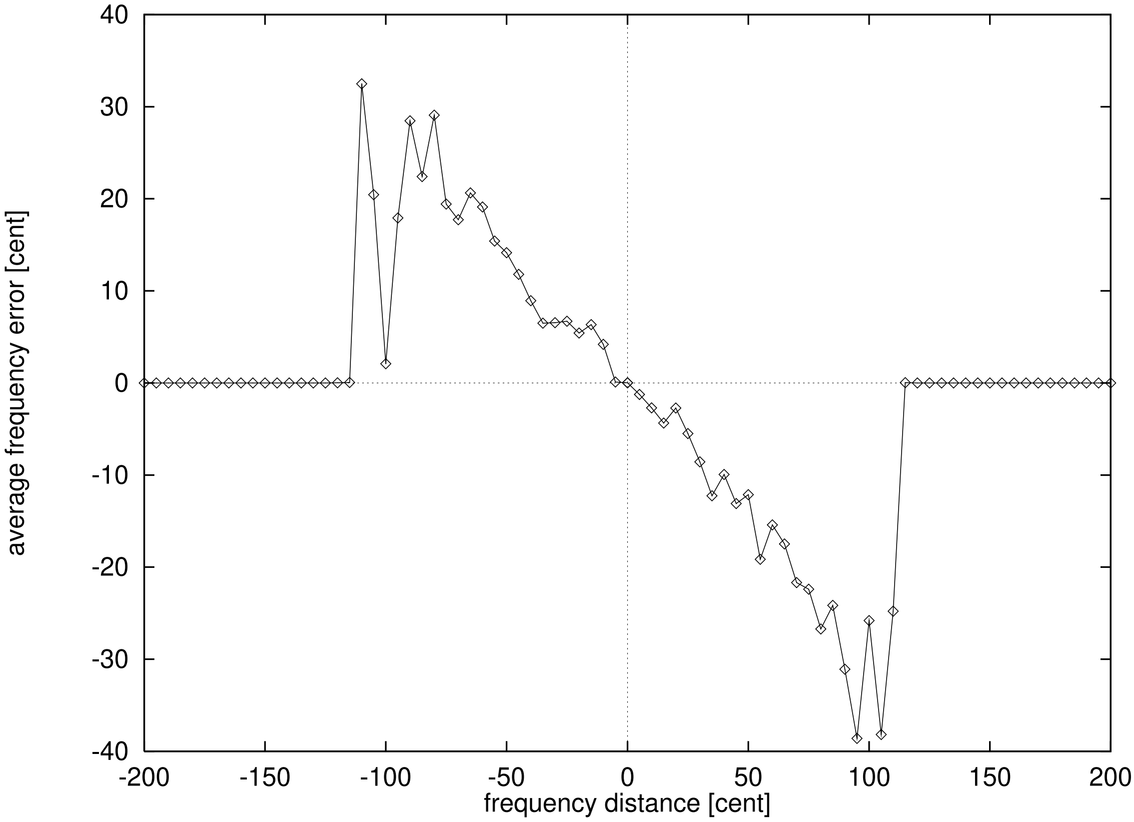

Figure 3.7: Frequency estimation error for superimposed sine at 1kHz and -10dB depending on partial distance. Parameters are $f_s = 44.1$kHz and $Q^{-1} = 0.05$.

get closer than that, the estimates are getting biased towards each other until they become indistinguishable at zero distance.

3.2.5 Nonstationary Performance

In this section the behavior of a PT in nonstationary environment is investigated. First, the settling time of a freshly installed PT in dependence on its initial frequency displacement is considered, then its ability of tracking two sweeps crossing each other.

3.2.5.1 Settling Time

The adaptation rule and the gammatone tracking filter dynamics are the determining factors for the settling time of a PT. For the unit step response of the gammatone filter we have for $n = 3$ with (2.65) and $\gamma(n, \lambda)$ set for amplitude normalization:

$$\epsilon(t) * g_{3,\lambda}(t) = 1 - e^{-\lambda t} \left(\frac{(\lambda t)^2}{2} + \lambda t + 1 \right). \tag{3.31}$$

The upper graph in Fig. 3.8 relates $\epsilon(t) * g_{3,\lambda}(t)$ to the development of amplitude estimates delivered by the PT as a response to a switched sine at -20 dB and 1 kHz. The parameter is the the initial displacement of the tracking filter's center frequency. For 0 Hz initial displacement the difference is negligible, for larger displacements it takes about 600 samples until the curves become indistinguishable. The lower panel in Fig. 3.8 shows the behavior of the frequency estimates in the same experiment. Note that at this fine scale a small oscillation becomes visible. A closer examination shows that the frequency of this oscillation is 1 kHz, i.e. the signal frequency. Its maximum relative amplitude is as low as 0.05%.

Attentive readers might have noticed that in spite of the PTs having a positive initial displacement, all trajectories start below the true sine frequency. This can be attributed to the fact that the buffer of the AR model estimator is empty initially, leading to a poor reliability of the first estimates. For the given parameters the first valid estimate does not appear before sample number 45, where the estimator window has just been filled. In order to account for this lack of reliability, it was decided to keep the center frequency fixed at the initial frequency until the estimator window is full. Once this is the case, the PT is released and center frequency adaptation as described in Section 3.2.3 is initiated. The initial frequency of each PT is determined before the stepback to the estimated onset sample is performed. This procedure will be explained in Section 3.3.7.

3.2.5.2 Partial Crossing

Figure 3.9 shows the results of experiments with two partials crossing each other in the time-frequency plane. One sweep is moving upwards from 300 Hz to 3500 Hz, the other one downwards from 3500 Hz to 500 Hz, each with a speed of 6 kHz/s. Both graphs look very similar at first sight. However, in the top graph the partials are successfully tracked beyond the intersection, while in the experiment shown below the PTs bounce back to where they came from. In fact, this *bouncing percept* is the one preferably reported by listeners, an effect attributed to the *frequency proximity principle* in human audition [Bregman, 1990]. The only difference between the two experiments shown is in the choice of the parameter κ in (3.26). This, however, does not imply that low values for kappa generally force the bouncing percept and higher ones the crossing percept. In fact, experiments with other values and the rapid oscillations apparent at the intersection suggest that it is difficult to find deterministic dependencies for this decision.

A feasible way to force deterministic behavior mentioned by Wang [1994] is making the adaptation rule take past center frequency changes into account. This measure would favor the crossing percept over the bouncing percept. However, as this decision seems to be somewhat arbitrary in the light of differing preferences of human listeners, it was not found worth the complications for stability analysis coming along with this change of tracking dynamics.

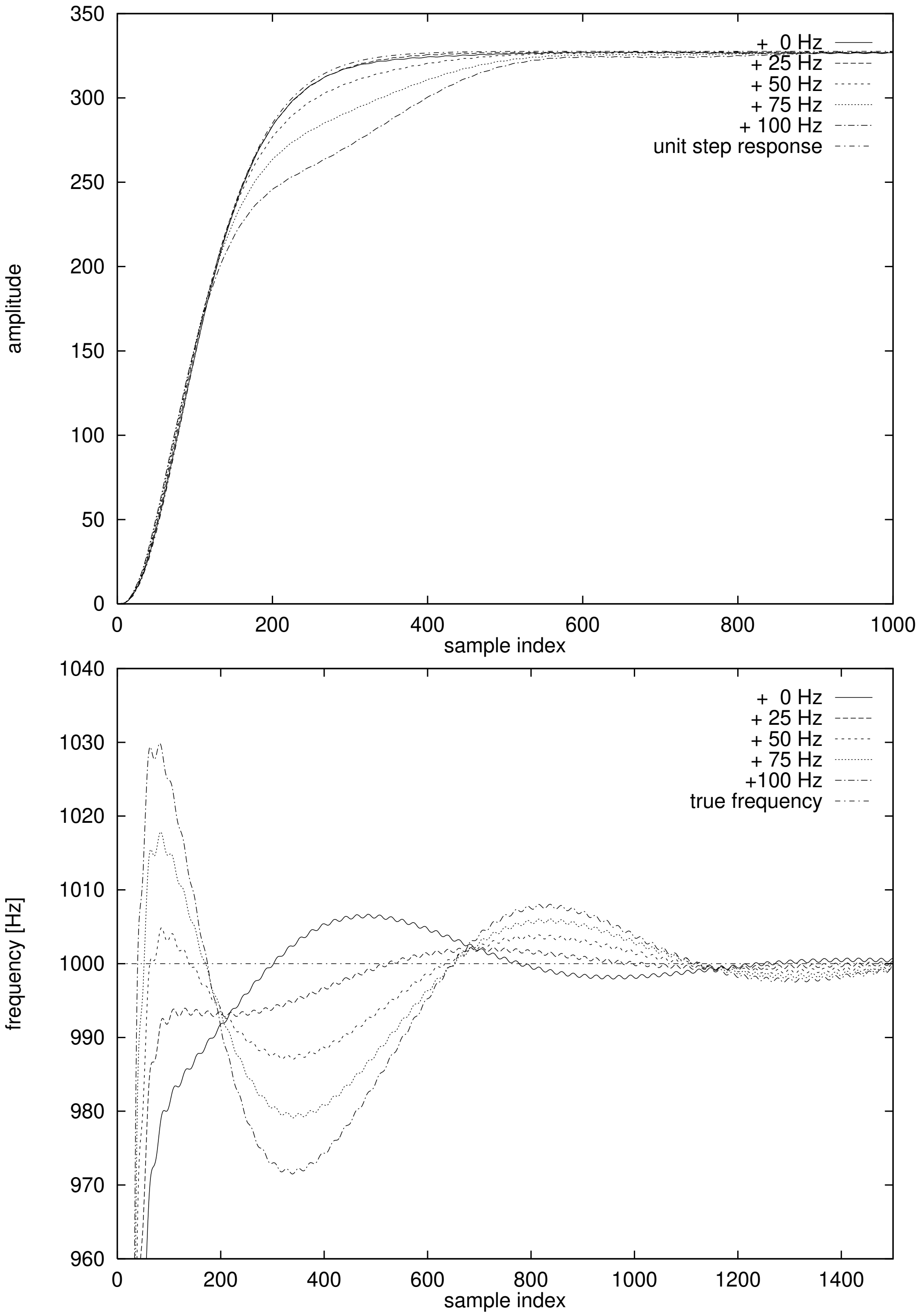

Figure 3.8: Amplitude response (top) and frequency response (bottom) for switched sine at various initial displacements of the tracking filter's center frequency. Parameters are $f_0 = 1$ kHz, $Q^{-1} = 0.1$, $f_s = 44.1$ kHz.

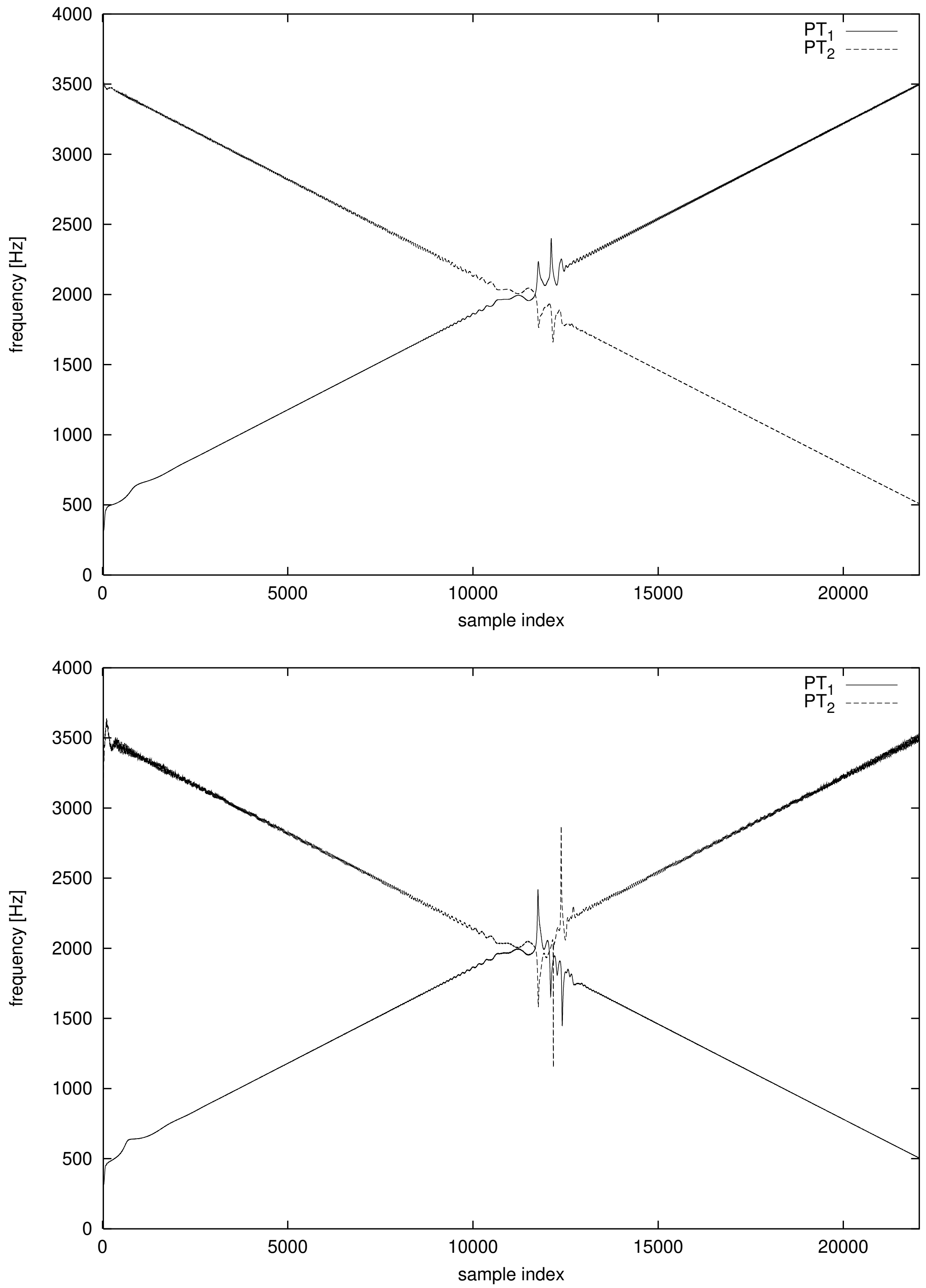

Figure 3.9: Partial crossing – crossing percept (top, $\kappa = 1.0$), bouncing percept (bottom, $\kappa = 0.25$). Parameters are $Q^{-1} = 0.1$, $f_s = 44.1$ kHz.

3.3 Master Module

The architecture of the master module is visualized in Fig. 3.10. At its entrance, a *residual signal* is formed by subtracting the estimated partial signals $\hat{s}_i(k)$ delivered by the TGs from the overall input signal $s(k)$. This residual is used as input to all

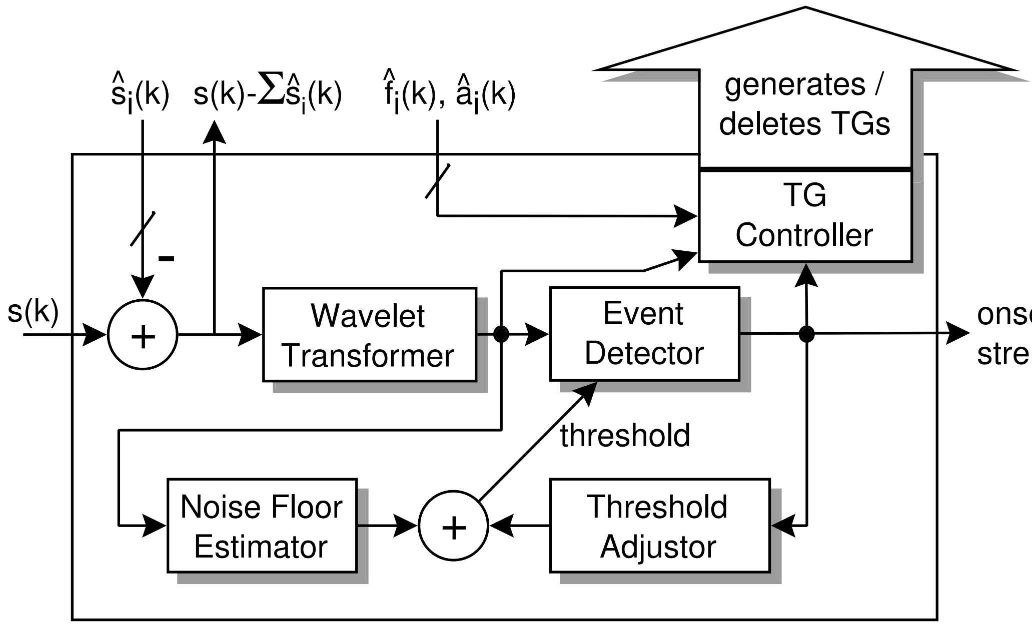

Figure 3.10: Master module.

other calculations, be it noise floor estimation, onset detection or partial tracking. The *wavelet transformer* is realized as a bank of quasi–analytic gammatone filters. Its purpose is providing suitable initial states for the partial tracking filters and pre-processing the signal for the *noise floor estimator* and the *event detector*. The event detector depends on the delivery of a threshold, which is calculated from the outputs of the noise floor estimator and the *threshold adjustor*. While the noise floor estimator contributes to the threshold by taking stochastic signal components into account, the threshold adjustor accounts for the finite velocity of the wavelet transform decay that follows if an onset has occurred. The *TG controller* is responsible for installing new groups of PTs and removing present PTs if a condition for partial death is met. These decisions are taken based on informations delivered by two different types of sources: first the event detector sending information about stochastic and transient components, second the PTs sending informations about the partials they are taking care of. Once an onset is detected, the partial locations in frequency are approximately identified by the TG controller and a tracker group with PTs at the proper frequency locations is installed after a stepback to the estimated onset sample. The details of the master module architecture concerning onset localization, PT initialization and PT death are described in the following subsections.

3.3.1 Forming a Broadband Resolution

From the result in Section 2.3.2 follows that it is advantageous for arrival time estimation to have a broadband target signal in a narrow observation time interval. It is for this reason that sound onset localization based on the interpretation of narrowband bandpass filter outputs (e.g. in [Baumann, 1995; Cooke, 1993; Moorer, 1975; Serra, 1989]) is inappropriate. Consequently, the architecture proposed in this thesis incorporates a broadband approach for partial onset detection instead. The optimum signal detector is the matched filter introduced in Section 2.3.2, requiring the target signal to be known precisely. Setting up a library of matched filters for all possible signals is intractable for obvious reasons, so the only viable way is giving up optimality for the sake of a higher degree of generality.

Signals satisfying the homogeneity Definition 2.12 cause a characteristic phase pattern in the time–scale plane. Isolated singularities can be detected through the integration of the wavelet transform along the curve of constant phase at which the impulse response amplitude is at its maximum, provided that the wavelet features a sufficient number of vanishing moments. The integration of wavelet coefficients along a line of constant phase in the time–scale plane can be regarded as the formation of a broadband signal from infinitely many delay–compensated narrow ones. In practice, this integration must be approximated by summing over discrete band locations in a limited scale range. The impulse response amplitude peak of the gammatone filter is at $t_p(n, \lambda) = \frac{n-1}{\lambda}$, which is $\frac{n-1}{n}$ times the group delay at the center frequency (see Sections 2.4.5 and 2.4.8). For the phase at this point of time we have

$$\begin{aligned} \psi_p(n) &= 2\pi f_0 t_p(n, \lambda) \\ &= \frac{Q \cdot (n-1)}{\sqrt{2n-3}} \end{aligned} \tag{3.32}$$

with (2.72) and the relative bandwidth being a constant of $Q^{-1} = \frac{\Delta f}{f_0}$. As expected, $\psi_p(n)$ is a constant for given choices of Q and n. Consider the operator

$$\mathcal{X}[s(t)] = \left| \sum_{i=0}^{N_b - 1} W_s\left(t + t_p\left(n,\, a_i^{-1}\, \lambda_{max}\right), a_i\right) \right| . \tag{3.33}$$

For a Dirac impulse located at t_0, every band will have the same phase and a maximum of its modulus at $t_0 + t_p(n, \lambda)$. If the i–th wavelet band with scale parameter a_i responds to $s(t) = \delta(t - t_0)$ with a maximum of its modulus at $t_0 + t_p\left(n,\, a_i^{-1}\, \lambda_{max}\right)$, then $\mathcal{X}[s(t)]$ will also have its maximum at t_0, as the complex coefficients $W_s(t + t_p(n, \lambda_i)$ sum up constructively along the phase line. An example for the Dirac impulse at $t_0 = 0$ is shown in Fig. 3.11.

Obviously, the unit step $\epsilon(t)$ is more appropriate as a model for partial onsets than the Dirac impulse. Consider a signal of the form

$$s(t) = \epsilon(t) \cdot e^{j 2\pi f_a t}. \tag{3.34}$$

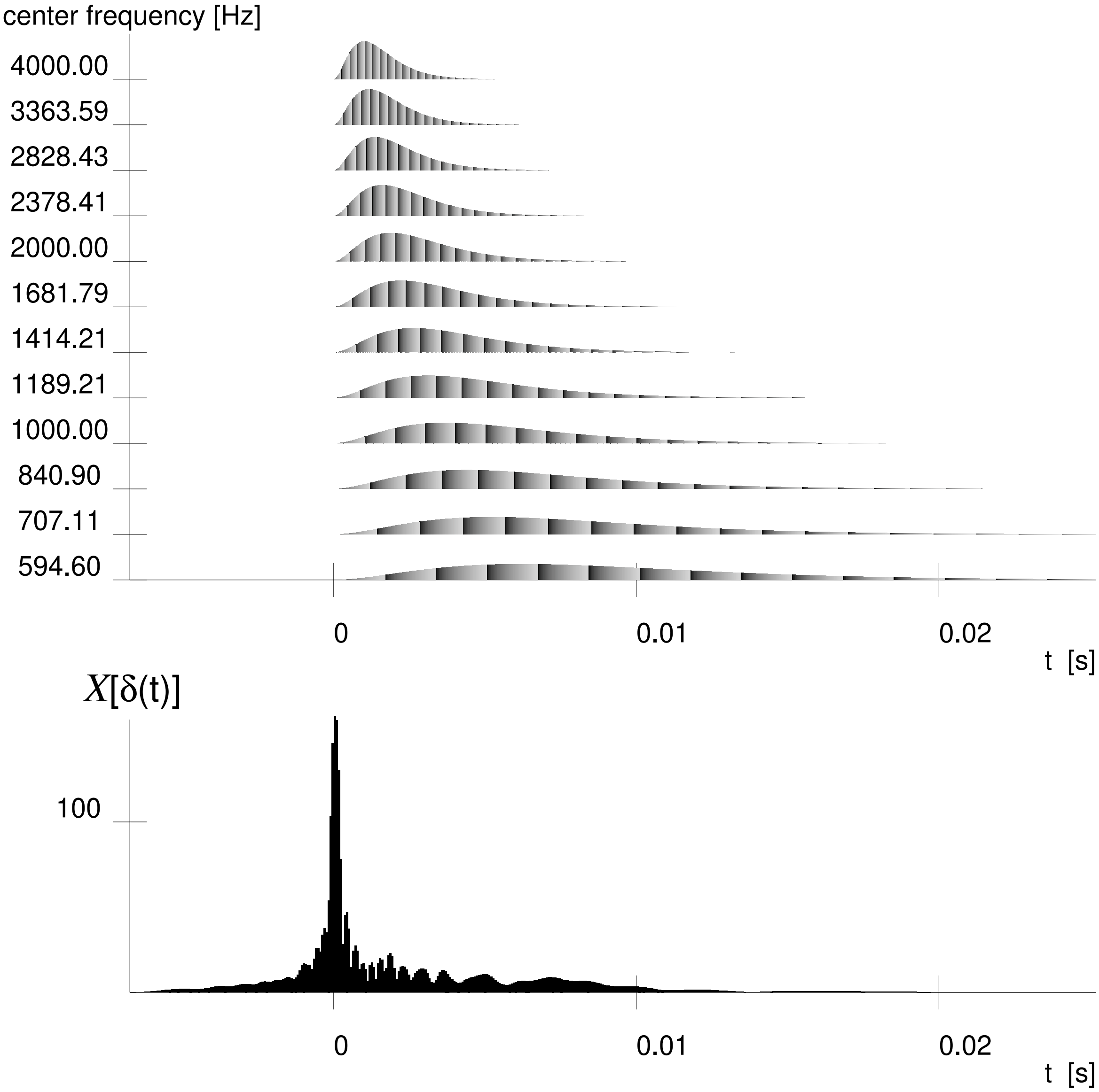

Figure 3.11: $\mathcal{X}[\delta(t)]$ for 3 octaves, 4 filters per octave.

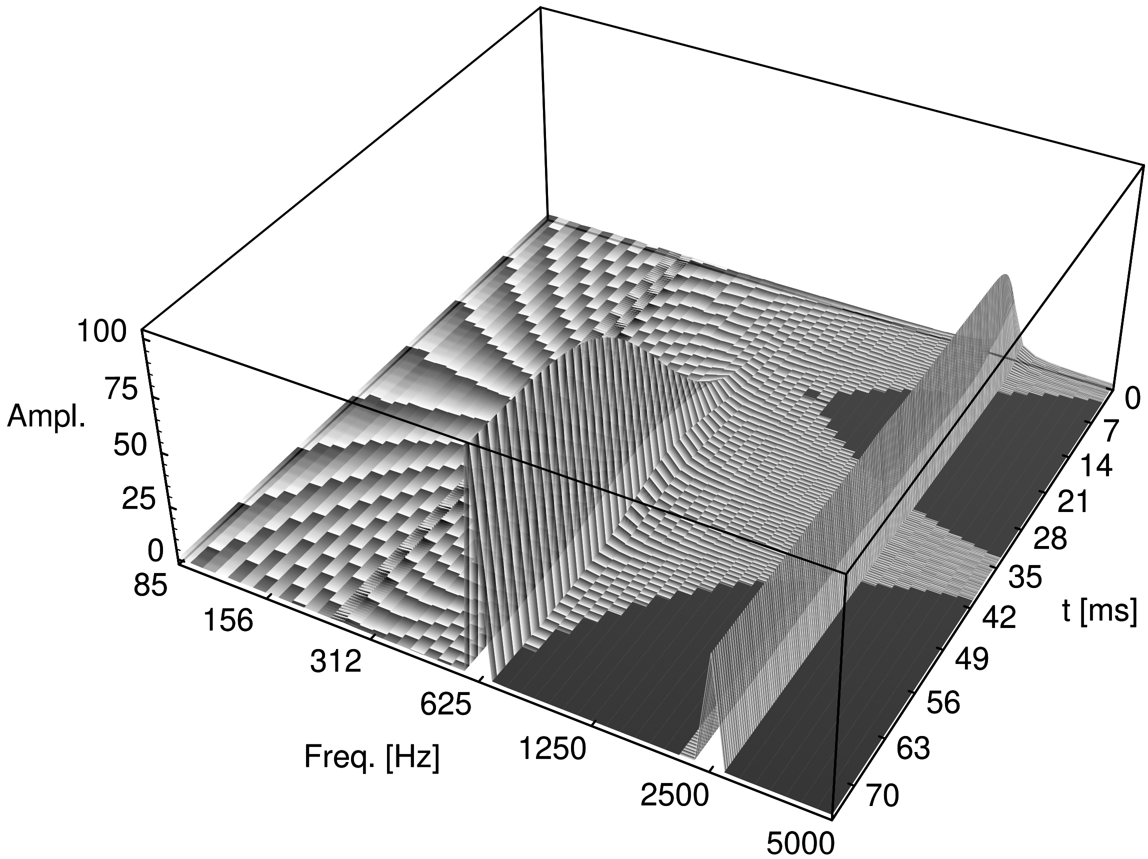

Figure 3.12: Two partials analyzed by an analytic Gaussian wavelet.

For the wavelet bands with $f_0 \gg f_a$ and those with $f_0 \ll f_a$, this signal is indistinguishable from $\epsilon(t)$ in terms of phase. This is clearly visible in Fig. 3.12 [Kliewer, 1993]: While the phase patterns of the two partials, one starting at 0 ms, the other at 37 ms, are distorted in the bands close to the signal frequencies 2500 Hz and 625 Hz, they remain intact in the remote bands.

The operator $\mathcal{X}$ is energy–normalized. For reasons that will become clearer in Section 3.3.2, it was found convenient to use amplitude normalization instead, so we arrive at the operator

$$\boxed{\mathcal{Y}[s(t)] = c_n \cdot \left| \sum_{i=0}^{N_b-1} \frac{1}{\sqrt{a_i}} \cdot W_s\left(t + t_p\left(n,\, a_i^{-1}\,\lambda_{max}\right), a_i\right) \right|}, \tag{3.35}$$

with c_n chosen such that

$$\forall f_a \in [f_{0_{min}}, f_{0_{max}}] \,:\, \lim_{t\to\infty} \mathcal{Y}\left[\epsilon(t) \cdot e^{j2\pi f_a t}\right] = 1. \tag{3.36}$$

The operator $\mathcal{Y}$ defined by (3.35) will serve as the basis of the onset detector described in what follows. One might ask, why the broadband signal for event detection is derived from a filter bank and not realized as a single filter with a broad passband. The reason is that the filter bank is also used for noise floor estimation and partial tracker initialization. Due to this multi–functionality, the high computational effort for the filter bank realization was considered worthwhile.

Even in the noiseless case, $\mathcal{Y}[s(t)]$ shows significant ringing clearly visible in Fig. 3.11 for $\mathcal{X}[\delta(t)]$. In the noisy case the situation becomes even more tedious, as further spurious peaks are added. In order to reduce the number of misses and false alarms, the following principles for peak selection are set into effect:

1. *Proper Choice of the Observation Interval*
   From the derivation of arrival time error variance for the matched filter given in Section 2.3.2, two conditions can be derived for the observation interval: first, it should be large enough to comprise the whole target signal, second, it must be short enough to guarantee a low error variance. The singularity of the unit jump has no extension in time. Due to the finiteness of realizable filter bandwidths, however, the effective signal duration is not infinitely short. For this reason, the lower bound of the observation interval will be derived from the overall bandwidth of the wavelet filter bank.

2. *Adaptive Resynthesis*
   The operator $\mathcal{Y}$ does not operate on the overall input signal $s(t)$, but on the residual signal, from which steady components are continuously removed through adaptive resynthesis. The adaptive signal cancellation be represented by the time–variant operator

   $$\mathcal{R}[s(t)] = s(t) - \sum_{i=1}^{N_{pt}(t)} \hat{s}_i(t), \tag{3.37}$$

   where $N_{pt}(t)$ is the number of PTs at time t and $s_i(t)$ is the signal delivered for residual calculation by the PT with index i.

3. *Threshold Adjustment*
   If a peak is identified as an onset, a time–varying threshold is computed taking finite tracking speed and ringing effects into account. Subsequent peaks are required to exceed this threshold in order to be considered as an onset candidate.

4. *Noise Floor Estimation*
   The noise floor is continuously estimated and incorporated into the threshold.

3.3.2 Threshold Adjustment

The easiest solution to cope with the ringing effects visible in Fig. 3.11 would be by introducing a fixed threshold and a recovery time, during which all peaks that follow an onset and lie above this threshold are ignored. However, these parameters would have to be chosen by trial and error in order to reduce misses and false alarms, so this approach is not satisfactory. In the following, a method for automated threshold adjustment is developed. The event parameters influencing the threshold adjustor are *onset time* and *onset strength*.

The assumption of infinitely fast tracking, i.e. zero group delay of the partial trackers, would be practical but is highly unrealistic. For the consideration of the finite tracking speed case it will be convenient to combine $\mathcal{Y}$ according to (3.35) and $\mathcal{R}$ according to (3.37) into a single operator

$$\mathcal{Z}[\bullet] = \mathcal{Y}[\mathcal{R}\,[\bullet]]\,. \tag{3.38}$$

For estimating an upper bound for $\mathcal{Z}[s(t)]$ on the right hand side of the detected peak we make the following approximative assumptions:

(a) Due to (3.36) the operator $\mathcal{Y}$ is equivalent to taking the modulus of the response of an ideal bandpass with bounding frequencies $f_{0_{min}}$, $f_{0_{max}}$ and bandwidth

$$\Delta f_b = \frac{1}{\tau_b} = f_{0_{max}} - f_{0_{min}}. \tag{3.39}$$

(b) The group delay d_{min} of the fastest partial tracker (equaling the group delay of the highest wavelet filter plus the one of the associated AR model estimator) satisfies

$$d_{min} > 2\tau_b. \tag{3.40}$$

Then, trivially, this condition also holds for all other partial trackers with group delay $d_i > d_{min}$. Using (3.39) and (2.80), (3.40) translates to $\frac{\Delta f_b}{\Delta f_{max}} > 7.26$ for a gammatone filter bank with order $n = 3$, with Δf_{max} denoting the bandwidth of the uppermost filter.

(c) The input is of the form $s(t) = q \cdot \epsilon(t) \cdot e^{j2\pi f_a t}$, with $f_{0_{min}} < f_a < f_{0_{max}}$.

(d) In $[0, 0 + d_i]$, the signal reaches the filter bank unaltered, while for $t > d_i$ it is completely canceled by the responsible partial tracker with group delay d_i.

Due to assumption (a), the lowpass equivalent of the filter bank has the impulse response

$$h_l(t) = \Delta f_b \cdot \mathrm{si}\,(\pi \Delta f_b t)\,, \tag{3.41}$$

with $\mathrm{si}(x) = \frac{\sin(x)}{x}$. As we are not interested in phases but amplitudes only, the class of signals defined by assumption (c) can be equivalently represented by the step function $q \cdot \epsilon(t)$. Fig. 3.13 shows the modulus of the ideal lowpass filter response to a unit step. With assumption (d), the resulting input function is the impulse response of a

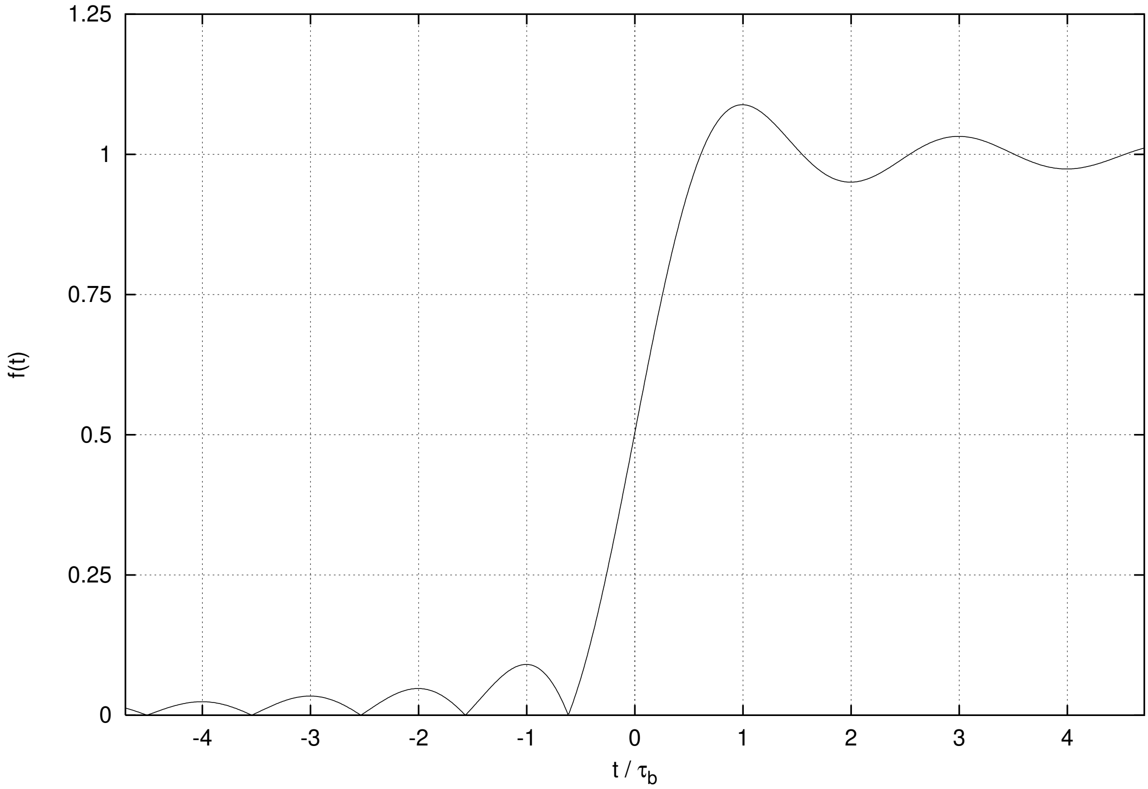

Figure 3.13: Modulus of the filter given by (3.41), responding to a unit step.

short–time integrator

$$s_l(t) = q \cdot (\epsilon(t) - \epsilon(t - d_i)), \tag{3.42}$$

so an upper bound $\Phi_p(t)$ for $\mathcal{Z}[s(t)]$ can be approximated as

$$\Phi_p[s(t)] = h_l(t) * s_l(t) = q \cdot \Delta f_b \cdot \int_{t-d_i}^{t} \mathrm{si}\,(\pi \Delta f_b t)\; dt. \tag{3.43}$$

The following properties of the *si*–function are important for our further considerations:

$$\max\left(\int_{-\infty}^{x} \mathrm{si}(\pi x)\, dx\right) = \int_{-\infty}^{1} \mathrm{si}(\pi x)\, dx \approx 1.0895, \tag{3.44}$$

$$\left|\int_{-\infty}^{x} \mathrm{si}(\pi x)\, dx\right| < 0.0895, \text{ for } x < -1, \tag{3.45}$$

$$\int_a^b \mathrm{si}(\pi x)\,dx < \int_a^b \frac{1}{\pi x}\,dx, \text{ for } 0 < a < b. \tag{3.46}$$

Due to (3.44), the ripple of the integrated *si*–function is invariant with respect to dilations of the integrand, a behavior commonly known as *Gibbs' phenomenon* [Papoulis, 1962; Fliege, 1991; Lüke, 1985]. Thus, for $t \approx \frac{1}{\Delta f_b} = \tau_b$ in (3.43) we find the peak of $\mathcal{Z}[s(t)]$ as being approximately equal to the input signal amplitude q, since with (3.43):

$$\begin{aligned}
\Phi_p[s(t)]|_{t=\tau_b} &= q \cdot \Delta f_b \cdot \int_{\tau_b - d_i}^{\tau_b} \mathrm{si}\left(\pi \Delta f_b t\right)\,dt \\
&= q \cdot \Delta f_b \cdot \left[\int_{-\infty}^{\tau_b} \mathrm{si}\left(\pi \Delta f_b t\right)\,dt - \int_{-\infty}^{\tau_b - d_i} \mathrm{si}\left(\pi \Delta f_b t\right)\,dt\right] \\
&\approx q \cdot \int_{-\infty}^{1} \mathrm{si}\left(\pi x\right)\,dx, \qquad \text{with (3.45) and assumption (b)} \\
&\approx q, \qquad \text{with (3.44).}
\end{aligned} \tag{3.47}$$

With the value of q found at the peak we arrive at the following upper bound for $t > t_x$, with a time $t_x > d_i$ yet to be determined:

$$\begin{aligned}
\Phi_p[s(t)]|_{t>t_x} &= q \cdot \Delta f_b \cdot \int_{t-d_i}^{t} \mathrm{si}\left(\pi \Delta f_b t\right)\,dt \\
&< q \cdot \int_{\Delta f_b \cdot (t-d_i)}^{\Delta f_b \cdot t} \frac{1}{\pi t}\,dt \qquad \text{with (3.46)} \\
&= -\frac{q}{\pi} \cdot \log\left(1 - \frac{d_i}{t}\right).
\end{aligned} \tag{3.48}$$

Finally, the upper bound for $\mathcal{Z}[s(t)]$ is

$$\boxed{\Phi_p[s(t)] = \begin{cases} q, & \text{for } \tau_b < t \le t_x, \\ -\frac{q}{\pi} \cdot \log\left(1 - \frac{d_i}{t}\right), & \text{for } t > t_x \end{cases}}, \tag{3.49}$$

with t_x such that $\Phi_p[s(t)]$ is continuous at t_x:

$$\begin{aligned}
q &= -\frac{q}{\pi} \cdot \log\left(1 - \frac{d_i}{t_x}\right) \\
\Longrightarrow t_x &= \frac{d_i}{1 - e^{-\pi}}.
\end{aligned} \tag{3.50}$$

Note that the overall bandwidth Δf_b does not appear in the final result (3.49), because assumption (b) ensures the domination of partial tracker group delays over the time constant of the main filter bank. Due to the presence of the tracker group delay d_i, the threshold depends on the frequency location of the partial originating at the onset.

This knowledge, together with initial partial amplitudes, is delivered by the partial initialization procedure to be described in Section 3.3.7.

Fig. 3.14 shows some examples for thresholds calculated by the method presented above. The true sine amplitude is 3277 for each of the three different partials, which is reflected in the asymptotic value in the lower right panel showing the output of $\mathcal{Y}$ for a sine of 500Hz with adaptive feedback cancellation disabled. The panel above shows the output of $\mathcal{Y}$ for the same signal with a tracker installed at the onset sample and adaptive feedback cancellation enabled. Obviously, the signal is gradually discarded. The panels with adaptive feedback cancellation enabled indicate that the threshold computed by the method presented in this section is appropriate for reducing false onset alarms while still maintaining a tight approximation to the true curve in order to reduce the probability of missing subsequent onsets.

So far only the case of the onset of a single partial has been discussed. The question arises, what to do in the case of multiple partials or no partial at all detected by the partial initialization procedure. If n_{pt} partials are detected at a particular onset, the peak value q is shared among them proportionally to their initial amplitudes. If no partial has been found, the minimum group delay of the bands in the wavelet filter bank, i.e. the group delay of the uppermost wavelet filter, is used in (3.49). As the system behavior in these cases can be demonstrated more easily after the introduction of the partial initialization procedure, the case of a signal consisting of a combination of all signals appearing separately in Fig. 3.14 is delayed to Section 4.1.

3.3.3 What is Noise?

In common language use, the word *noise* is far from being defined unambiguously. Of all the definitions listed in Fig. 3.15, the entries 2b–2e are the ones getting nearest to what could be used in a technical formulation. But still, words like *undesired*, *unwanted*, *disturbing*, *irrelevant*, *meaningless* leave us with the open question of how to find a definition that does not essentially rely on individual preferences. The word *random* appearing in Definition 2d opens a road in this direction. In the extreme case of randomness, each instantaneous value of a signal is everywhere independent of all other values. This, however, is a property difficult to verify. Instead, we are satisfied with the requirement of different samples to be *uncorrelated.* In this case and assuming wide–sense stationarity[1] we have for the autocorrelation function

$$\phi_{xx}(\tau) = E\{x(t+\tau)x(t)\} = \delta(\tau), \tag{3.51}$$

translating to the demand of a stationary flat (i.e. white) spectral density in the frequency domain. Thus, noise due to this definition is a phenomenon infinitely extended in *both* time and frequency. The third condition we assume is a Gaussian probability density. This assumption is justified by the central limit theorem [Papoulis, 1990], stating that the probability density of a process resulting from the additive superposition

[1] i.e. the autocorrelation only depends on the lag parameter τ.

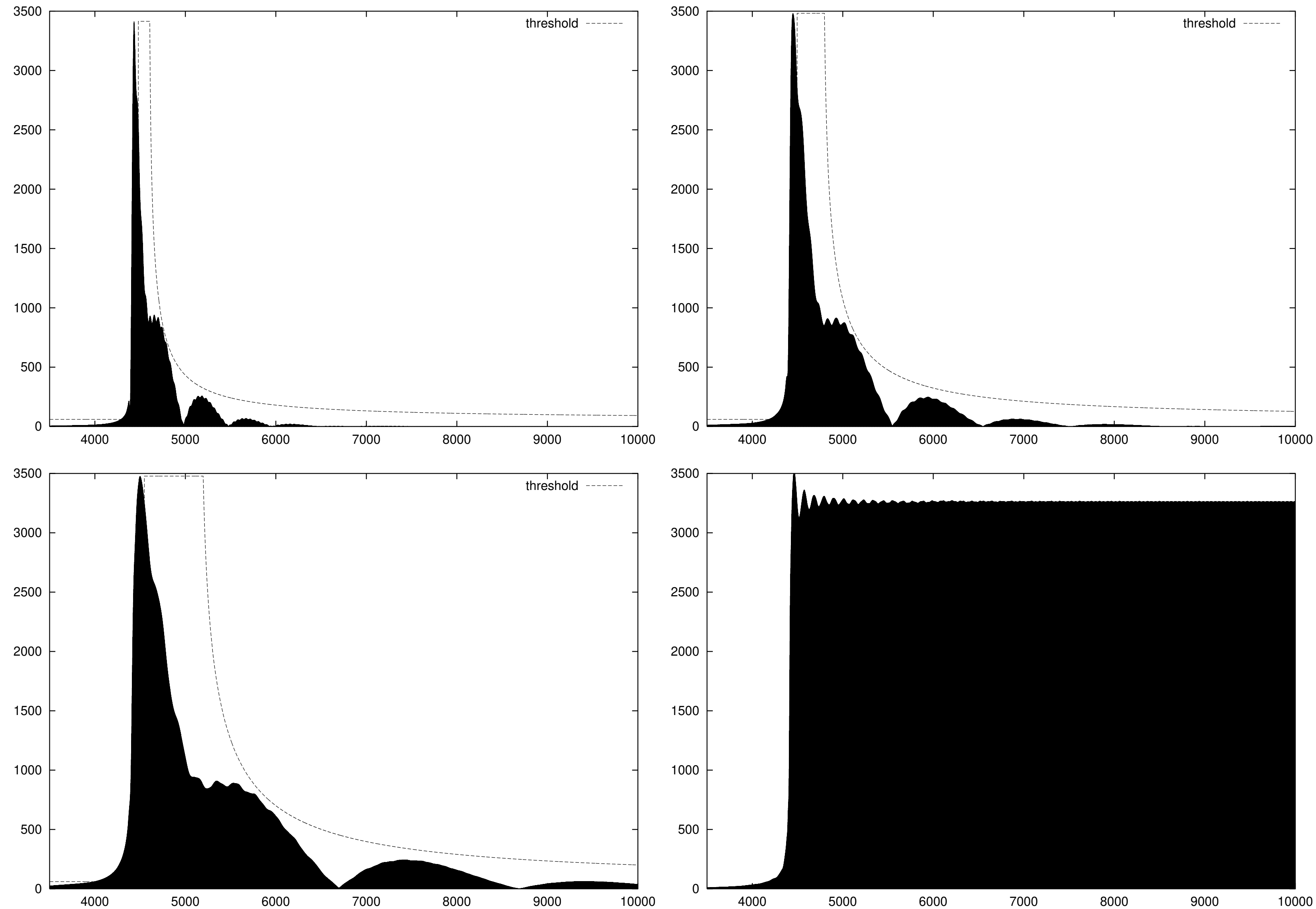

Figure 3.14: Onset thresholds for partials of $f_0 = 1\text{kHz}$ (upper left), $f_1 = 500\text{Hz}$ (upper right), $f_2 = 250\text{Hz}$ (lower left), all with the same amplitude $a_0 = 3277$ and onset sample 4410. The lower right panel shows $\mathcal{Y}[a_0 \cdot \sin(2\pi\, 500\text{Hz}\, k\, T_s)]$ for the case of adaptive feedback cancellation switched off. Analysis parameters are $f_{0_{min}} = 49$ Hz, $f_{0_{max}} = 1480$ Hz, $\frac{\Delta f}{f_0} = 0.1$.

[1]**noise** \ˈ noˑiz\ *n*
[ME, fr. OF, strife, quarrel, noise, fr. L *nausea* nausea]
1: loud, confused, or senseless shouting or outcry
2a: SOUND; *esp***:** one that lacks agreeable musical quality or is noticeably unpleasant
b: any sound that is undesired or interferes with one's hearing of something
c: an unwanted signal or a disturbance (as static or a variation of voltage) in an electronic communication system (as radio or television); *broadly***:** a disturbance interfering with the operation of a usu. mechanical device or system
d: electromagnetic radiation (as light or radio waves) that is composed of several frequencies and that involves random changes in frequency or amplitude
e: irrelevant or meaningless bits or words occurring along with desired information (as in a computer output)
3: common talk**:** RUMOR; *esp***:** SLANDER
4: something that attracts attention ⟨the play...will make little *noise* in the world —Brendan Gill⟩
5: something spoken or uttered

Figure 3.15: Meanings of the word *noise* quoted from *Webster's Ninth New Collegiate Dictionary.*

of m random processes with similar variances approaches the Gaussian distribution as m grows to infinity, even if the variables themselves are not Gaussian.

While the definition of noise as a *wide–sense stationary Gaussian white process* is satisfactory in terms of precision, it is much more restrictive than any of the definitions in Fig. 3.15. On the other hand, in applications where *noise* comprises a larger class of signals, the class of non–noise signals (i.e. the target class) must necessarily be restricted. An example for such applications is voice activity detection (VAD) in wireless personal communication systems [El-Maleh and Kabal, 1997]. Here, the knowledge of some characteristics of the wanted signal component (i.e. speech) is used in order to make *noise* definable as *everything that is not speech.* In this sense a trumpet melody would have to be considered as noise. In the context of the work presented here, we take the approach of using a very narrow definition, in order not to restrict the target class right from the start.

3.3.4 Noise Floor Estimation

In the presence of noise, $\mathcal{Y}[s(t)]$ exhibits further spurious peaks that we wish to eliminate. In the following, an algorithm for noise floor estimation is presented. The noise floor is added to the event detection threshold (3.49), both together forming the total threshold employed for onset detection.

Due to the proof of Corollary 2.3, we have for the modulus along lines of constant phase in the wavelet transform of homogeneous signals

$$|w_i| \sim a_i^{\mu+\frac{1}{2}}, \tag{3.52}$$

where w_i is the value of the wavelet transform along the phase line at band index i, a_i is the scale parameter and $\mu \in \mathbb{R}$ the degree of homogeneity. Unfortunately, the least squares fit of the parameter μ to the measured data leads to a nonlinear optimization problem [Friedman and Kandel, 1994] and the linearization via logarithmic transformation yields a solution which is highly dominated by the most unreliable small values. As the estimation of μ is difficult, we aim at finding a method for detecting noise–only time intervals without having to determine this parameter.

In the case of white noise of variance σ_n^2, we have due to the Wiener–Lee theorem and the energy normalization of the wavelet filters

$$E\left\{|w_i|^2\right\} = \sigma_n^2. \tag{3.53}$$

From this follows that for white noise, we have $\mu = -0.5$ in (3.52) on average.

If the noise is not only white but also Gaussian and if the real and imaginary filter responses are orthogonal[2] the modulus exhibits a Rayleigh distribution (see Appendix C), so the expectation value of $|w_i|$ is

$$E\{|w_i|\} = \frac{\sqrt{\pi}}{2} \cdot \sigma_n. \tag{3.54}$$

Thus, for Gaussian white noise the expectation value of the modulus is a constant given by (3.54). We make use of this fact for noise floor estimation by fitting a straight line of the form

$$g(i) = a + b \cdot i, \quad i \in \{0..N_b - 1\}, \tag{3.55}$$

where N_b is the number of wavelet bands and i the band index running from the highest to the lowest band, to the wavelet transform modulus along the selected phase line via least squares linear regression. For the approximation error we have

$$e(i) = g(i) - |w_i| = a + b \cdot i - |w_i|. \tag{3.56}$$

For least squares approximation we need to solve

$$\|e(i)\|^2 = \sum_{i=0}^{N_b-1} e^2(i) \rightarrow \min. \tag{3.57}$$

From $\frac{\partial\|e(i)\|^2}{\partial a} = 0$ and $\frac{\partial\|e(i)\|^2}{\partial b} = 0$, we obtain the matrix equation

$$\begin{pmatrix} \sum_{i=0}^{N_b-1} 1 & \sum_{i=0}^{N_b-1} i \\ \sum_{i=0}^{N_b-1} i & \sum_{i=0}^{N_b-1} i^2 \end{pmatrix} \cdot \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{N_b-1} |w_i| \\ \sum_{i=0}^{N_b-1} i \cdot |w_i| \end{pmatrix}. \tag{3.58}$$

[2]This condition holds for analytic filters.

With

$$m_1 = \frac{1}{N_b} \sum_{i=0}^{N_b-1} |w_i| \tag{3.59}$$

and

$$m_2 = \frac{1}{N_b} \sum_{i=0}^{N_b-1} i \cdot |w_i| \tag{3.60}$$

we have

$$\begin{pmatrix} 1 & \frac{(N_b-1)}{2} \\ \frac{(N_b-1)}{2} & \frac{(2N_b^2-3N_b+1)}{6} \end{pmatrix} \cdot \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}. \tag{3.61}$$

The matrix in (3.61) is the *Fisher information matrix* (see Appendix E) for line fitting in Gaussian white noise of unit variance [Kay, 1993]. Matrix inversion yields

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \frac{2}{N_b+1} \begin{pmatrix} 2N_b - 1 & -3 \\ -3 & \frac{6}{N_b-1} \end{pmatrix} \cdot \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}. \tag{3.62}$$

As the matrix in (3.61) is the Fisher information matrix, the diagonal elements of its inverse in (3.62) would be the *Cramér–Rao bounds* for the error variances $\sigma_{\hat{a}}^2$ and $\sigma_{\hat{b}}^2$. However, as our wavelet filters are not mutually orthogonal, the estimates obtained by least squares estimation cannot be expected to be optimal in the maximum likelihood sense.

As the ideal case of $\hat{b} = 0$ will hardly ever be observed, a tolerance must be granted. For the variance of the modulus σ_w^2 we have with (3.53), (3.54) and $m_1 \approx E\{|w_i|\}$:

$$\begin{aligned} \sigma_w^2 &= E\left\{(|w_i| - E\{|w_i|\})^2\right\} \\ &= E\{|w_i|^2\} - E^2\{|w_i|\} \\ &= \sigma_n^2 \left(1 - \frac{\pi}{4}\right) \\ &\approx m_1^2 \cdot \left(\frac{4}{\pi} - 1\right). \end{aligned} \tag{3.63}$$

The acceptable tolerance for $\hat{b}$ is set proportional to σ_w, i.e. if

$$|b| \leq \eta \cdot m_1 \cdot \sqrt{\left(\frac{4}{\pi} - 1\right)}, \tag{3.64}$$

with some constant $\eta > 0$, we consider the signal as Gaussian white noise.

In case this condition holds, a new value for the noise–induced component of the onset detection threshold is set. With c_n and a_i as in (3.35) for the calculation of $\mathcal{Y}$, we set

$$m(kT_s) = c_n \cdot \frac{1}{N_b} \sum_{i=0}^{N_b-1} \frac{1}{\sqrt{a_i}} \cdot |w_i(kT_s)| \tag{3.65}$$

We see that $m(kT_s)$ is calculated from the very same coefficients as $\mathcal{Y}[s(kT_s)]$. However, while $\mathcal{Y}[s(kT_s)]$ is the modulus of the sum of these coefficients, $m(kT_s)$ is the sum of the moduli. In order to take finite word length effects into account, we set a lower bound of 1 to the noise floor, so we finally arrive at

$$\Phi_n(kT_s) = \max(m(kT_s),\, 1). \tag{3.66}$$

If (3.64) does not hold, Φ_n is kept unaltered, i.e. $\Phi_n(kT_s) = \Phi_n((k-1)T_s)$. The initial value is $\Phi_n(0) = 1$.

A problem with the criterion (3.64) arises from the fact that $\hat{b}$ would be close to zero not only for white noise but for all spectra being symmetric with respect to the midmost frequency band. This problem can be solved by calculating $N_{lf} > 1$ line fits for maximally shifted band index origins and requiring the fulfillment of (3.64) for each of them. Another problem is the possibility of a flat modulus occurring in the neighborhood of an onset. If the noise floor was derived from such a line, it would be erroneously high. In order to prevent the noise floor to be calculated from such a location, we additionally require that

$$\sum_j \Phi_{p_j}[s(kT_s)] \leq \frac{\Phi_n(kT_s)}{2}, \tag{3.67}$$

i.e. the noise floor may not be updated before the threshold adjustment resulting from previous onsets given by (3.49) has fallen below 50% of the last valid noise floor estimate.

3.3.5 Onset Detection Algorithm

For the sake of convenience we combine the onset–induced threshold $\sum_j \Phi_{p_j}[s(kT_s)]$ and the noise–induced threshold $\Phi_n(kT_s)$ according to (3.66) into the total threshold

$$\Phi_t(kT_s) = \Phi_n(kT_s) + \sum_j \Phi_{p_j}[s(kT_s)]. \tag{3.68}$$

The second term in (3.68) accounts for the ripple at the right hand side of an onset but not for the ripple on the left (see Fig. 3.11). This problem is solved by introducing a time constant τ_m denoting the distance the algorithm awaits before a peak is identified as an onset. If there appears a higher value before this time has run out, the previous peak is no longer considered as an onset candidate. As the maximum distance between two adjacent peaks is $2 \cdot \tau_b$ (see Fig. 3.13), τ_m should be chosen such that

$$\tau_m \geq 2 \cdot \tau_b = \frac{2}{f_{0_{max}} - f_{0_{min}}}. \tag{3.69}$$

If there is a lower peak within $[k_0 T_s - \tau_m, k_0 T_s]$, it is considered as caused by the ripple phenomenon. Hence, in case this peak should have been caused by a true onset, it is lost. The system constant τ_m can be regarded as defining the observation interval used for onset detection. As was shown for the matched filter in Section 2.3.2, a short observation interval is advantageous, the lower limit given by the target signal duration. However, as the finite bandwidth of $\mathcal{Y}$ overrules the infinite bandwidth of the target signal, it is the determining factor for the minimum length of the observation interval.

With $\Phi_n(0) = 0$ as initial value, the complete algorithm for event detection at sample index k is as follows:

1. Calculate $\sum_j \Phi_{p_j}[s(kT_s)]$, the threshold adjustment resulting from previous onsets, by (3.49).
2. Estimate b in (3.55) by (3.62) for N_{lf} different, maximally shifted band index origins.
3. If (3.64) holds for all N_{lf} origin shifts and (3.67) also holds, set $\Phi_n((k+1)T_s)$ according to (3.66).
4. If (3.64) does *not* hold for every index origin shift, the signal does not consist of white noise only. Then
   (a) $\Phi_n(kT_s) := \Phi_n((k-1)T_s)$.
   (b) Sample index $k_0 := k - 1$ is an onset candidate, if
      - $\mathcal{Z}[s(k_0 T_s)] > \Phi_t(k_0 T_s)$ and
      - $\mathcal{Z}[s(k_0 T_s)]$ is a local maximum.
5. If $\mathcal{Z}[s(kT_s)] > \mathcal{Z}[s(k_0 T_s)]$, cancel k_0 as an onset candidate.
6. If $k - k_0 > f_s \cdot \tau_m$, k_0 is confirmed as an onset. Then
   - Step back to k_0.
   - $q_{j+1} := \mathcal{Z}[s(k_0 T_s)] - \Phi_t(k_0 T_s)$ is used as initial value for $\Phi_{p_{j+1}}[s(kT_s)]$, henceforth developing according to (3.49).
7. $k := k + 1$, repeat.

Figure 3.16 shows an example for sinusoid onset detection at two different signal–to–noise rates. The analysis parameters are $f_s = 22.05$ kHz, $n = 3$, 4 octaves, $N_v = 12$,

$N_{lf} = 4$, $f_{0_{min}} = 98$ Hz, $Q^{-1} = 0.1$, $\tau_m = \tau_b$, $\eta = 2.0$. The true onset is at sample index 2205 in both cases. For 20 dB SNR the mean onset index estimate is at 2214.8, for 10 dB at 2217.0. At the given sample rate this translates to a bias of less than 0.6 μs. The standard deviation from the estimated mean onset index is 6.8 samples for 20 dB SNR and 73.97 samples for 10 dB. The number of outliers increases quickly towards an SNR below 10 dB.

3.3.6 Temporal and Spectral Masking

The introduction of a time constant τ_m into the onset detection algorithm described in Section 3.3.5 leads to a phenomenon similar to one that can also be observed in psychoacoustic experiments, where a strong onset covers a previous smaller one due to so–called *pre–masking*. Among other types of masking, this effect is exploited for bit rate reduction in lossy audio coding algorithms such as MPEG, DTS or AC-3. Figure 3.17, which was reproduced from [Noll, 1997], shows a rough sketch of masking effects playing a role in the human auditory system [Zwicker and Fastl, 1990].
The second temporal masking effect is that of *post–masking*. Post–masking also plays a role in the proposed architecture due to the time–varying threshold on the right hand side of an onset (see Section 3.3.2, most notably Fig. 3.14). It must be stressed, however, that the purpose of introducing τ_m and the time–varying threshold is detection robustness and not modeling a psychophysical phenomenon, but the coincidence is certainly worth noting. The third type of masking in Fig. 3.17 is termed *simultaneous masking*. As the underlying mechanism is that of proximity in frequency, we prefer to call this effect *spectral masking*. This type of masking had to be introduced into the architecture for tracking stability reasons. It will play an important role in the following sections about partial tracker initialization and death.

3.3.7 Partial Tracker Initialization

In [Solbach and Wöhrmann, 1996] the partial tracking method based on adaptive gammatone filtering and first order AR model estimation was dependent on a–priori knowledge about the initial frequency locations. In the following, a method for partial localization removing this dependency is presented.

Let i denote the wavelet band index starting from low scales (i.e. high frequencies). We assume partial parameter estimation being performed in the i–th band with center frequency $f_0(i)$. If the partial frequency estimates $\hat{f}(i)$ satisfy $\hat{f}(i) < f_0(i)$, the i–th band resides above the partial, if we have $\hat{f}(i) > f_0(i)$ it lies below. Thus, in order to localize partial candidates, we may look for the positive zero crossings of the *detuning function*

$$\boxed{\chi(i) = \hat{f}(i) - f_0(i)}\,. \tag{3.70}$$

The method used for frequency estimation in the i–th band is identical to the one described in Section 3.2.1, with a time window size $\Delta t(i)$ according to (2.69). In order

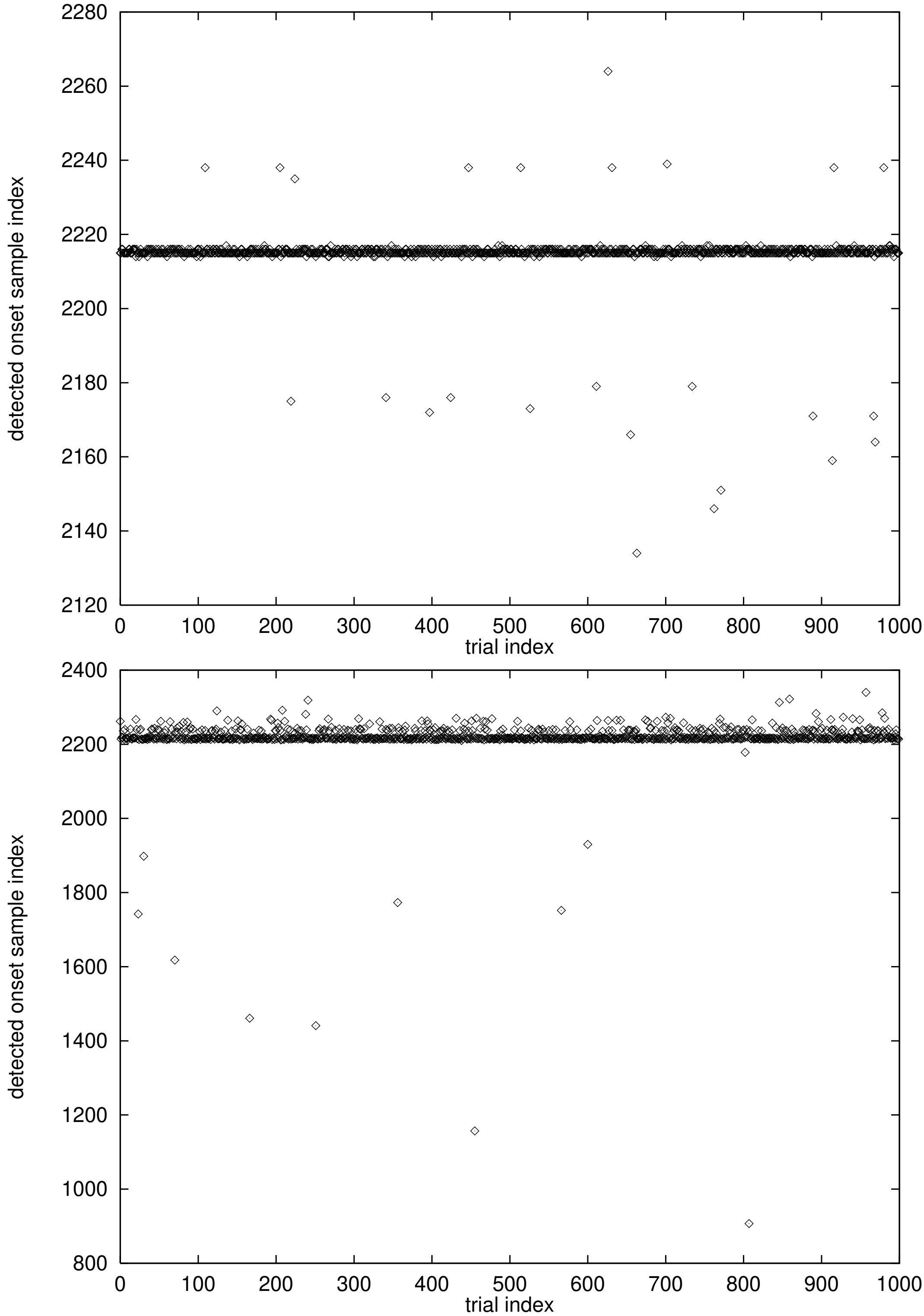

Figure 3.16: Onset detection at 20 dB SNR (top) 10 dB SNR (bottom), true onset at sample index 2205.

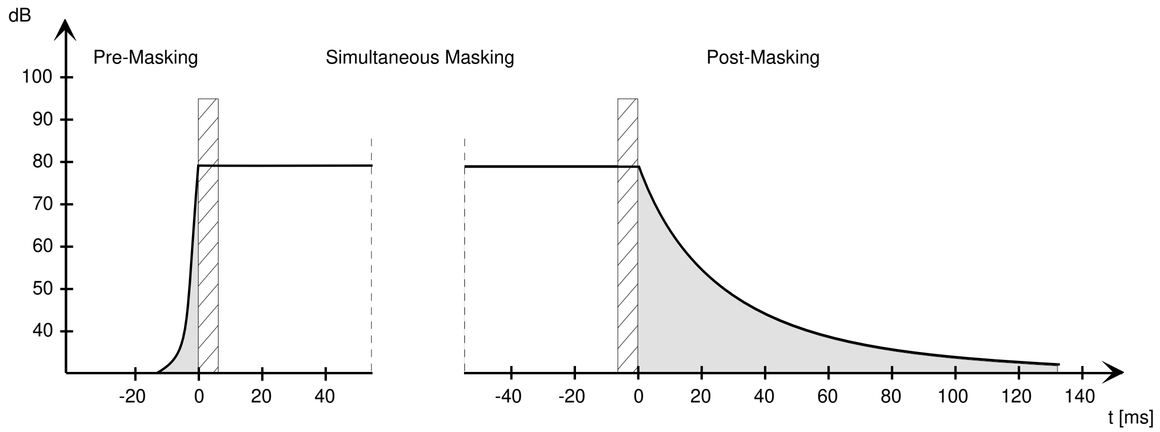

Figure 3.17: Masking effects in the human auditory system.

to account for the differing estimator window sizes across the wavelet filter bank, $\chi_i(k)$ is evaluated along a curve defined by

$$\hat{t}_0 + t_p(i), \tag{3.71}$$

where $\hat{t}_0$ is the estimated onset time and $t_p(i)$ is the impulse response amplitude peak location given by (2.66). With (2.69) we find that

$$t_p(i) = \frac{2 \cdot (n-1)}{\sqrt{2n-1}} \cdot \Delta t(i). \tag{3.72}$$

From this follows that $t_p(i) \approx 1.79 \cdot \Delta t(i)$ for $n = 3$, i.e. the AR estimator windows are indeed filled with samples past the estimated onset location. With (2.73) and $Q^{-1} = \frac{\Delta f}{f_0}$ we get

$$f_0(i) \cdot t_p(i) = \frac{(n-1)}{2\pi \cdot \sqrt{2n-3} \cdot Q^{-1}}. \tag{3.73}$$

For $n = 3$ and $Q^{-1} = 0.05$, the i–th filter has passed 3.676 cycles at its center frequency when arriving at $\hat{t}_0 + t_p(i)$. This was found to yield a fast but nevertheless sufficiently reliable estimate for approximate partial localization.

The detuning function (3.70) was also used by Cooke [1993] as a basis of a permanent partial localization process. In his system, $\chi(i)$ is evaluated along vertical lines in the time–frequency plane and the frequency estimator is realized in a different manner. More importantly though, there is no partial feedback cancellation performed, so $\chi(i)$ is always calculated from the output of a filter bank operating on the unaltered input signal. By contrast, in the architecture proposed in this thesis partial localization via $\chi(i)$ is performed immediately after onset detection only. Moreover, it is not calculated for the overall system input but for the residual signal, which is approximately devoid of older partials.

However, a severe problem still appears in the presence of noise. In this case, $\chi(i)$ exhibits many spurious zero crossings even without the actual presence of a partial. It should, however, not be surprising that making a partial localization procedure rely on frequency estimates only, while completely ignoring the amplitudes, causes susceptibility to noise. A suitable amplitude criterion can be derived from the assumption of the noise being Gaussian and white. Then, in noise–only periods the amplitude exhibits a Rayleigh distribution (see Appendix C). Setting a certain threshold a_t, leads to a false alarm probability of $P(a > a_t)$ due to (C.2), that a exceeds a_t although there was in fact nothing but Gaussian white noise present. Inversely, we can choose a certain false alarm probability $P(a > a_t)$ to calculate the related threshold a_t. For this, however, we would need knowledge about the noise power. Clearly, in the situation at hand, the wavelet bands we should *not* take into consideration for noise power estimation are the ones that passed the zero crossing test described above, because these are partial candidates. What we do instead is determine the average $\bar{a}_\chi$ of the bands that were no zero crossing candidates and use this value as a coarse estimate for the expectation value of the noise amplitude[3]. Thus, with (C.2) and (C.4) we get for the threshold:

$$\boxed{a_t = \frac{2 \cdot \bar{a}_\chi}{\sqrt{\pi}} \cdot \sqrt{-\log P(a > a_t)}}\,. \tag{3.74}$$

In practice, a value of 0.01% was found to be a suitable default value for $P(a > a_t)$ and was used throughout the examples given in Chapter 4.

After the zero crossing identification procedure, it is impossible that there are two adjacent bands that are both partial candidates. This can already be considered as a kind of a spectral masking effect. It is, however, still possible that new PTs are installed at some frequency location already claimed by an older tracker. If more than one PT were tracking the same partial, the system could easily become unstable, because of adaptive feedback being involved. In order to prevent this happening, the global initialization threshold (3.74) is complemented by an individual component depending on a wavelet band's present neighborhood in frequency space. If there were no feedback cancellation, the power inferred by a PT with index j into the passband of a band with index i with the transfer function $G_{n,\lambda_i}(f)$ would be

$$P_{s_i} = \sum_j \hat{a}_j^2 \cdot \left| G_{n,\lambda_i}(\hat{f}_j) \right|^2 , \tag{3.75}$$

where $\hat{a}_j$ and $\hat{f}_j$ are the amplitude and frequency estimates of PT j. If the amplitude estimate $\hat{a}_i$ indicated a partial in band number i residing at its center frequency f_{0_i}, the power inferred at the filter output would be

$$P_i = \hat{a}_i^2 \cdot \left| G_{n,\lambda_i}(f_{0_i}) \right|^2 . \tag{3.76}$$

[3]Because of the onset proximity the expectation value will be biased upwards.

We require that

$$P_i > p_s \cdot P_{s_i}, \quad 0 < p_s \leq 1 \tag{3.77}$$

with some constant p_s. With (2.64), this requirement translates to

$$\boxed{\hat{a}_i^2 > \Phi_{s_i}^2 = p_s \cdot \sum_j \frac{\hat{a}_j^2}{\left(1 + \left(\frac{2\pi(\hat{f}_j - f_{0_i})}{\lambda_i}\right)^2\right)^n}}, \tag{3.78}$$

with each of the variables indexed with i and j evaluated at $\hat{t}_0 + t_p(i)$ and $\hat{t}_0 + t_p(j)$, respectively. Partial candidates that have passed the zero crossing test are checked against $a_t + \Phi_{s_i}$. If they also pass this test, a partial tracker is instantiated with its tracking filter being a copy of the i–th filter.

3.3.8 Partial Tracker Death and Offset Time Localization

The first condition for a partial tracker to stay alive at sample index k is

$$a_i(k) \quad > \quad \Phi_t(k) + \Phi_{s_i}(k) \tag{3.79}$$

with $\Phi_t(k)$ being the total global threshold given by (3.68). $\Phi_{s_i}(k)$ is the local spectral masking threshold as given by (3.78), with the difference that both i and j are PT indices and that the summation is carried out for all $j \neq i$ PTs. In order to avoid the possible case of all PTs masking each other, they are first sorted from low amplitudes to high ones and (3.78) is evaluated in this order. If a PT fails to satisfy (3.78) it is masked and its index is excluded for every i following.

Once (3.79) is violated, the partial tracker dies. At very small amplitudes, shortly before a PT is deleted, rounding errors were sometimes found to induce large frequency errors. In order to prevent this happening, it was found advantageous to also impose the following frequency condition for PT survival:

$$\left|\hat{f}_i(k) - f_{0_i}(k)\right| < 2 \cdot \Delta f_i(k). \tag{3.80}$$

It is required that the filter spacing in the wavelet filter bank is sufficiently dense, since otherwise, with the PTs being installed as copies of the wavelet filters closest to the initial estimate, they run danger of being immediately removed as soon as the estimation window is full and adaptation is enabled.

With the PT death conditions (3.79) and (3.80), the precise localization of abrupt partial offsets would be impossible, because the relative slowness of tracker dynamics causes strong smearing of signal discontinuities. Nevertheless, precise offset localization can be achieved by combining the amplitude information delivered by the PTs with the precise timing of the onset detector. As a result of the adaptive feedback cancellation mechanism, offsets appear as onsets to the event detector. With the precisely located offsets and the gammatone unit step response (2.65), an offset of partial i is localized at $\hat{t}_0$, if

$$\boxed{\hat{a}_i\left(\hat{t}_0 + t_p(i)\right) < a_i\left(\hat{t}_0\right) \cdot e^{-\lambda t_p(i)} \cdot \sum_{i=1}^{n} \frac{(\lambda t_p(i))^{n-i}}{(n-i)!}}. \tag{3.81}$$

3.4 Distributed Computation

Like all nontrivial subtasks in artificial intelligence, auditory cognition requires considerable computation power. For this reason the development of concepts for distributed computation is mandatory. However, the task of effectively making use of the computational power offered by systems like the ER II parallel machine consisting of a 256 node crossbar network hosting 128 floating point DSP modules [Mayer-Lindenberg, 1997a; 1997b] is far from being trivial. Innovative object–oriented DSP operating system concepts like the one presented in [Meyer, 1994; Reekie and Meyer, 1994] can be helpful for achieving satisfactory performance. Moreover, on processors operating with integer data types or single–precision floating point numbers, implementations might exhibit unexpected effects due to rounding errors if no software compensations are provided. One example is the possibility of IIR filters becoming unstable (see Appendix D). In the following, general expressions for memory requirements, communication and computation load in dependence on the system parameters are given. Figures for $f_s = 22.05$ kHz and the default parameter set given in Appendix B.1 are calculated.

3.4.1 Memory Requirements

The memory requirement of an algorithm is an important issue. This is especially true for implementations on DSP modules like the ones based on the TMS320C40 by TEXAS INSTRUMENTS with 1Mbyte memory attached, which are working in the PENTAGON parallel machine, a 3–D torus of 5x5x5 nodes [Mayer-Lindenberg, 1995]). As in most DSP applications, the memory space needed for data buffering is the most significant.

3.4.1.1 Master Module

In the proposed architecture the master module is the most demanding component in terms of memory requirements and computation load. A filter realization in *direct form II* [Oppenheim and Shafer, 1975] assumed,

$$l_c = 2 \cdot (2n + 1)\, N_b \tag{3.82}$$

words must be stored for the filter coefficients of the wavelet transform module, where N_b is the number of wavelet bands and n is the gammatone filter order. The factor of 2 is due to the fact that the coefficients in (2.93) are complex numbers.

Due to the causality restriction imposed on realizable filter structures, the operator $\mathcal{Y}$ defined by (3.35) is calculated along an exponential curve in time. Thus, a delay compensation for each filter is necessary, conveniently provided by using proper offsets while reading from the input buffer. The length of this buffer may not be shorter than the maximum distance between the curve of impulse response maxima and the suspected onset sample, i.e.

$$l_p = \mathrm{ceil}\left(f_s \cdot (t_p(n, \lambda_{min}) - t_p(n, \lambda_{max}))\right), \tag{3.83}$$

where $\mathrm{ceil}(x), x \in \mathbb{R}$ is a function rounding x up to the closest integer. This length, however, is not sufficient. As soon as an onset has been identified and initial partial locations have been estimated, a stepback in time to the estimated onset location is performed. Due to the parameter τ_m determining the time the onset algorithm awaits for confirmation after peak detection (see Section 3.3.5), the total buffer length is

$$l_t = \mathrm{ceil}(f_s \cdot (\tau_m + t_p(n, \lambda_{min}))). \tag{3.84}$$

If a stepback occurs, the algorithm must have access to not only the most recent l_t input words, but also to the corresponding state variables of each filter within the main filter bank. A single filter needs

$$l_i = \mathrm{ceil}(2n \cdot f_s \cdot (\tau_m + t_p(n, \lambda_i))) \tag{3.85}$$

words. Thus, the total buffer comprises

$$\begin{aligned} l_{sv} &= \sum_{i=0}^{N_b-1} l_i \\ &= \mathrm{ceil}\left(2n \cdot f_s \cdot \left(N_b \tau_m + \sum_{i=0}^{N_b-1} t_p(n, \lambda_i)\right)\right) \end{aligned} \tag{3.86}$$

words. For the default parameter set given in Appendix B.1 we get $N_b = 60$, $\lambda_{min} = 53.33\frac{1}{\mathrm{s}}$ with (2.72), $t_p(3, \lambda_{min}) = 37.5$ ms with (2.66), resulting in $l_c = 840$, $l_t = 843$ and

$$\begin{aligned} l_{sv} &= \mathrm{ceil}\left(5547.1 + 4961.2 \cdot \sum_{i=0}^{59} 2^{-\frac{i}{12}}\right) \\ &= \mathrm{ceil}\left(5547.1 + 4961.2 \cdot \frac{1 - 2^{-5}}{1 - 2^{-\frac{1}{12}}}\right) \\ &= 174373. \end{aligned}$$

Thus, in this case the total memory requirement in data words is $l_c + l_t + l_{sv} = 176056$. Obviously, this figure would be much lower, if the gammatone filters were realized as FIR filters, since in this case there would not be any state variables to store. On the other hand, this advantage would have to be paid by a considerable increase of computation load.

3.4.1.2 Partial Trackers

The stepback procedure must also be performed by the PTs. If the partial trackers are restricted to the frequency range of the wavelet filter bank, there is no need for another input buffer. The state variables, however, must be memorized separately. As opposed to the wavelet filters, the tracking filters are adaptive, so the effective time window size for partial parameter estimation can vary over time. In order to avoid difficulties with fractionalized buffers in the case of signals with time–varying frequency, each PT is instantiated with the maximum buffer length, which is the length of the state variable buffer of the lowermost filter in the master's main filter bank, i.e. the filter with $\lambda_i = \lambda_{min}$ in (3.85). As the PTs are adaptive, not only the state variables, but also the center frequencies must be stored, so we have a factor of 3 instead of 2 in (3.85):

$$l_{pt} = \mathrm{ceil}\left(3f_s \cdot (\tau_m + t_p(n, \lambda_{min}))\right). \tag{3.87}$$

For the default parameter set in Table B.1 we obtain $l_{pt} = 2527$ words for the buffer size of each PT.

3.4.2 Computation Load

In the following, the computation load of the proposed architecture is roughly estimated. The values given are computations per sample. As each system component operates at the sampling rate, the number of operations per second is obtained by multiplying with f_s. Complex–valued multiplications are counted as 4 real–valued multiplications plus 2 real–valued additions, complex–valued additions as 2 real–valued additions. Complex–valued divisions are counted as 6 real–valued multiplications, 1 division and 3 additions. For $\arctan(x)$, $\sqrt{x}$ and $\log(x)$, each with $x \in \mathbb{R}$, we assume a polynomial of 5–th order to be evaluated, requiring 9 real–valued multiplications and 5 additions (shift operations for normalization neglected) [Ana, 1992].

3.4.2.1 Master Module

With N_b denoting the number of bands in the main filter bank and N_{pt} the current number of PTs, the number of operations per sample associated with the master module is shown in Table 3.2. Calculations performed casually after peak detection are neglected. This results in a total of

$$\begin{aligned}
N_{mul} &= (16n + N_{lf} + 21)N_b + 3N_{lf} + (N_{pt}^2 - N_{pt})(n + 12) + 11N_{pt} + 2, \\
N_{div} &= 2N_{pt}^2 - N_{pt}, \\
N_{add} &= (8n + N_{lf} + 15)N_b + 7N_{pt}^2 - (N_{lf} + 3).
\end{aligned}$$

For the parameter set given in Table B.1 we have $n = 3$, $N_b = 60$ and $N_{lf} = 4$, so 4394 real–valued multiplications, no division and 2573 additions must be calculated

computation		multiply	divide	add
$\mathcal{R}$	(3.37)	–	–	N_{pt}
w_i		$(16n+8)N_b$	–	$(8n+6)N_b$
$\lvert w_i \rvert$		$11N_b$	–	$6N_b$
$\mathcal{Y}$	(3.35)	$2N_b{+}2$	–	$2(N_b-1)$
Φ_p	(3.49)	$10N_{pt}$	N_{pt}	$6N_{pt}$
Φ_{s_i}	(3.78)	$(N_{pt}^2-N_{pt})(n+12)+N_{pt}$	$2(N_{pt}^2-N_{pt})$	$7(N_{pt}^2-N_{pt})$
$N_b \cdot m_1$	(3.59)	–	–	N_b-1
$N_b \cdot m_2$	(3.60)	$N_{lf}N_b$	–	$N_{lf}(N_b-1)$
b	(3.62)	$2N_{lf}$	–	–
Φ_n	(3.64),(3.66)	$N_{lf}+1$	–	–

Table 3.2: Number of operations per sample for the master module.

for each sample at $f_s = 22.05$ kHz, if no PT is present. For a reasonable number of PTs, e.g. $N_{pt} = 10$, we arrive at 5854 real–valued multiplications, 190 divisions and 3273 additions.

Due to a stepback occurring for each onset at which alternations of N_{pt} have been detected, the system must be capable of catching up the time loss, if real–time computation is desired. The minimum distance between two onsets that can be resolved is given by τ_m. The stepback in each band is $\tau_m + t_p(n, \lambda_i)$, summing up to $N_b\tau_m + \sum_{i=0}^{N_b-1} t_p(n, \lambda_i)$. In the worst case of a sequence of onsets at distance τ_m, the system must be capable of calculating $N_b + \sum_{i=0}^{N_b-1} t_p(n, \lambda_i) \cdot \tau_m^{-1}$ times faster than real–time, in order not to fall behind. For the default parameters in Table B.1, this amounts to a factor of

$$N_b + \frac{t_p(n, \lambda_{min})}{\tau_m} \cdot \frac{1-2^{-5}}{1-2^{-\frac{1}{12}}} = 986.25. \tag{3.88}$$

Obviously, even though the case of partial onsets constantly occurring at a distance as narrow as $\tau_m = 698.8$ μs is very unlikely to occur in reality, the master module cannot run in real–time on a single processor of contemporary technology. Fortunately, workload distribution over several network nodes is straight–forward, since each band of the wavelet transform can be executed independently of all others. The only data that would have to be collected within the master module nodes is the residual and intermediate sums for the calculation of $\mathcal{Y}$, $N_b \cdot m_1$ and $N_b \cdot m_2$. Spectral masking can be performed separately in each master module node, if adjacent bands are grouped together. However, master module nodes holding adjacent bands would have to exchange coefficients from those wavelet filters lying close to the boundary.

3.4.2.2 Partial Trackers

The number of computations needed for each PT is shown in Table 3.3. For $n = 3$ we

computation	equation	multiply	divide	add
x	(3.2)	$8n+4$	–	$8n+5$
$\hat{h}_1$	(3.5)	14	1	9
$\hat{f}$	(3.7)	10	1	5
$\hat{s}$	(3.8)	12	–	6
$\hat{a}$	(3.10)	11	–	6
f_0	(3.16)	1	–	2
Δf	$f_0 \cdot Q^{-1}$	1	–	–
λ	(2.72)	1	–	–
$\gamma_a(n,\lambda)$	(2.58)	n	–	–
a_i, b_i	(2.92)–(2.94)	$8n+10$	1	$4n+8$
$d_{n,\lambda}$	(2.80)	–	1	–
Δt	(2.69)	1	–	–
N	(3.26)	1	–	–
g	(3.25)	2	1	2
total		$17n+68$	5	$12n+43$

Table 3.3: Number of operations per sample for a single PT.

arrive at 119 real–valued multiplications, 79 additions and 5 divisions for each sample and each PT. With the TMS320C40 being able to perform 40 MFLOPs, each of the TGs could run on a single node of the PENTAGON network at $f_s = 22.05$ kHz, if we assume a tracker group to contain no more than roughly 8 PTs.

3.4.3 Communication Load

The logical communication structure resulting from the proposed architecture leads straight to the classical master/slave concept, with the master module as master and the tracker groups as slaves. In the setting illustrated in Fig. 3.1, the residual is distributed from the master to each of the PTs and a global residual is formed in the master module by subtracting N_{pt} estimates from the input signal. Furthermore, each PT sends the amplitude and frequency information needed by the master to perform spectral masking. This sums up to N_{pt} messages originating from the master and $3N_{pt}$ messages returning from the PTs. A more efficient way in terms of communication bandwidth would be having the PTs send the complex–valued amplitude, thus saving N_{pt} messages. In this case the master must have memorized the previous values in order to reconstruct amplitude, frequency and real–valued prediction, so this method only works at the expense of additional operations and memory capacity. The setting considered in the following will be the one illustrated in Fig. 3.1. Two important cases of physical communication structures can be distinguished:

1. **Bus structure**: In this case, the master is able to broadcast the residual signal to all PTs listening on the bus with a single physical data packet. Then all PTs send their prediction, amplitude and frequency estimates back to the master, resulting in $3N_{pt} + 1$ messages on the bus. As each participant of the communication is entitled a constant, known message size per sample, collision–free communication can be most conveniently provided by a *round robin* scheme. A possible realization is the following: as a slave receives a command for TG instantiation, it is assigned a unique integer number as an identifier for each PT. Each slave owns a copy of a list containing all identifiers currently present. The sequence of talking on the bus is determined by the numerical order of identifiers. The master owns the identifier 0 and is always the first to talk. Before distributing the next sample of the residual, the master sends control information for the instantiation or deletion of TGs. As partial instantiations and deaths get announced on the bus, each slave performs an update of its copy of the identifier list. With real time computation capability of all components and 16 bit data words transmission assumed, a bandwidth of $f_s \cdot 2$ byte/s is needed for a single logical data link between the master and each of the slaves. With this data rate, the bandwidth of a dedicated 100 Mbit/s Ethernet bus would be sufficient for a reasonable number of PTs, since $(3N_{pt} + 1) \cdot 2\ \text{byte} \cdot 22.05\ \text{kHz} \ll 100\ \text{Mbit/s}$ for $N_{pt} \ll 94$, leaving enough bandwidth for data packaging and tracker control.

2. **Discrete communication channels in a regular network**: It is assumed that the messages cannot be broadcast on a hardware bus. Instead, they have to travel through discrete communication channels, as is the case in most tightly coupled multiprocessor networks. In the case of the Pentagon network, the links run at 4 Mwords/s and each node has six of them. They can be considered as fully bidirectional as they have an 8–word FIFO buffer in each direction and 4 million direction changes per second can performed. As the master module is involved in N_{pt} times as much logical data packet transmissions as any other participant, the communication load in a regular network would be very unevenly distributed. The master must be able to receive $3N_{pt}$ message packets and send the residual out via each of its communication links. Assumed the master module resided on a single Pentagon node with 6 communication links, a total of $3N_{pt}$ messages would have to be received and 6 would have to be sent at the sampling rate f_s. In this case, for a reasonable number of PTs, the bandwidth is sufficient to carry the communication load, since $(3N_{pt} + 6)\text{words} \cdot 22.05\ \text{kHz} \ll 6 \cdot 4\ \text{Mwords/s}$ for $N_{pt} \ll 360$.

3.5 Comparison to Related Approaches

In this section the properties of the architecture presented in this thesis are set into relation to important previous works in the field of multicomponent signal separation.

A conclusive summary is spared for Chapter 5.

[Moorer, 1975]

Moorer's dissertation is a pioneering work about music analysis by computational means. The approach is strictly bottom–up from audio signal level to musical score level, without any feedback involved. There is severe restrictions on the signal class: it must be a strictly harmonic duet performance with the presence of strong fundamentals and without any vibrato, glissando, partial collision and stochastic components involved.

Moorer reports bad experiences with low–level processing based on the discrete Fourier transform. Instead, his system is based on real–valued bandpass filters of constant bandwidth and the *comb filter*, a system with the transfer function $1 - z^{-m}$. Any signal having a periodicity which is an integer multiple of m is canceled out by such a filter. In Moorer's system, the presence of a k–periodic signal is indicated by the amplitude average at the output of the comb filter being close to zero for $m = k$. In the first step, Moorer uses the comb–filter as a periodicity detector in frames of 10 ms duration. Then, fixed 4–th order Chebychev bandpass filters of a constant 20 Hz bandwidth are set to all harmonics of each periodicity detected. In a third step, a comb–filter is used on each bandpass filter output in order to extract the dominant frequency.

As stochastic signal components are not accounted for, it is not surprising that the author reports that "*over 90% of the traces produced by the filtering and pitch detection must be discarded.*" (p.127) The conditions for such a trace to be kept are various heuristical assumptions like a minimum length of 80 ms and some smoothness constraints for amplitude and frequency. Subsequently, traces "*that overlap significantly in time and whose pitches are within a few percent of one another*" are merged and the ones that are left are grouped to notes according to yet another heuristic incorporating assumptions like the requirement for the fundamental to be of "*substantial strength and quality*" (p. 136). In a last step, notes are grouped to melodies by a simple criterion based on a minimum pitch difference.

The transcriptions of the two example duet extracts are amazingly close to the original, but will be difficult to reproduce as many of the system parameters and methods used seem to be ad hoc heuristics and most of the thresholds and parameters remain unspecified. As the author puts it:

> "*It seems to be a property of machine perception programs that they get more and more heuristic and less and less defensible on theoretical bases as they proceed to higher and higher levels of processing, away from the low–level, signal–processing techniques.*" (p.135)

Moorer's work has been very influential to various later approaches. In his outlook, we even find early reasoning about the potentials of innovative concepts like adaptive and multiresolution filtering.

[McAulay and Quatieri, 1986], [Serra, 1989], [Maher, 1990]

McAulay and Quatieri propose an analysis model based on the frame–wise calculation of STFTs, from which significant peaks are selected. Adjacent peaks are heuristically fused to form time–varying tracks. Resynthesis is performed by a bank of oscillators, each oscillator reproducing the frequency and amplitude trajectory of a single track. There is no feedback of the resynthesized signal. Phase accuracy is improved using cubic phase interpolation between adjacent frames. The window size is continuously adapted to 2.5 times the average period of the fundamental found in voiced sections while remaining unaltered in unvoiced ones. The thresholds for birth and death of tracks must be chosen by the user.

Based on the work of McAulay/Quatieri, Serra [1989] proposes a system for decomposing a signal into deterministic and stochastic components. As a consequence of being STFT–based, this architecture has the same disadvantages like constant bandwidth, non–logarithmic distribution of frequency bins and the necessity of phase and amplitude interpolation between frequency bins and time frames. To save computing time, Serra proposes to just ignore the phases. A residual spectrum is calculated by subtraction of the amplitudes of the deterministic components. The extraction and modeling of stochastic components by fitting line segments to the maxima of the residuum is done for the purpose of manipulation and resynthesis only and does not have any influence on the earlier processing stages. This results in the requirement of user interaction for setting thresholds and system parameters in order to reduce the number of spurious partial traces.

Another extension of the McAulay/Quatieri architecture is proposed by Maher [1990]. His aim is to achieve voice separation in a duet performance. The modified algorithm iteratively estimates a pair of fundamentals for each frame by minimizing the weighted partial frequency mismatch. The weighting function is heuristically derived from the associated amplitudes. Maher also tries to cope with the difficulty to separate closely spaced partials. All strategies he proposes rely on the idealized assumption that the fundamental frequencies are resolved exactly, so that the frequencies of the colliding partials can be assumed known. The remaining task then is to estimate the amplitudes. In the first strategy the solution of a two–dimensional linear equation is calculated. A problem with this approach is, that the equation to be solved becomes singular as the components approach each other. The second strategy involves the analysis of the beating frequency. As Maher points out, the results by this strategy are useless "*for notes with duration less than than the beat period and for notes with significant vibrato, tremolo or other amplitude–frequency modulation*". The third strategy involves the use of simple source models and spectral templates. Maher discourages the use of this strategy by terming the results "*not sufficiently encouraging to merit further investigation*". These problems, however, may have been caused by over–simplification of the signal models.

[Cooke, 1993]

Like the architecture presented in this thesis, Cooke's approach is based on an analytic version of the gammatone filter. He uses an order of $n = 4$ as opposed to $n = 3$ which was shown in Section 2.4.8 to yield minimum group delay at a given bandwidth. The filter spacing is $0.3 \cdot ERB$ which is considerably narrower than the default value proposed in this thesis. Cooke's approach does provide resynthesis, but it is only used for model quality evaluation and not for making use of the benefits offered by adaptive feedback cancellation.

The basic building block in Cooke's system is the so–called *place group*. Place groups are obtained by a refined variant of the procedure for zero crossing detection in the detuning function described in Section 3.3.7. As this method is *local in time* by nature, coherence must be achieved by a kind of temporal aggregation heuristic. Cooke calls a sequence of place groups that was grouped together a *synchrony strand*. Strand parameters (frequency, frequency change, amplitude, amplitude modulation rate, "dominance") are updated every millisecond. There is no explicit account for stochastic signal components. The grouping principles employed are harmonicity and common amplitude modulation for combining strands into groups and finally pitch contour similarity for combining subsequent group fragments into trajectories. Onset and offset synchrony remain unused.

[Ramalingam and Kumaresan, 1994], [Wang, 1994], [Nakatani *et al.*, 1995a; 1995b]

The common feature between the architecture presented in this thesis and approaches proposed by [Ramalingam and Kumaresan, 1994] and [Wang, 1994] is, that they comprise an adaptive feedback cancellation mechanism. The most important difference to the system presented here is their inability to handle signal onsets, thus requiring proper tracker initialization through user interaction. Moreover, the loop filters in these architectures have a constant absolute bandwidth as opposed to the constant *relative* bandwidth in our approach.

As already pointed out in Section 2.2.2 the amplitude estimator used by Wang is biased, even in the case of white noise. The frequency estimator he proposes is due to [Kay, 1993]. This estimator is computationally expensive because of its FIR nature. Moreover, it is questionable if these costs are worth the effort, because this estimator is optimal in the maximum likelihood sense only in the high SNR limit of Gaussian white noise. In Wang's proposal, however, Kay's estimator is never used in this context due to the coloring effect of the loop filter.

The architecture proposed in [Nakatani *et al.*, 1995a; 1995b] is also based on adaptive feedback cancellation. The use of STFT–based parameter estimation implies the known disadvantages: the parameters of partials residing between the frequency bins must be interpolated and the compromises caused by constant time–frequency resolution may degrade the results in the high and low frequency limits. As a plus with respect to [Ramalingam and Kumaresan, 1994] and [Wang, 1994], this proposal does not ignore the importance of noise floor estimation.

Common to all three approaches is the suggestion of exploiting partial harmonicity for source separation. The usefulness of this grouping mechanism is well documented in psychoacoustics experiments [Bregman, 1990] and can also be supported by signal–theoretic considerations (e.g. by comparing the CRBs (2.34) and (2.36)). However, it can well be debated if such a mechanism should already be introduced into low–level processing stages. As the partials originating from a sound source do not necessarily have to be strictly harmonic, an artificial listening system making use of this assumption would also have to incorporate a mechanism for dealing with cases in which it does *not* hold. This is especially true for systems containing an adaptive feedback mechanism, where the pursuit of erroneous hypotheses introduces artifacts that might even render the system unstable.

[Baumann, 1995]

In this work, the author departs from psychoacoustic experiments that aim at quantifying the parameters of human auditory perception derived from a *Gestalt*–theoretic viewpoint. The algorithm developed subsequently are strictly feed–forward and do not involve feedback loops. There is no account for stochastic signal components. The spectral analysis method favored by Baumann is a variant of the STFT, the so–called *Fourier t–Transform*. A close inspection reveals that this method is a different way of realizing a gammatone filter bank with constant relative bandwidth. Usually, methods based on the Fourier transform are considered advantageous for computational effectiveness reasons. In case of the *Fourier t–Transform*, however, a different window must be realized for each frequency bin to maintain the constant–Q property. This makes the disadvantage of a linear tiling of the frequency axis instead of a logarithmic one a questionable price to pay.

In several postprocessing steps, partial trajectories are extracted and grouped to partial clusters. These computations are performed within time frames of constant length, thus losing one of the beneficial properties of constant–Q resolution. Numerous parameters are introduced for implementing contrast enhancement, recognition of partial collisions, sequential and harmonic integration. Partial continuity alone is implemented by 6 different parameters, some of them requiring a considerable variability of the settings depending on the properties of the sound material. The decisions have to be made by user interaction.

[Ellis, 1996]

In Ellis' front–end a bank of real–valued, 4–th order, constant–Q gammatone filters is used. The amplitude intensity envelope for each band is derived by half–wave rectification, squaring and smoothing with a one–pole lowpass filter with a fixed time constant of 25 ms. As a consequence, the constant–Q design is degraded at this stage, most notably towards higher frequencies, where the time–constant dominating the intensity response is a fixed 25 ms. Phase information is ignored in this architecture. The onset times are derived from the weighted averaging of times where sudden amplitude changes occur in each band. A short–time autocorrelation is computed for the smoothed, half–wave rectified filter outputs of each band separately. Afterwards a 'summary' autocorrelation is formed.

1. 'Noise clouds' (colored noise with a fixed power spectral density and slowly varying amplification),

2. Transient elements (characterized by onset time, initial spectrum and decay rates for each band),

3. 'Wideband' periodic elements.

These three basic types roughly correspond to noise floor, onset and steady part considered in the thesis presented here. The attribute 'wideband periodic' used by Ellis in conjunction with the third type is somewhat misleading, since if a signal is periodic in the time domain it necessarily consists of distinct narrowband components in the frequency domain. Thus, the 'wideband' nature of such a signal can only be attributed to the spread of these narrowband components over the frequency axis. In Ellis' architecture such a set of narrowband sound elements is grouped by their common contribution to the same peak in the signal's summary autocorrelation, which is a consequence of harmonicity. As a drawback, this grouping principle is likely to fail for signals with strong non–harmonic components, such as bell–like sounds.

Ellis' work is distinct from all approaches mentioned so far, in the sense that it comprises a mechanism for considering multiple hypotheses concurrently. If the actual signal intensity surpasses the expectations resulting from the current hypothesis, the generation of a new sound element is considered, if it is lower than expected the deletion of an existing element is envisaged by the system. At the time where the system recognizes the necessity of introducing or deleting an element, it may fork into different hypothesis branches, each branch with a different set of elements trying to characterize the signal. Discrepancies are found as the differences between measured and predicted signal intensity. The quality measure for each hypothesis branch is the number of bits required to represent the input signal to a fixed level of accuracy. This results in the favoring hypotheses describing the signal with the fewest parameters at the lowest error. As Ellis' system does not incorporate the use of higher–level knowledge yet, some sounds consisting of several different elements are not recognized as a unity and have to be grouped by hand (p.121). Also, thresholds have to be set manually to produce a reasonable output according to the author's judgments (p. 151). Despite this need for manual interception and some shortcomings of the front–end processing, Ellis' work represents a promising first step towards the support of multiple hypotheses in computational auditory scene analysis.

Chapter 4

Results

In this chapter, results illustrating the properties and capabilities of the proposed architecture are presented. For each example, sounds of the originals, the residuals and the resyntheses can be retrieved via `http://www.tu-harburg.de/ti6/pub/diss/`. Unless stated otherwise, the system parameters are the default parameters listed in Appendix B.

4.1 Three Partials with Identical Onset Time

In the first example, the three partials appearing separately in Section 3.3.2 for demonstrating threshold calculation are combined into a single signal. The resulting sound consists of three partials with frequencies $f_0 = 1\text{kHz}$, $f_1 = 500\text{Hz}$, $f_2 = 250\text{Hz}$, all with the same amplitude of 3277 and onset sample of 4410. The top panel of Fig. 4.1 shows the estimated frequency trajectories. The bottom panel shows the two functions determining the onset detection algorithm given in Section 3.3.5, $\mathcal{Z}[s(kT_s)]$ according to (3.38) and the total threshold $\Phi_t(kT_s)$ according to (3.68). All three partials are discovered at the first peak residing at sample index 4420. Several false alarms appear between sample indices 6000 and 8000 and between 10000 and 12000, but are rejected as spurious by the partial initialization procedure. As these spurious onsets slow down computation, it is worthwhile to ask for the reason why the spurious alarms appear and if they can be avoided. As mentioned in Section 3.3.2, the peak level is shared among the detected partials for threshold calculation proportionally to their initial amplitudes. In the multiple partial case, however, the onset does not obey to the unit jump model the threshold adjustment algorithm is designed for. One way to avoid false alarms would be simply raising the threshold, thus making the system more insensitive. This measure, however, might cause misses of onsets in other cases. An alternative approach would be raising the value of the pre–masking time constant τ_m. If τ_m is large enough, the amplitude of the second peak at sample index 4583 would be shared among the partials for threshold calculation instead of the first peak at 4420, but this would make the system localize the onset yet another 163 samples later. It was decided that the disadvantages of these measures for avoiding spurious onset detection are not compensated for by the only benefit, which is a reduced computation time.

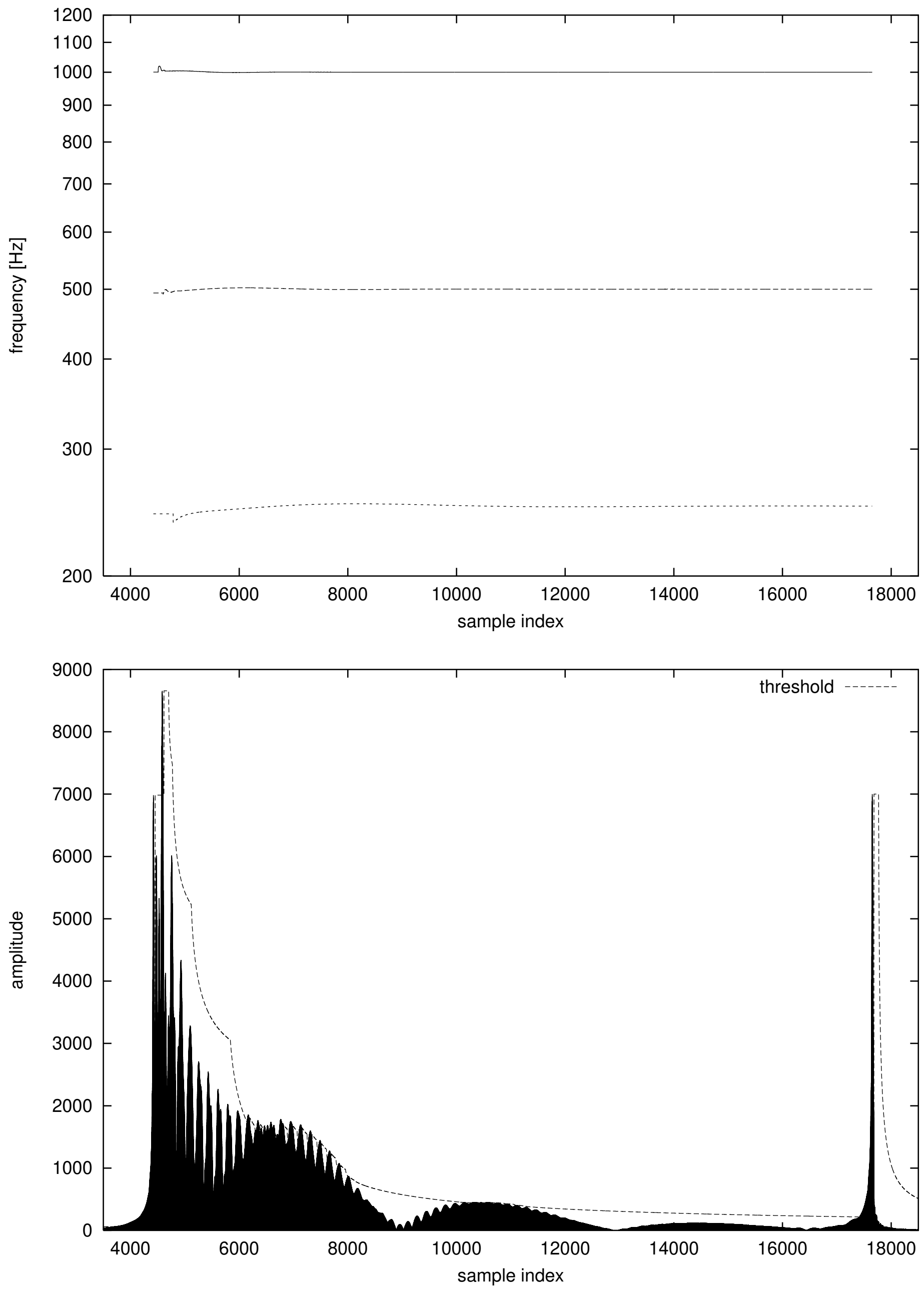

Figure 4.1: Example 4.1 – frequency trajectories (top), $\mathcal{Z}[s(kT_s)]$ according to (3.38) and total threshold $\Phi_t(kT_s)$ according to (3.68) (bottom).

4.2 A Mix of Partials in Noise

Four brief partials immersed in noise appear at three different onset times. This example is the one considered in [Solbach and Wöhrmann, 1996] with use of previous knowledge about the partial locations in frequency. It is a brief sound of 6600 samples at $f_s = 22.05$ kHz. Because of the brevity and the noise that is superimposed, it is at least difficult for human listeners to identify the sound components.

The top panel of Fig. 4.2 shows the development of the frequency estimates over time. In the bottom panel of Fig. 4.2 we see the two functions determining the onset detection algorithm. Sound parameters and analysis results are listed in Table 4.1. While the offsets are localized precisely for all partials, the two low–level partials with onset sample 3308 are detected late. As the analysis log–file indicates, there is indeed an onset detected at sample 3307. However, with tight misses of surpassing the local threshold for both partials, this onset is rejected as spurious by the partial initialization procedure. The partial at 400 Hz is finally accepted 68 samples later at the next crossing of the threshold level, the partial at 700 Hz another 69 samples later. Similar to the example of the previous section, the offsets have a cleaner appearance than the onsets, which is not surprising, since PT removal is instant, whereas initialization needs a certain settling time, even though the initialization procedure involving a stepback allows the maximum relative error of the initial frequency estimates to be as low as 2.1%.

No.	SNR	start sample true/estimated	end sample true/estimated	frequency true/mean/initial [Hz]
1	13.7 dB	1103/1135	6600/6603	500/500.0/493.9
2	13.7 dB	2205/2222	5512/5522	600/600.2/587.3
3	3.2 dB	3308/3376	4410/4405	400/398.6/392.0
4	3.2 dB	3308/3445	4410/4405	700/699.4/698.5

Table 4.1: Signal parameters vs. analysis results.

Fig. 4.3 shows a two–dimensional projection of the partial trajectories in the time–frequency plane with the amplitude on the z–axis. It shows that the discontinuities of partial onsets or offsets cause a slight oscillation in the estimates of concurrent partial trajectories. The analysis time was 1:07 minutes on a Pentium–166 workstation running NeXTStep, the biggest part of the time being required for file–I/O to create numerous log, sound and data files. Although this figure is rather unprecise, it might give an approximative idea about the algorithm's computation load. For a more detailed discussion of this issue, see Section 3.4.2.

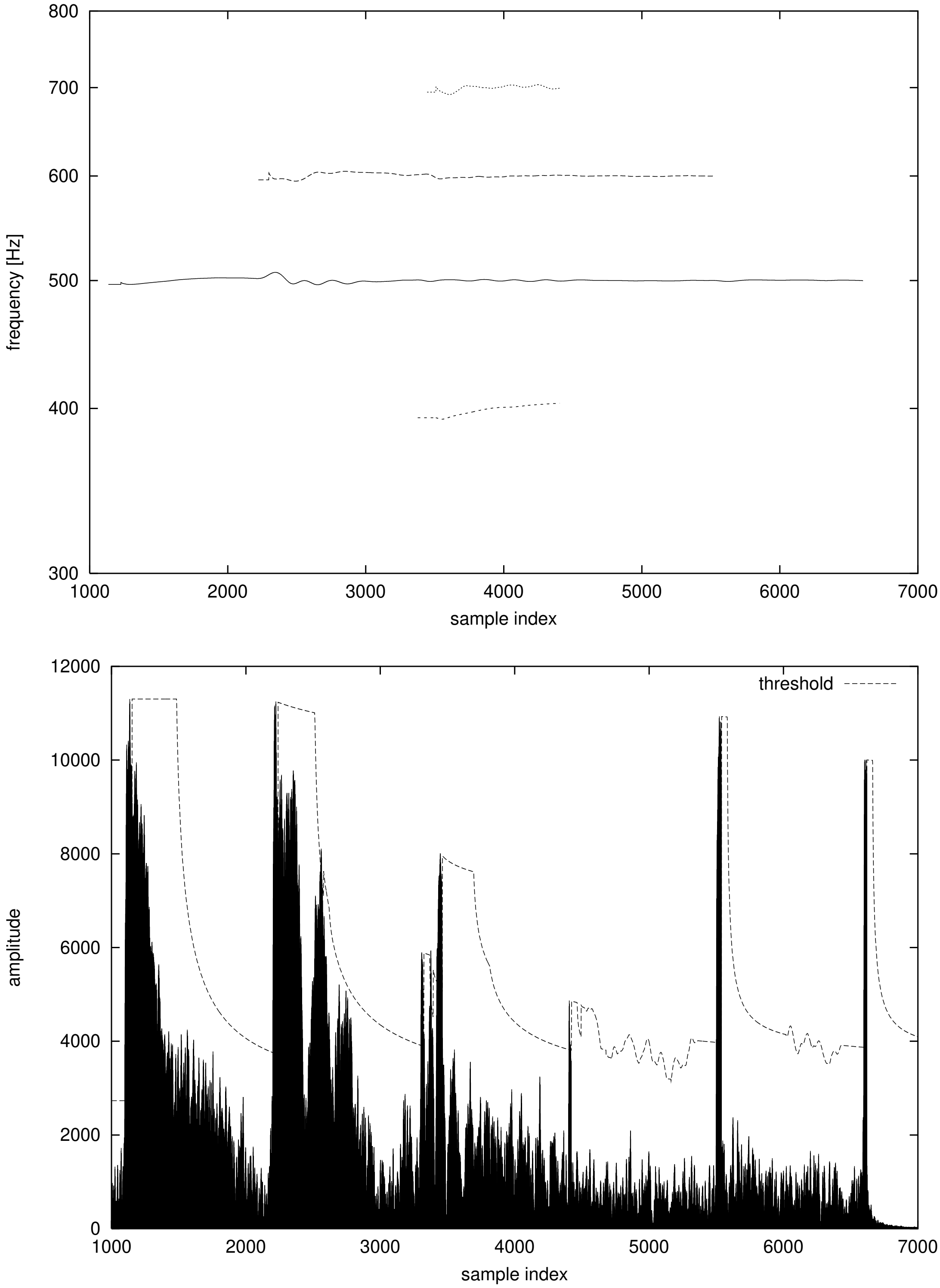

Figure 4.2: Example 4.2 – frequency trajectories (top), $\mathcal{Z}[s(kT_s)]$ according to (3.38) and total threshold $\Phi_t(kT_s)$ according to (3.68) (bottom).

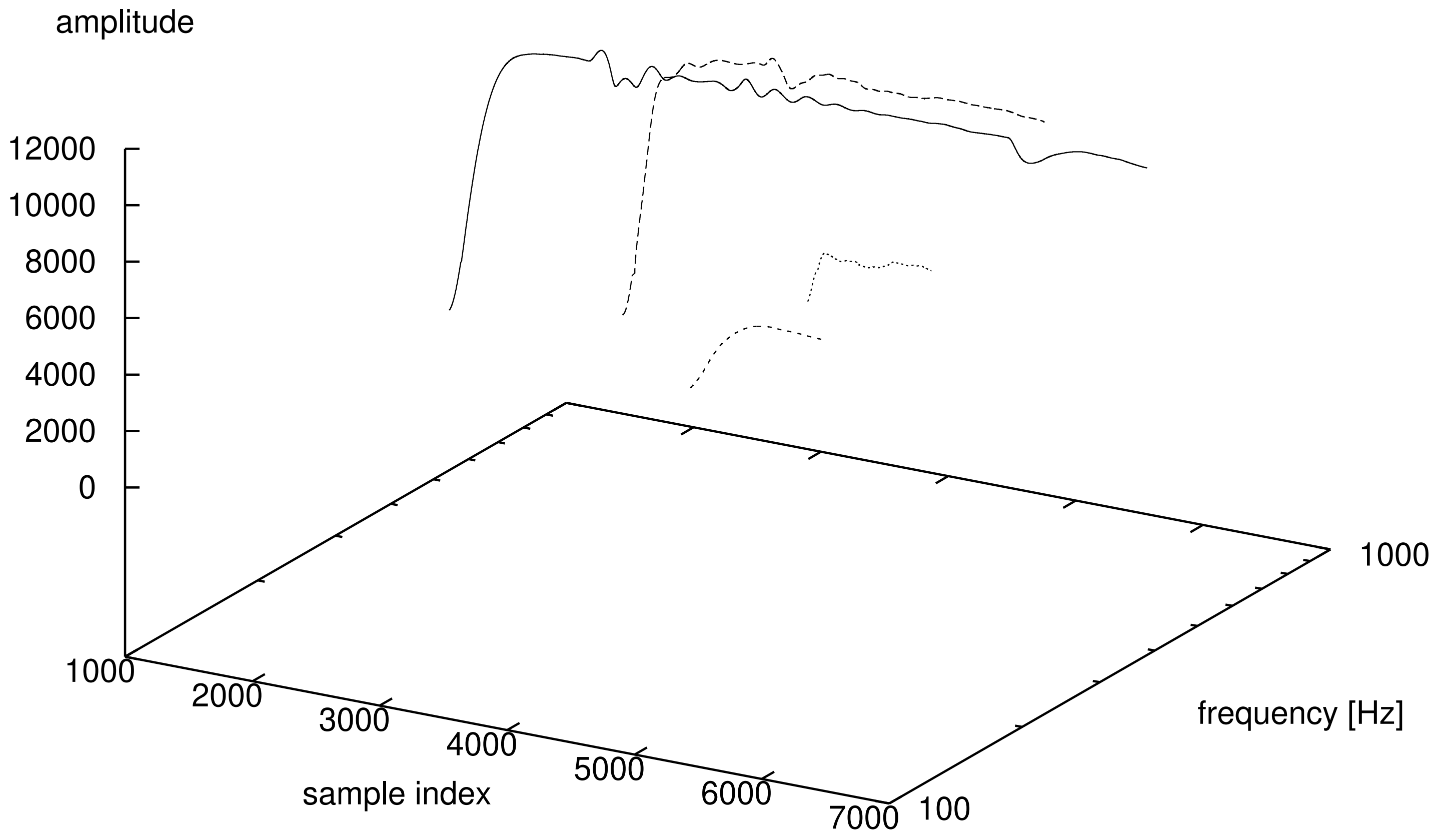

Figure 4.3: Example 4.2 – partial trajectories.

4.3 Partials with Exponential Decay in Noise

J.S. Bach's fugue in C–major from the first volume of the *Well–tempered Clavier* (BWV 846) was synthesized with partials of exponentially decaying amplitude. The synthesis was based on the MIDI file coming with the *Csound*–distribution. See Appendix G for a description of the synthesis details. The resulting partials are of the form

$$a(t) = a_0 \cdot e^{-0.0125 f_0 t} \cdot \sin(2\pi f_0 t). \tag{4.1}$$

where f_0 equals the pitch of the MIDI–note at well–tempered tuning and a_0 depends on the MIDI onset velocity v_m in an exponential fashion with $a_0 = 625$ for $v_m = 1$ and $a_0 = 5000$ for $v_m = 128$. The note duration information contained in the MIDI file remained unused. After synthesis of the sound file, Gaussian white noise of -30 dB power was superimposed. Figure 4.4 shows the partial trajectory estimates for the first four notes. It shows that as the partials fade away, the noise–induced oscillations gradually increase until the PTs are finally removed. Apart from noise removal and softened attack phases, the resynthesized sound is perceptually identical to the original.

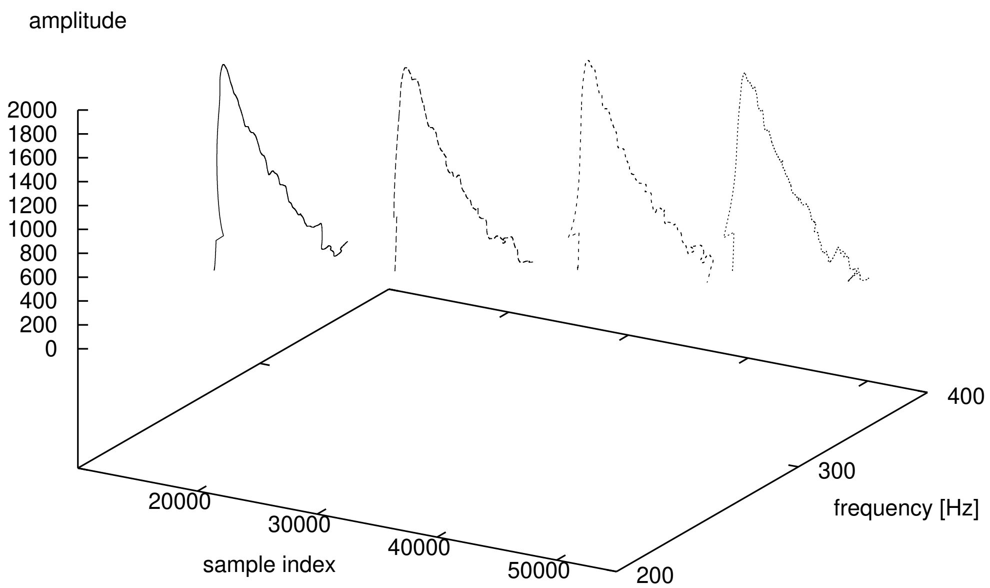

Figure 4.4: Example 4.3 – partial trajectory estimates for the first four notes.

Figure 4.5 relates the score derived from the MIDI file to the mean of the partial frequency estimates for the first two bars of this piece. Although (4.1) makes us expect the durations to be constant for tones residing at the same frequency, this is not the case in the analysis. The durations are shortened in the context of rapid successions of tones lying close to each other in frequency. A close inspection of the analysis log file reveils that this phenomenon is caused by the spectral masking property of the algorithm.

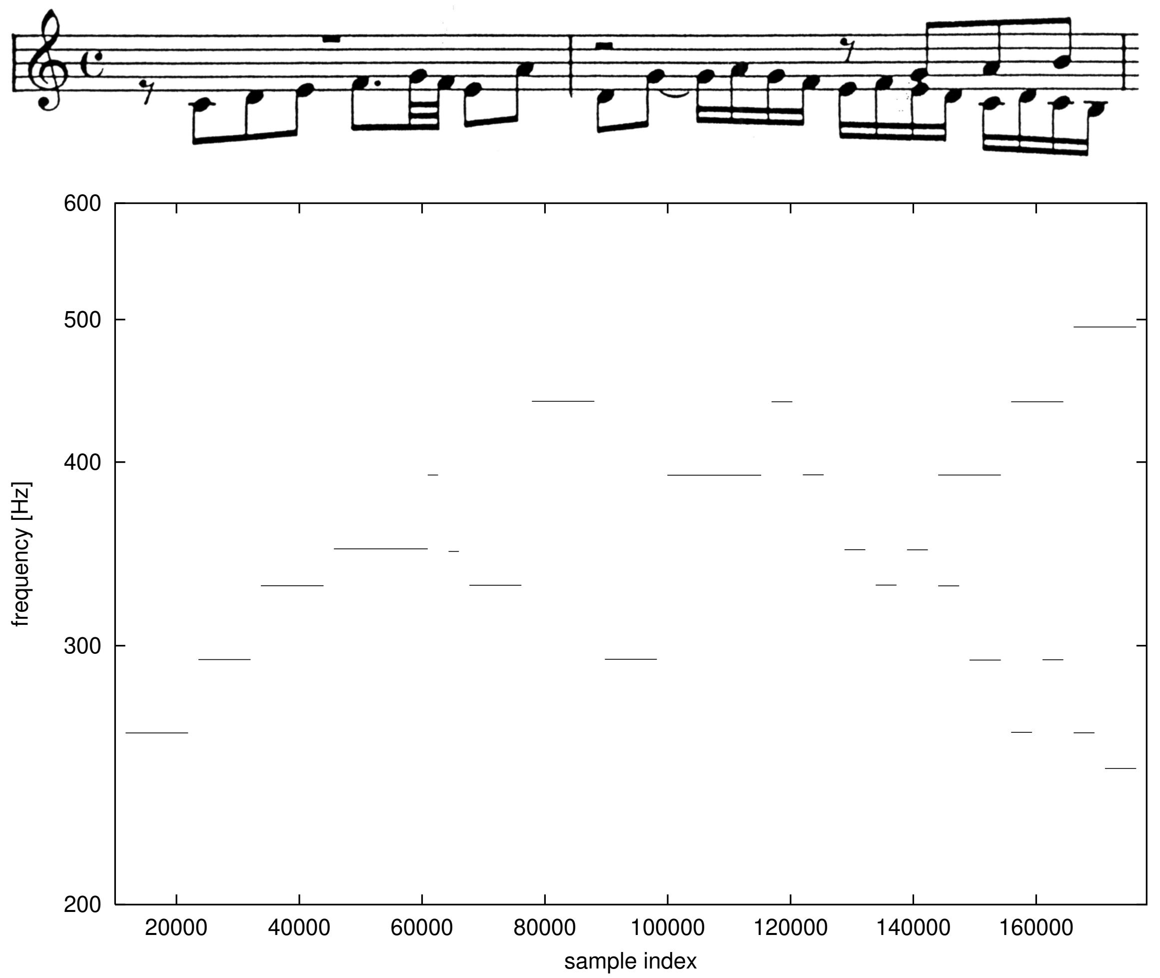

Figure 4.5: Example 4.3 – score and estimated partial frequency means, bar #1 to #2.

The top panel of Fig, 4.6 shows the frequency estimates for the time when the 4–th voice joins in, which is in the middle of bar #5. In the bottom panel we see $\mathcal{Z}[s(kT_s)]$ according to (3.38) and total threshold $\Phi_t(kT_s)$ according to (3.68) for the same time frame.

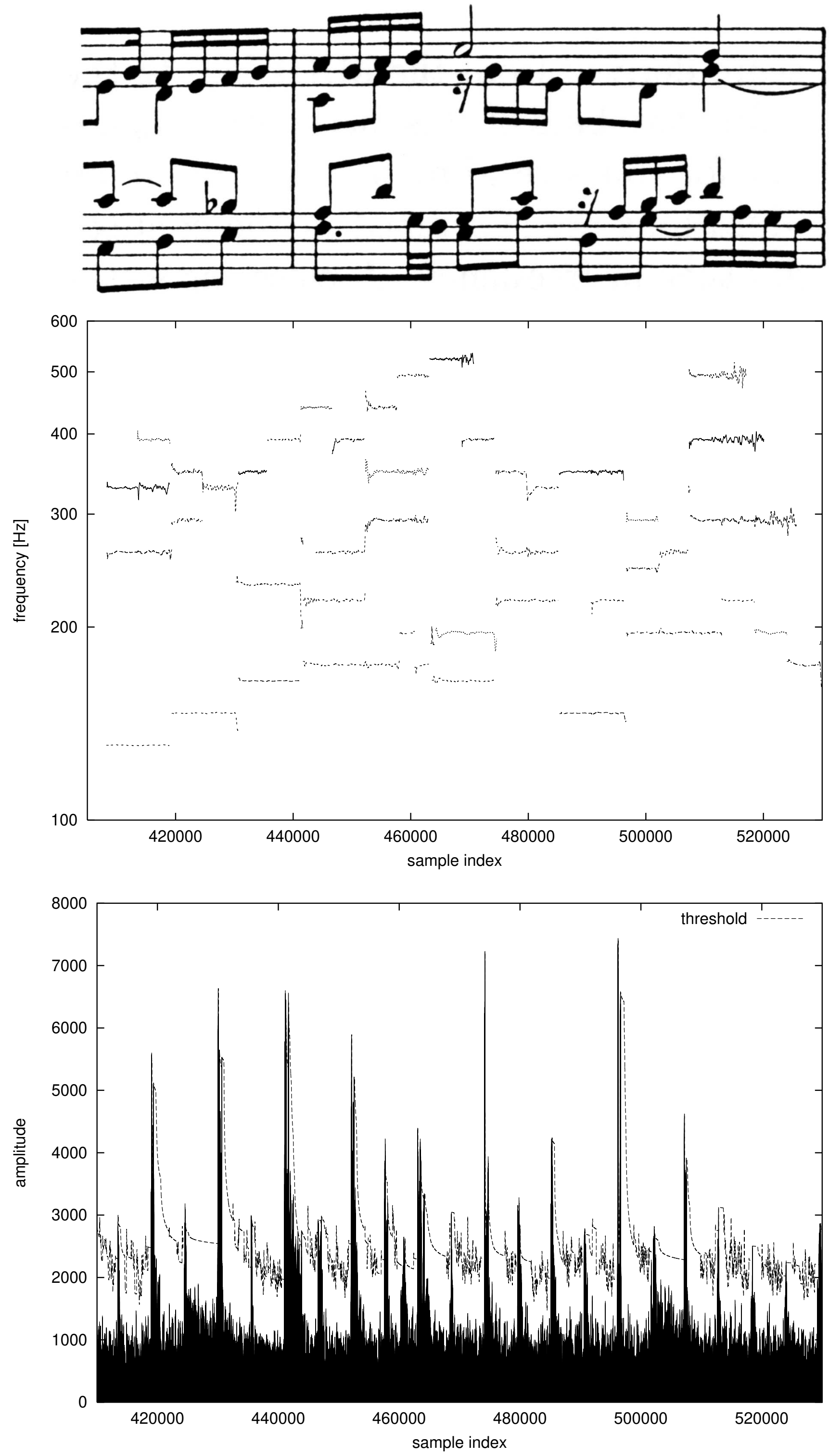

Figure 4.6: Example 4.3 – score, estimated partial frequencies and threshold, bars #5 and #6.

4.4 Piano Tones

The sounds produced by a piano are rich and complex [Rigden, 1976; Hall, 1980]. Neither are the frequencies of the overtones strictly harmonic, nor does the amplitude envelope adhere to rules that can be formulated by simple linear models. Further complications involved are body resonances and mechanical noises in the attack phase. Figure 4.7 shows the analysis of the first measure of Glenn Gold's famous recording of the *Goldberg Variations* by Johann Sebastian Bach (BWV 988) [Gould, 1955], sampled at 44.1 kHz. It consists of two notes being played, G_{-1} and G_{+1}. For both tones, the fundamental is the first partial detected, the upper fundamental at sample index 4343, the lower one at 6172. The detection of the upper harmonics is delayed. In the given example, the 4–th harmonic of the lower fundamental interferes with the fundamental of the upper note. This is the well–known octave problem. Such interferences can only be resolved by the application of suitable sound source models.

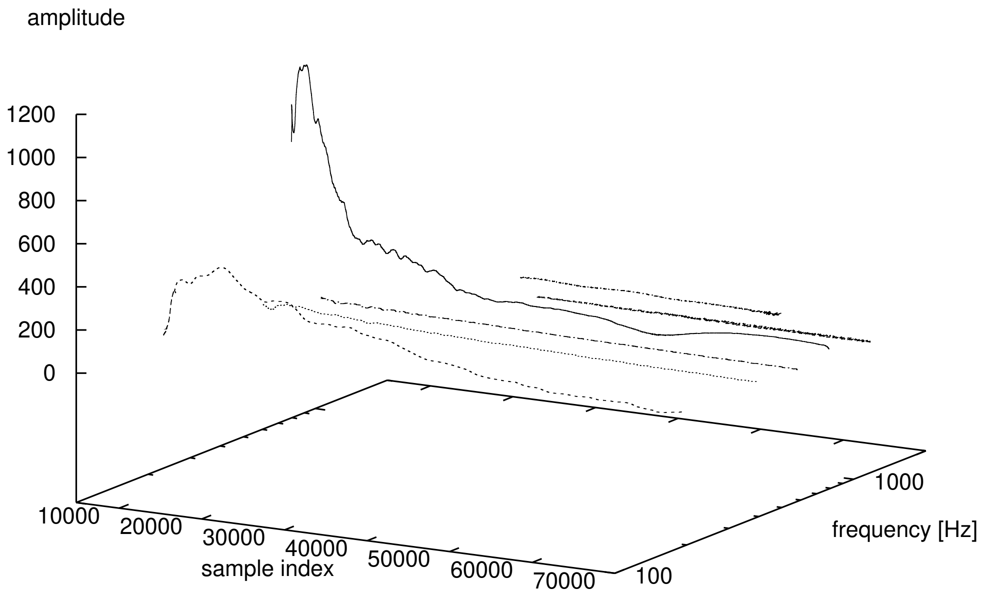

Figure 4.7: Example 4.4 – partial trajectories for the first measure of the Goldberg variations.

Figure 4.8 shows the frequency trajectories for the first few seconds of the piece. The components below 180 Hz can be attributed to the performer's unusual vocal accompaniment.

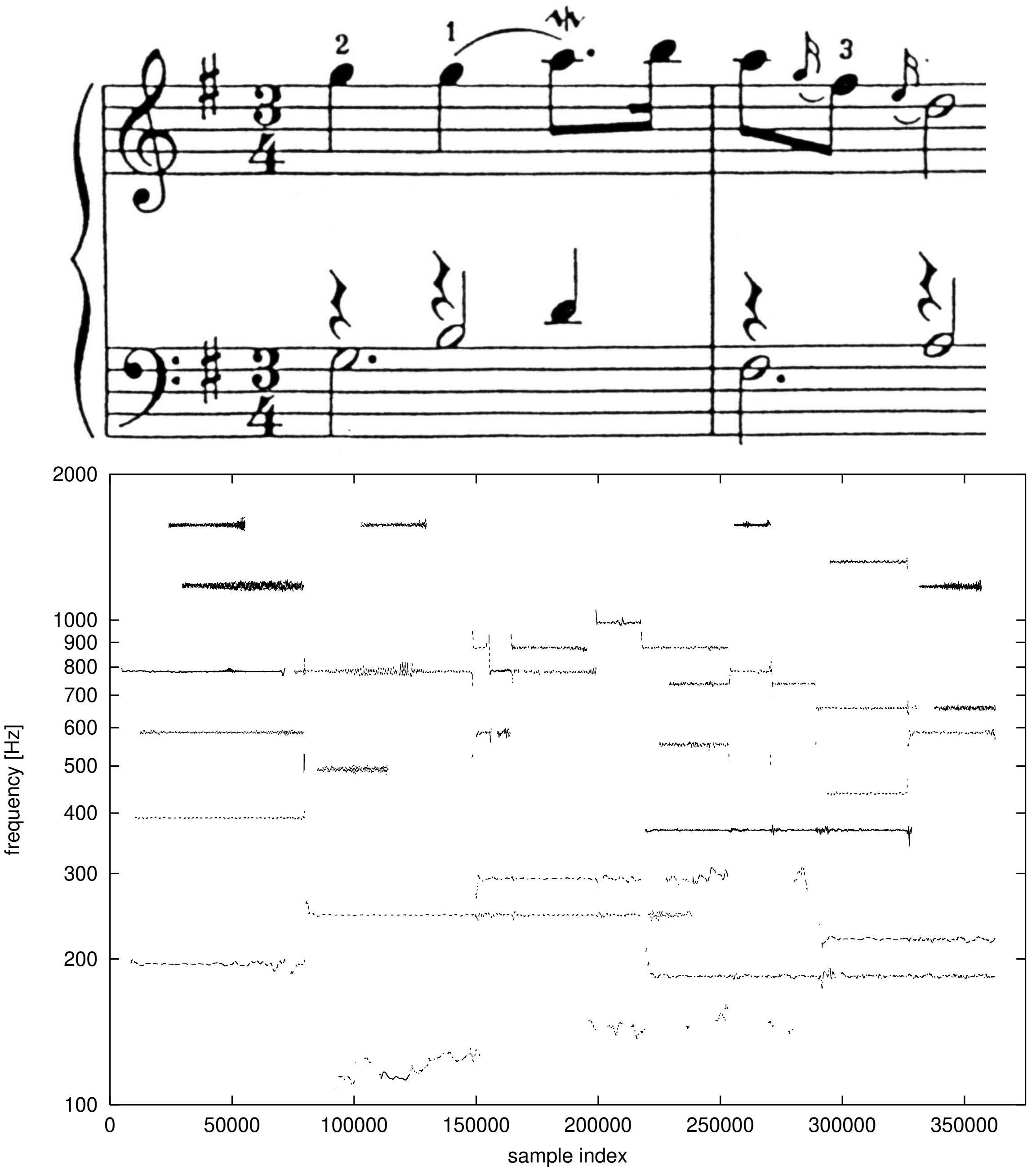

Figure 4.8: Example 4.4 – frequency trajectories for the first few seconds of the Goldberg variations.

4.5 Speech

Although the architecture was not particularily optimized for speech applications, it is interesting to examine its response to such a stimulus. The phrase used in an exemplary speech signal is '*A quick brown fox jumps over the lazy dog*' uttered by a male speaker.

The sampling rate is 44.1 kHz. The default analysis parameters were altered such that the main filter bank comprised 7 octaves starting from $f_{0_{min}} = 49$ Hz. Amazingly, the representation of the fricatives in "fox" and "jumps" is better than would be expected from a representation based on partials. Apparently, the system manages to build noisy components by rapid generation and deletion of partial trackers. A total of 352 partial trackers was produced in this example.

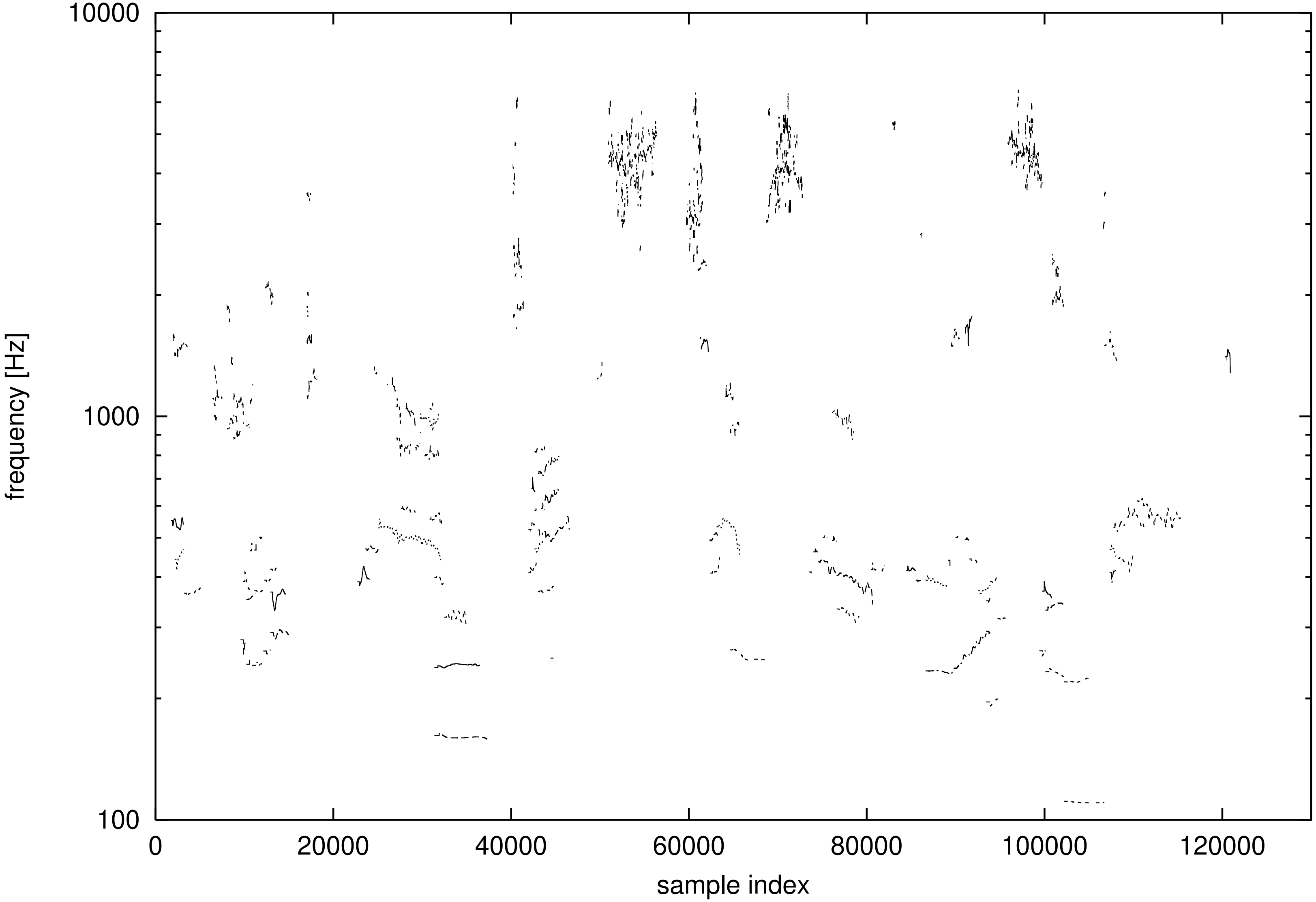

Figure 4.9: Example 4.5 – frequency trajectories for male speech.

Chapter 5

Conclusion and Outlook

In this thesis a new architecture for robust partial tracking and sound onset localization in single channel audio signal mixes has been presented. Two different time–frequency resolutions are evaluated in parallel, a broadband one for onset localization and a narrowband one for partial tracking. The fusion of the two yields a better time–frequency resolution than conventional linear representations. Further advantageous features of the proposed architecture include:

- *Two–pass processing and signal resynthesis with adaptive feedback cancellation.* As opposed to the strict feed–forward, single–pass transformation strategies proposed so far (e.g. [Moorer, 1975; Baumann, 1995; Cooke, 1993]), the system is neither strictly feed–forward nor single–pass. After the detection of new partials subsequent to a hypothesized onset, the system steps back to the onset location and runs a second pass taking the newly gained insight into account. Moreover, adaptive feedback cancellation facilitates noise floor estimation, onset detection and the separation of partials lying close to each other in frequency (see Section 3.2.2).

- *Automated threshold adaptation and continuous noise floor estimation.* These mechanisms serve for keeping the rate of false onset alarms low. The threshold is continuously updated taking previous signal onsets and noise floor estimates into account. By contrast, most of the approaches proposed so far either completely ignore stochastic signal components or treat them as a mere byproduct without any influence on system thresholds (e.g. [Serra, 1989]). The relevance of signal transients for the derivation of suitable threshold levels seems to have been generally disregarded in previous CASA works.

- *Partial trackers are realized with a tracking filter of variable bandwidth.* As opposed to conventional fixed bandwidth realizations (e.g. [Wang, 1994]), the tracking filter bandwidths are kept as a fixed fractional of the center frequencies, which is in accordance with the frequency resolution properties of human auditory perception.

- *No necessity for interpolation post–processing.*
  Being stream–based and frequency–adaptive, the approach avoids the necessity for interpolation between adjacent time frames or frequency bins as given in [McAulay and Quatieri, 1986; Serra, 1989; Maher, 1990] and many others.

- *A small number of system parameters.*
  In the previous approaches mentioned in Section 3.5, the number of parameters that have to be set by the user is often considerable. By contrast, the number of parameters used in the architecture presented in this thesis is comparatively small. Except for the number of octaves in the main filter bank, all examples given in Chapter 4 were obtained with the default parameter settings (see Appendix B).

In the future, the following extensions of the architecture might prove to be useful:

- As the partial phases are maintained, phase–lock detection between partial trackers can be used for partial grouping.

- With the proposed architecture realized separately for the channels of a stereo signal, the high precision of onset estimates can be used to measure inter–channel time differences, an important clue for partial grouping.

- Operating the partial trackers backwards to the onset location instead of a simple step–back would lower the initial tracking error.

- Tightening the tracking filter bandwidth around the current signal bandwidth estimate given by (3.9) can be used for partial pinning. This would result in a yet increased precision of frequency estimates. On the other hand, the partial tracker's ability to follow sudden frequency changes would be deteriorated.

The inclusion of higher–level knowledge in conjunction with top–down information flow has the potential of leading to a leap of quality in signal analysis. The advantages of making an analysis rely on correct hypotheses are increased precision and robustness against disturbing signal superpositions. An example is frequency estimation in the case of harmonic signals: The Cramér–Rao bound (2.34) for the single partial case is higher than (2.36) for the case of a fundamental accompanied by higher harmonics. Thus, if the assumption of a set of partials forming an harmonic set actually holds, the fundamental frequency can be determined more precisely than without this assumption. Yet more specialized assumptions are source models. It would certainly help to have a model of the sounds produced by a Cello if the task is to segregate it from its companions in a string quartet. Also, the octave problem appearing in Section 4.4 for two concurrent piano notes could be solved. As a drawback, however, such a hypothesis will introduce strong artifacts if it is wrong. For this reason, the performance gain by including off–hands heuristics into an analysis system, is generally paid by a loss of generality. The apparent variety of signals in natural environment calls for systems pursuing a multitude of hypotheses in parallel. Research towards such multi–hypothesis systems (e.g. [Ellis, 1996]) is still in its infant stadium. Powerful parallel architectures will be necessary to perform such computations in reasonable time.

Appendix A

List of Symbols and Acronyms

$\forall$	for all
$\exists$	exists
$*$	convolution
$\sim$	is proportional to
$\hat{\bullet}$	estimate of a stochastic variable $\bullet$
AR model	autoregressive model
ASA	auditory scene analysis
$\arg[z]$	phase of $z \in \mathbb{C}$
$\mathbb{C}$	the set of complex numbers
CASA	computational auditory scene analysis
CWT	continuous wavelet transform
$\mathrm{ceil}(x)$	function rounding x upwards to the nearest integer
Δf	frequency window width as defined by (2.16)
Δt	time window width as defined by (2.17)
$\delta(t)$	Dirac impulse
$E\{x\}$	expectation value of a stochastic variable x
$\epsilon(t)$	unit step at $t = 0$
$\mathcal{F}[\,]$	Fourier transform operator as given by Definition 2.1
f	frequency variable
f_0	center frequency as defined by (2.15)
f_s	sampling rate
FIR	finite impulse response
$\Gamma(x)$	gamma function as defined by (2.59)
$\gamma(n, \lambda)$	gammatone filter normalization constant
$\gamma_a(n, \lambda)$	$\gamma(n, \lambda)$ set for amplitude normalization due to (2.58)
$\gamma_e(n, \lambda)$	$\gamma(n, \lambda)$ set for energy normalization due to (2.62)
$\phi_{xx}(\tau)$	autocorrelation function of a time function $x(t)$
$\mathcal{H}[\,]$	Hilbert transform operator as defined by Definition 2.4
IIR	infinite impulse response

$\mathrm{Im}[z]$	imaginary part of $z \in \mathbb{C}$
j	imaginary unit
$L_2(\mathbb{R})$	the space of square integrable functions with real–valued argument
log	natural logarithm
λ	damping constant
$\mathbb{N}$	the set of natural numbers
N_b	number of bands in a wavelet filter bank
N_{lf}	number of line fits for noise floor estimation (see Section 3.3.4)
N_{pt}	number of PTs
n	gammatone filter order
PT	partial tracker
Q	filter quality $\frac{f_0}{\Delta f}$
Q^{-1}	relative bandwidth $\frac{\Delta f}{f_0}$
$\mathbb{R}$	the set of real numbers
$\mathrm{Re}[z]$	real part of $z \in \mathbb{C}$
SNR	signal–to–noise ratio
STFT	short–time Fourier transform
$\mathrm{sgn}(x)$	sign function for $x \in \mathbb{R}$: -1 for $x < 0$, 1 for $x > 0$, 0 for $x = 0$
$\mathrm{si}(x)$	$\frac{\sin(x)}{x}$
σ_x	standard deviation of a stochastic variable x
T_s	sampling interval
TFD	time–frequency distribution
TG	tracker group
t	time variable
τ_m	pre–masking time constant (see Section 3.3.6)
$\mathbb{Z}$	the set of entire numbers

Appendix B

The AnnaLisa Program

All results of this thesis have been obtained with a program named AnnaLisa, that was gradually evolving as the work went on. The authors of the *C++* code are Rolf Wöhrmann and the author of this thesis. In its present incarnation (version 1.0), AnnaLisa runs on NeXTStep, Linux and SGI machines but should be easily portable to any other UNIX–compatible platform.

B.1 AnnaLisa Calling Sequence

The most simple way to call the AnnaLisa exectutable is

```
annalisa <sound_path> <anna_dir>
```

where

- `sound_path` is the path to the sound file to be analyzed. The sound must have 16 bit linear Sun/NeXTStep format (.au or .snd).
- `anna_dir` is the directory into which AnnaLisa puts its output.

The default parameters can be altered using optional arguments. The complete calling syntax of the AnnaLisa exectutable is:

```
usage: annalisa [-T <start> <duration>] [-n order] [-t <tau_m>] [-o <octaves>]
                [-v <voices>] [-r <rel_bw>] [-f <f_0min >] [-N <N_lf>] [-e <eta>]
                [-p <p_s>] [-P <P_a>] <sound_path> <anna_dir>
```

This output is produced by calling `annalisa` without any parameter. The relation between these parameters and the notation used in this thesis is given in Table B.1, together with the respective default values. The number of bands N_b is the number of octaves times the number of voices per octave N_v. The resulting maximum center frequency $f_{0_{max}}$ is

$$f_{0_{max}} = f_{0_{min}} \cdot 2^{\frac{N_b - 1}{N_v}}. \tag{B.1}$$

parameter	denotes	description in Sec.	default value
`start`	start sample for analysis		start of sound file
`duration`	duration of analysis		duration of sound file
`order`	gammatone filter order n		3
`octaves`	number of octaves		5
`voices`	number of bands per octave N_v		12
`f_0min`	$f_{0_{min}}$		98 Hz
`rel_bw`	relative bandwidth $Q^{-1} = \frac{\Delta f}{f_0}$		0.05
`tau_m`	pre–masking time constant τ_m	3.3.6	698.8 μs (see text)
`N_lf`	N_{lf}	3.3.4	4
`eta`	η	3.3.4	1.5
`p_s`	p_s	3.3.6	1.0
`P_a`	$P(a > a_t)$	3.3.7	0.01%

Table B.1: Default parameters.

For the given default parameters we arrive at $f_{0_{max}} = 2960.0$ Hz. If not altered manually, τ_m is calculated from $f_{0_{min}}$ and $f_{0_{max}}$ according to (3.69). With the given default parameters, τ_m is 698.8 μs.

B.2 AnnaLisa Analysis Directory

The main directory structure of the result produced by AnnaLisa is as follows:

```
<anna_dir>/
    ...
    tracker/
        info
        ...
```

where `<sound>` and `<anna_dir>` are the parameters given to the AnnaLisa executable (see calling sequence above). The `info` file in the `tracker/` directory contains several lines, each line corresponding to one tracker group. The syntax is as follows:

```
<sampling rate>
<tag> <onset index> <velocity> <tracker_count>
<tag> <onset index> <velocity> <tracker_count>
<tag> <onset index> <velocity> <tracker_count>
.....
```

where

- `<sampling rate>` is the sampling rate of the sound file,

- `<tag>` is a unique TG number,
- `<onset index>` is the sample index at which the TG was instantated,
- `<velocity>` is the onset velocity $\mathcal{Z}(k_0 T_s)$ (see Section 3.3.5),
- `<tracker_count>` is the number of PTs in the TG.

AnnaLisa puts files with the extension `.fa` into `<anna_dir>/tracker/`. Each of them contains frequency and amplitude information of a single PT. Each line of the `info`–file corresponds to a number of `<tracker_count>` files named `<tag>_1.fa` through `<tag>_<tracker_count>.fa`. These files contain two columns in `ASCII` format, each corresponding to one sample, the first column denoting the estimated frequency, the second column the amplitude estimate.

B.3 AnnaLisa Tools

The programs described in this section were used to further evaluate the data that is left by the AnnaLisa program in the analysis directory `<anna_dir>`. These programs were compiled and tested on NeXTStep, Linux and SGI machines, except `anna2snd`, which does not run on any other system but NeXTStep. The programs `anna2tracker` and `anna2onset` require the *gnuplot* program, version 3.5 (pre 3.6) or later[1].

B.3.1 anna2tracker

Calling Sequence

```
anna2tracker [-f] [-a] [-m] <anna_dir> [<start> <end> [<fmin> <fmax>] ]
```

Description

A partial trajectory graph is created, either as a two dimensional plot with sample index (optionally from `<start>` to `<end>`) vs. either amplitude (option `-a`) or frequency (option `-f`, from `<fmin>` to `<fmax>`, if desired), or as a 3D plot in which both amplitudes and frequencies are jointly displayed (default option). The graph is generated by creating a file named `<anna_dir>/freqPT.gnuplot` (respectively `ampPT.gnuplot` or `3dPT.gnuplot`) and calling the *gnuplot* program. This call creates an instant display and leaves a PostScript file with `.eps` appended to the filename for later use. If `-m` is added to `-f`, the true frequency trajectories are replaced by their mean value of the partial frequency.

[1] *Gnuplot* is a public domain program available from ftp://cmpc1.phys.soton.ac.uk/pub/, a related USENET newsgroup is comp.graphics.apps.gnuplot.

B.3.2 anna2onset

Calling Sequence

```
anna2onset <anna_dir> [<start_sample> <end_sample>]
```

Description

A graph of both the total threshold and $\mathcal{Z}[s(kT_s)]$ according to (3.38) is generated by creating a file named `<anna_dir>/onset.gnuplot` and calling the *gnuplot* program. This call creates an instant display and leaves a file named `<anna_dir>/onset.eps` for later use.

B.3.3 anna2txt

Calling Sequence

```
anna2txt <anna_dir> <midi-info-file> <output file>
```

Description

A MIDI–equivalent text file is generated from the analysis directory. The syntax of this file is such that it can be converted to a true MIDI file by the use of `txt2midi`, a program written by Guenter Nagler (gnagler@ihm.tu-graz.ac.at). It is available from http://hgiicm.tu-graz.ac.at/Cpub. The co–author of `anna2txt` is Dirk Bächle.

B.3.4 anna2snd

Calling Sequence

```
anna2snd <anna_dir>
```

Description

A NeXT sound file named `<anna_dir>/resynth.snd` is generated from the partial tracker data. This sound file contains the original sound minus the residual, which is found in `<anna_dir>/residuum.snd`. As `anna2snd` makes use of the NeXTStep sound library routines, it does not compile on any other operating system.

Appendix C

Rice Distribution

Let $z = x + jy$ be a complex random variable where the real part x and the imaginary part y are uncorrelated Gaussian distributed random variables with zero mean and variances $\sigma_x^2 = \sigma_y^2 = 1$. The distribution of the modulus $r = \sqrt{x^2 + y^2}$ is given by

$$p_0(r) = \epsilon(r) \cdot r \cdot e^{\frac{-r^2}{2}}, \tag{C.1}$$

which is the *Rayleigh distribution* [Kammeyer, 1992]. The probability that r exceeds a certain value r_t is

$$P(r > r_t) = \int_{r_t}^{\infty} p_0(r)\, dr = e^{\frac{-r_t^2}{2}}. \tag{C.2}$$

If a constant $w = a \cdot e^{j\phi}$ is superimposed, the resulting distribution is the *Rice distribution* given by

$$p_a(r) = \epsilon(r) \cdot r \cdot e^{\frac{-(r^2+a^2)}{2}} \cdot I_0\left(ar\right), \tag{C.3}$$

where I_0 is the modified Bessel function of the first kind and zero order. For $a \gg 1$ this distribution approaches the Gaussian, for $a = 0$ it equals the Rayleigh distribution. Fig. C.1 shows the Rice distribution for some parameter values. With I_1 being the modified Bessel function of the first kind and order one, the expectation value $E_a\{r\}$ is [Helstrom, 1995]

$$E_a\{r\} = \sqrt{\frac{\pi}{2}} \cdot \left[\left(1 + \frac{a^2}{2}\right) \cdot I_0\left(\frac{a^2}{4}\right) + \frac{a^2}{2} \cdot I_1\left(\frac{a^2}{4}\right)\right] \cdot e^{-\frac{a^2}{4}} \tag{C.4}$$

shown in Fig. C.2. For $a = 0$ we have the expectation value of the Rayleigh distribution

$$E_0\{r\} = \sqrt{\frac{\pi}{2}}. \tag{C.5}$$

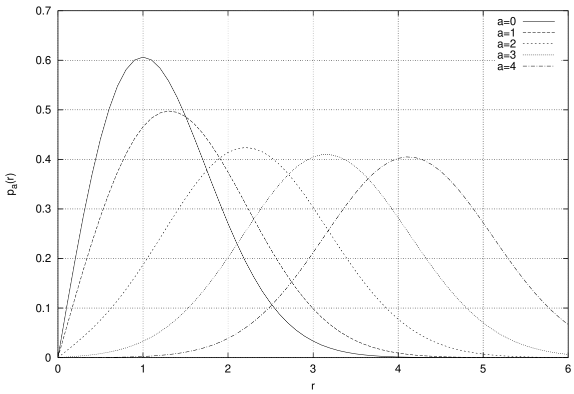

Figure C.1: Rice distribution.

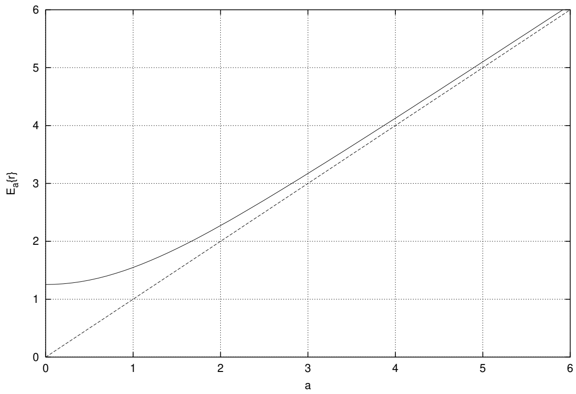

Figure C.2: $E_a\{r\}$ due to (C.4)

Obviously, averaging moduli of a complex variable $w = a \cdot e^{j\phi}$ in Gaussian white noise, yields a biased estimate of the amplitude a. In case the noise power is known, the bias can be removed from an estimate $\hat{a}$ by solving the equation

$$E_a\{r\} = \hat{a} \Longleftrightarrow E_a\{r\} - \hat{a} = 0 \tag{C.6}$$

for the unknown variable a. This can be done by applying the Newton–Raphson iteration [Friedman and Kandel, 1994], with the first derivative of the left side of (C.6)

$$\frac{d}{da}E_a\{r\} = \sqrt{\frac{\pi}{2}} \cdot \frac{a}{2} \cdot \left[I_0\left(\frac{a^2}{4}\right) + I_1\left(\frac{a^2}{4}\right) \right] \cdot e^{-\frac{a^2}{4}}. \tag{C.7}$$

With $f(r) = E_a\{r\}$, $g(r) = \frac{d}{da}E_a\{r\}$, `observation` being the normalized observed value $\frac{\hat{a}}{\sigma_x}$ (assumed $\sigma_x = \sigma_y$) and `error` being the allowed residual approximation error, the algorithm for obtaining an approximation of the true value a can be written in C–language notation as follows:

```
double newton_raphson(double observation, double error)
{
    double s, r = 0.0;

    if (measurement > sqrt(pi/2)) {
        r = measurement;
        do {
                s = (f(r) - measurement) / g(r);
                r = r - s;
        } while(s > error);
    }
    return(r);
}
```

Appendix D

Discrete–Time Approximations of Continuous–Time Systems

There are many different ways of transforming a continuous–time system $H(s)$ to a discrete time system $H(z)$. In the following considerations we restrict ourselves to the methods most frequently appearing in literature, being

- the *impulse–invariant method*,
- the *backward difference method*,
- the *bilinear transform*.

There is a wealth of other methods [Wan and Schneider, 1997], which are not considered here for the sake of brevity.

D.1 Impulse–Invariant Method

With this procedure a discrete–time impulse response is obtained by sampling the continuous time impulse response on a regular grid at the sampling frequency $f_s = \frac{1}{T_s}$, yielding

$$H_d(s) = T_s \cdot \sum_{k=-\infty}^{+\infty} h(kT_s) \cdot e^{-sT_s}. \tag{D.1}$$

Substituting $z = e^{sT_s}$, the transfer function of the discrete–time system is

$$H(z) = T_s \cdot \sum_{k=-\infty}^{+\infty} h(kT_s) \cdot z^{-k}. \tag{D.2}$$

As $z = e^{sT_s}$ holds, the left s–halfplane is mapped into the unit circle, whereas the right half plane is mapped outside. Thus, the impulse invariant transform maintains system stability. The imaginary axis is mapped onto the circle with one cycle corresponding to a frequency interval of bandwidth f_s. This is the reason why for systems with a real–valued impulse response, this method can only be used if the Fourier transform of the impulse response has a sufficient fall–off towards $\frac{f_s}{2}$. In these cases the impulse–invariant transform is the method of choice, because it does not introduce any distortion in the system's frequency response.

D.2 Backward Difference Method

In many cases the impulse–invariant method cannot be applied due to insufficient damping of the system's continuous–time frequency response at half the sampling rate. In these cases the *backward difference method* is one of the possible alternatives.

The derivative operation in the time domain is equivalent to a multiplication with s in the Laplace domain. In the backward difference method the derivative is approximated as

$$x'(t) \approx \frac{x(kT_s) - x((k-1)T_s)}{T_s}, \tag{D.3}$$

corresponding to the substitution

$$s = \frac{1 - z^{-1}}{T_s} \tag{D.4}$$

in the transfer function. Solving for z and setting $s = j\omega$ yields

$$z = \frac{1}{1 - j\omega T_s}. \tag{D.5}$$

From this identity follows that the whole imaginary axis in the s–plane is mapped onto a single cycle of a circle of radius $\frac{1}{2}$ passing through $z = 0$ and $z = 1$. Thus, aliasing is no problem anymore, but this is paid by a heavily distorted frequency response.

D.3 Bilinear Transform

Applying the trapezoid rule for the integration of $x'(t)$ yields

$$x(kT_s) \approx x((k-1)T_s) + (x'(kT_s) + x'((k-1)T_s)) \cdot \frac{T_s}{2}$$

corresponding to the substitution

$$s = 2f_s \cdot \frac{1 - z^{-1}}{1 + z^{-1}}, \tag{D.6}$$

which is the so–called *bilinear transform*. Solving for z and setting $s = j2\pi f$ yields

$$z = \frac{f_s + j\pi f}{f_s - j\pi f}. \tag{D.7}$$

The left s–halfplane is mapped into the unit circle, so stability is maintained. As with the backward difference method, the whole imaginary axis in the s–plane is mapped onto a single cycle of a circle in the z–plane, so aliasing is avoided. As the circle is the unit circle, the bilinear transform introduces much less distortion than the backward difference method. For $f = \frac{f_s}{\pi}$ we get

$$z_0 = \frac{f_s + jf_s}{f_s - jf_s} = j, \tag{D.8}$$

so $f = \frac{f_s}{\pi}$ is mapped onto the vertex of the unit circle. The distortion introduced by the bilinear transform can be calculated by inserting $z = e^{j2\pi f' T_s}$ and $s = j2\pi f$ into (D.6), yielding

$$f' = \frac{f_s}{\pi} \cdot \arctan \frac{\pi \cdot f}{f_s}. \tag{D.9}$$

An example for $f_s = 44.1\text{kHz}$ is shown in Fig. D.1.

D.4 Rounding Errors

Rounding errors have a negative effect on the fidelity of a discrete–time approximation [Zölzer, 1997]. If the sampling rate is very high compared to the absolute value of the real part of a pole location s_0, we have $|e^{s_0 T_s}| \approx 1$, i.e. the pole lies close to the stability boundary in the z–plane. For filters with such poles finite word lenght effects might even lead to a loss of system stability.

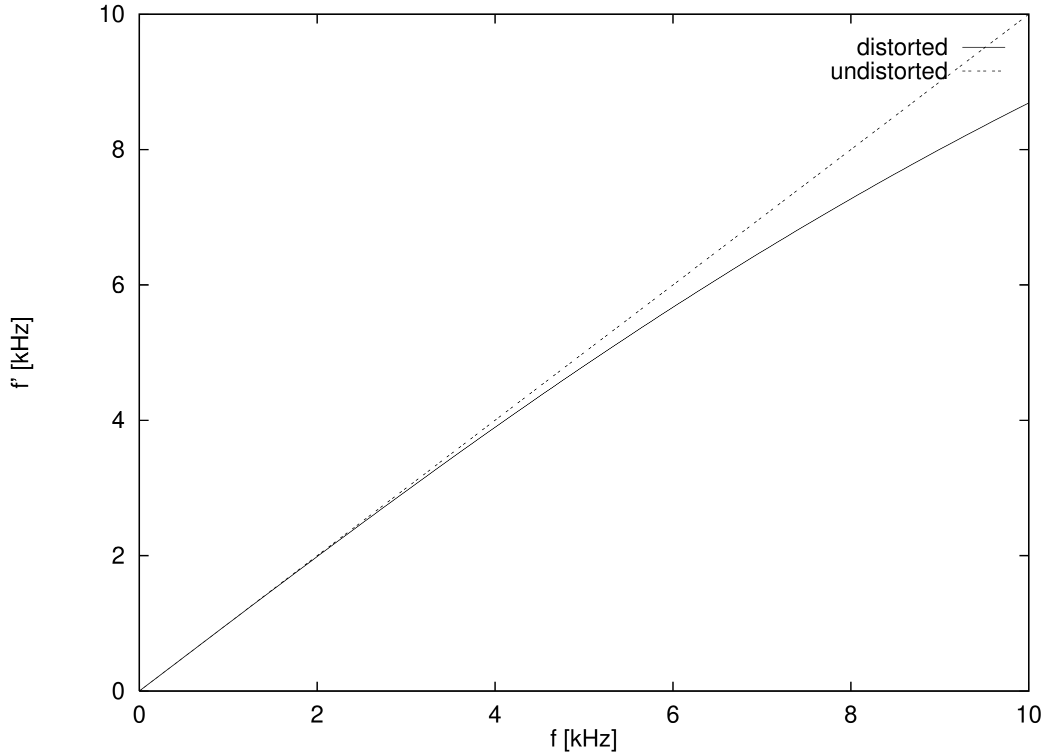

Figure D.1: Frequency distortion caused by the bilinear transform at $f_s = 44.1$kHz.

Appendix E

Maximum Likelihood Estimation

E.1 General Concept

For an unknown parameter vector $\mathbf{x}$ and an observation vector $\mathbf{v}$ we have the joint probability densities

$$\begin{aligned} p(\mathbf{x}, \mathbf{v}) &= p(\mathbf{v}, \mathbf{x}) \\ \Longleftrightarrow\ p(\mathbf{v}) \cdot p(\mathbf{x}|\mathbf{v}) &= p(\mathbf{x}) \cdot p(\mathbf{v}|\mathbf{x}), \end{aligned}$$

where $p(\mathbf{x}|\mathbf{v})$ and $p(\mathbf{v}|\mathbf{x})$ are conditional densities. Rearrangement yields

$$p(\mathbf{x}|\mathbf{v}) = \frac{p(\mathbf{x}) \cdot p(\mathbf{v}|\mathbf{x})}{p(\mathbf{v})}, \tag{E.1}$$

which is an equation commonly known as *Bayes rule*[1]. The term $p(\mathbf{x}|\mathbf{v})$ is called *a posteriori* density, $p(\mathbf{v}|\mathbf{x})$ is the *likelihood*. If $\hat{\mathbf{x}}$ is found such that

$$p(\hat{\mathbf{x}}|\mathbf{v}) = \max\, p(\mathbf{x}|\mathbf{v}), \tag{E.2}$$

the solution vector $\hat{\mathbf{x}}$ is called the *maximum a posteriori* (MAP) estimate. Using (E.1), (E.2) can be written as

$$p(\hat{\mathbf{x}}) \cdot p(\mathbf{v}|\hat{\mathbf{x}}) = \max(p(\mathbf{x}) \cdot p(\mathbf{v}|\mathbf{x})). \tag{E.3}$$

If $p(\mathbf{x})$ is unknown or evenly distributed the criterion reduces to maximizing the likelihood

$$p(\mathbf{v}|\hat{\mathbf{x}}) = \max\, p(\mathbf{v}|\mathbf{x}). \tag{E.4}$$

In order to find the solution we need to solve

$$\frac{\partial}{\partial x_i} p(\mathbf{v}|\mathbf{x}) = 0, \tag{E.5}$$

[1] *An Essay toward Solving a Problem in the Doctrine of Chances*, Reverend Thomas Bayes, 1763

or often more conveniently

$$\frac{\partial}{\partial x_i} \log p(\mathbf{v}|\mathbf{x}) = 0, \tag{E.6}$$

which is a set of conditions equivalent to (E.5), since the logarithm is a monotonic function for positive arguments. The solution $\hat{\mathbf{x}}$ for (E.5) or (E.6) is called *Maximum Likelihood* (ML) estimate. With the *Fisher information matrix* $\mathbf{J}$, whose elements are given by

$$J_{ij} = -E\left\{\frac{\partial^2 \log p(\mathbf{v}|\mathbf{x})}{\partial x_i \partial x_j}\right\}, \tag{E.7}$$

the lower bound for the error variance $E\left\{(x_i - \hat{x}_i)^2\right\}$ is given by the i–th diagonal element of $\mathbf{J}^{-1}$. This lower bound is known as the *Cramér–Rao bound* (CRB). The CRB is a lower bound for the achievable variance of any unbiased estimate. An estimator is called *efficient*, if the CRB is met. If the Fisher information matrix is diagonal (i.e. $J_{ij} = 0$ for $i \neq j$) the CRB is given by

$$\begin{aligned} E\left\{(x_i - \hat{x}_i)^2\right\} &\geq E\left\{\left[\frac{\partial \log p(\mathbf{v}|\mathbf{x})}{\partial x_i}\right]^2\right\}^{-1} && \text{(E.8)} \\ &= -E\left\{\frac{\partial^2 \log p(\mathbf{v}|\mathbf{x})}{\partial x_i^2}\right\}^{-1}. && \text{(E.9)} \end{aligned}$$

E.2 Gaussian Noise Case

If the dependency of the observation vector $\mathbf{v}$ on the parameter vector $\mathbf{x}$ can be written as

$$\mathbf{v} = \mathbf{S} \cdot \mathbf{x} + \mathbf{n} \tag{E.10}$$

where $\mathbf{S}$ is a matrix and $\mathbf{n}$ a vector with Gaussian distributed coefficients, we maximize

$$\begin{aligned} p(\mathbf{v}|\mathbf{x}) &= p(\mathbf{v} - \mathbf{S} \cdot \mathbf{x}) \\ &= \frac{1}{\sqrt{(2\pi)^n \cdot |\mathbf{C_n}|}} \cdot e^{-\frac{1}{2}[\mathbf{v}-\mathbf{S}\mathbf{x}]^T \cdot \mathbf{C_n}^{-1} \cdot [\mathbf{v}-\mathbf{S}\mathbf{x}]}. && \text{(E.11)} \end{aligned}$$

where $\mathbf{C_n}$ is the covariance matrix of the noise. Solving

$$\begin{aligned} \frac{d}{d\mathbf{x}} \log\, p(\mathbf{v}|\mathbf{x}) &= \frac{d}{d\mathbf{x}}\left\{[\mathbf{v} - \mathbf{S}\mathbf{x}]^T \cdot \mathbf{C_n}^{-1} \cdot [\mathbf{v} - \mathbf{S}\mathbf{x}]\right\} \\ &= \mathbf{0} && \text{(E.12)} \end{aligned}$$

yields

$$\hat{\mathbf{x}} = [\mathbf{S}^T \mathbf{C_n}^{-1} \mathbf{S}]^{-1} \mathbf{S}^T \mathbf{C_n}^{-1} \cdot \mathbf{v}. \tag{E.13}$$

If the noise is not only Gaussian but also white, the ML estimator is identical to the least squares estimator, since (E.12) becomes

$$\frac{d}{dx}\left\{[\mathbf{v}-\mathbf{S}\mathbf{x}]^T \cdot [\mathbf{v}-\mathbf{S}\mathbf{x}]\right\} = \mathbf{0}, \tag{E.14}$$

which is solved by

$$\hat{\mathbf{x}} \;=\; [\mathbf{S}^T\mathbf{S}]^{-1}\mathbf{S}^T \cdot \mathbf{v}. \tag{E.15}$$

Appendix F

AR Model Parameter Estimation and Linear Prediction

The underlying assumption in AR model parameter estimation is that the system under consideration can be characterized as an all–pole model of the form

$$F(z) = \frac{G}{1 - \sum_{i=1}^{p} h(i) \cdot z^{-i}}, \quad h(i) \in \mathbb{C}, G \in \mathbb{R}, \tag{F.1}$$

excited by Gaussian white noise of unity variance. The parameter p is called *model order*. After the decision for this model is made, the question how to choose the model order p is still open. If p is too low, the model misses important features of the original, if it is too high, undesired artifacts might appear. Neglecting the constant G in (F.1), the inverse of $F(z)$ is

$$H(z) = 1 - \sum_{i=1}^{p} h(i) \cdot z^{-i}, \tag{F.2}$$

for which Fig. F.1 shows a possible realization.

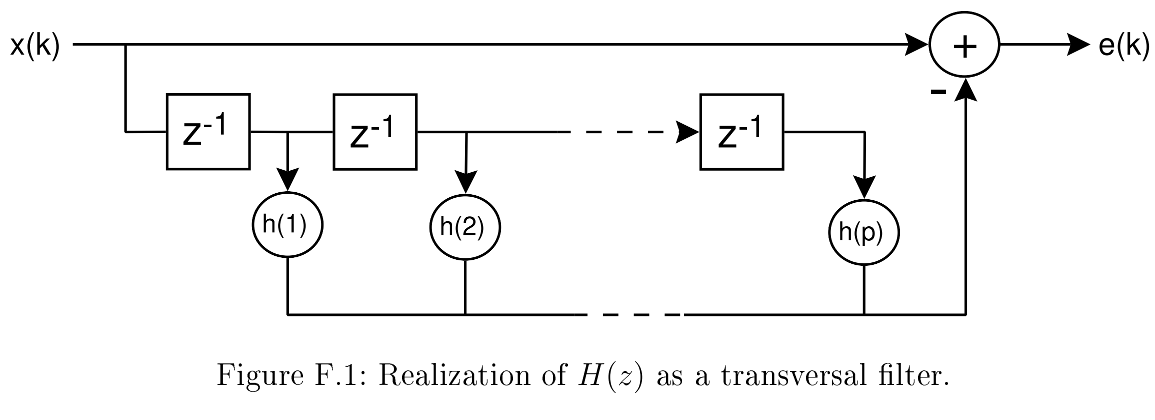

Figure F.1: Realization of $H(z)$ as a transversal filter.

The goal of AR model estimation is to adjust the *predictor coefficients* $h(i)$ of the system shown in Fig. F.1 such that

$$e(k) = x(k) - \sum_{i=1}^{p} h(i) \cdot x(k-i), \tag{F.3}$$

with $x(k)$ being the output of the original noise–driven system given by (F.1), is a white process. It can be shown, that this requirement is equivalent to minimizing the signal power of $e(k)$.

In practice, only a limited sample number of $x(k)$ can be observed. The objective is to calculate the predictor coefficients such that the square of $e(k)$, averaged over the most recent observations, is minimized. Considering the last $N \geq 2p$ samples, the $N-p$ most recent prediction errors can be taken into account. Thus, the optimization criterion is

$$\eta(n) = \sum_{k=n-N+p+1}^{n} |e(k)|^2 = \sum_{k=n-N+p+1}^{n} \left| x(k) - \sum_{i=1}^{p} h(i) \cdot x(k-i) \right|^2 \longrightarrow \min, \tag{F.4}$$

which is the criterion employed in the so–called *covariance method*. There are several other least squares criteria differing with respect to the summation range. The criterion employed in the so–called *autocorrelation method*[1] $\sum_{k=n-N+1}^{n} |e(k)|^2 \to \min$ uses $2p$ excess data points beyond the N measured ones, which are usually assumed to be zero. This criterion always yields minimum phase solutions at the expense of poorer spectral resolution and a bias of the pole locations towards the center of the unit circle [Kay, 1988; Marple, 1987]. While stability is essential in cases where the identified filter is actually synthesized, the covariance method is preferable for parameter estimation in nonstationary environments.

In order to minimize (F.4), the partial derivatives with respect to the $h(k)$ are determined as

$$\begin{aligned}
\frac{\delta\eta(n)}{\delta h(k)} &= \sum_{k=n-N+p+1}^{n} \Bigg\{ -\left(x(k) - \sum_{i=1}^{p} h(i) \cdot x(k-i) \right) \cdot x^*(k-\kappa) \\
&\quad - \left(x(k) - \sum_{i=1}^{p} h(i) \cdot x(k-i) \right)^* \cdot x(k-\kappa) \Bigg\} \\
&= \sum_{i=1}^{p} h(i) \cdot \sum_{k=n-N+p+1}^{n} x(k-i)x^*(k-\kappa) - \sum_{k=n-N+p+1}^{n} x(k)x^*(k-\kappa) \\
&\quad + \sum_{i=1}^{p} h^*(i) \cdot \sum_{k=n-N+p+1}^{n} x^*(k-i)x(k-\kappa) - \sum_{k=n-N+p+1}^{n} x^*(k)x(k-\kappa)
\end{aligned}$$

[1] Although the names *autocorrelation method* and *covariance method* are not related to the autocorrelation and covariance of a stochastic process, these notions are adopted for historical consistency with the signal processing literature.

with $\kappa = 1, 2, \ldots, p$. Demanding $\frac{\delta\eta(n)}{\delta h(k)} = 0$ yields the following matrix equation:

$$\sum_{i=1}^{p} h(i) \cdot \sum_{k=n-N+p+1}^{n} x(k-i)x^*(k-\kappa) = \sum_{k=n-N+p+1}^{n} x(k)x^*(k-\kappa), \quad \kappa = 1, 2, \ldots, p. \tag{F.5}$$

With $\phi_{xx}(\kappa, i) = \sum_{k=n-N+p+1}^{n} x^*(k-\kappa)x(k-i)$ (F.5) can be written as

$$\begin{pmatrix} \phi_{xx}(1,1) & \phi_{xx}(1,2) & \phi_{xx}(1,3) & \ldots & \phi_{xx}(1,p) \\ \phi_{xx}(2,1) & \phi_{xx}(2,2) & \phi_{xx}(2,3) & \ldots & \phi_{xx}(2,p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{xx}(p,1) & \phi_{xx}(p,2) & \phi_{xx}(p,3) & \ldots & \phi_{xx}(n,p) \end{pmatrix} \cdot \begin{pmatrix} h(1) \\ h(2) \\ \vdots \\ \vdots \\ h(p) \end{pmatrix} = \begin{pmatrix} \phi_{xx}(1,0) \\ \phi_{xx}(2,0) \\ \vdots \\ \vdots \\ \phi_{xx}(p,0) \end{pmatrix} \tag{F.6}$$

or short in matrix notation as

$$\underline{\Phi} \cdot \mathbf{h} = \Phi. \tag{F.7}$$

The vector $\mathbf{h_f(n)}$ solving this equation is the *forward predictor*. By making use of the hermitian structure of $\underline{\Phi}$ the solution can be found in $O(p^2)$ [Marple, 1987]. The same procedure can be carried out for the complex–conjugate time–reversed data set yielding the *backwards predictor* $\mathbf{h_b(n)}$. Both sets of coefficients characterize an all–pole linear system in the form of (F.1). In the so–called *modified covariance method* the forward and backward prediction errors are jointly minimized [Kay, 1988]. Neither variant of the covariance method is guaranteed to yield minimum phase parameters, which is not only acceptable but unavoidable in nonstationary environments. It is important to note that – as with the autocorrelation method – the presence of observation noise[2] negatively affects the estimation. The reason for this is easy to see as, when compared to the noiseless case, the spectrum appears as an AR process with widened passbands around the maxima [Marple, 1987; Kay, 1988].

The covariance method is closely related to the *least squares Prony method*, where the data is assumed to consist of a sum of frequency modulated damped exponentials. After calculating the coefficients of the predictor filter, the roots z_i of the associated polynomial (i.e. the poles of the inverse all–pole model) are computed. Once the roots are found, the frequencies f_i are estimated as

$$\hat{f}_i = \frac{f_s}{2\pi} \cdot \arg(z_i), \tag{F.8}$$

and the damping constants λ_i as

$$\hat{\lambda}_i = \log|z_i| \cdot f_s. \tag{F.9}$$

[2]Not to be confused with the excitation noise of the model.

In a final step, with f_i and λ_i assumed known as computed, the initial complex amplitudes can be calculated by solving a linear matrix equation with the measured data as the target vector.

Appendix G

Synthesis of a Sound File from MIDI Data Using *Csound*

Csound is a language for sound synthesis and processing developed by Berry Vercoe at the MIT media lab. It is available for a multitude of different computer platforms. The most recent version (3.47 as of the writing of this thesis) plus documentation can be downloaded from `ftp://ftp.maths.bath.ac.uk/pub/dream/`. The fugue example of Section 4.3 was synthesized from the MIDI file `Fugue#1.mf` accompanying the *Csound*–distribution with the following command line:

```
csound -o fugue1.snd -F Fugue#1.mf exp.orc exp.sco
```

The files `exp.orc` and `exp.sco` are given below.

```
;;;; CSOUND score file
;;;  exp.sco

;;; function table 1 contains a single sine
;;  16384 samples, starting at 0
f 1 0 16384 10 1

;;; function 2 is an exponential
;;  used for mapping midi velocities to amplitudes,
;;  128 samples, starting at 0, from 1/8 to 1 in 128 steps
f 2 0 128 5 1 128 8

;;; function 0 defines start and end time
f 0 600

;;; end
e
```

```
;;;; CSOUND orchestra file
;;;  exp.orc

;;; sampling rate
sr = 22050

;;; control rate
kr = 2205

;;; sampling rate / control rate
ksmps = 10

;;; number of channels
nchnls = 1

;;; instrument is a sine with exponentially decaying amplitude
instr 1
    ;; frequency is pitch of midi note
    ifreq     cpsmidi

    ;; max amplitude depends on midi velocity through
    ; function table 2, normalization to 5000
    iamp      ampmidi 5000, 2

    ;; time constant for amplitude
    itau = 80/ifreq

    ;; frequency / time_const = const.
    ;           start_val, delta_t , val_at_delta_t
    amp expon    iamp    , itau    , iamp*exp(-1)

    ;; generate signal using function table 1
    asignal      oscil amp, ifreq, 1

    ;; connect oscillator with output
    out asignal
endin
```

Bibliography

[Allen, 1985] J. Allen. Cochlear Modelling. *IEEE ASSP Magazine*, 2(1):3–29, January 1985.

[Ana, 1992] Analog Devices. *Digital Signal Processing Applications Using the ADSP–2100 Family, Volume 1*, 1992. Prentice Hall, Englewood Cliffs.

[Baumann, 1995] U. Baumann. *Ein Verfahren zur Erkennung und Trennung multipler akustischer Objekte.* PhD thesis, Technische Universität München, 1995.

[Beckenbauer, 1989] Th. Beckenbauer. *Spektrale Inhibition als Mittel zur Sprachverarbeitung.* PhD thesis, Technische Universität München, 1989.

[Boashash, 1992] B. Boashash. Estimating and Interpreting the Instantaneous Frequency of a Signal – Part 1: Fundamentals. *Proceedings of the IEEE*, 80 (4):520–538, April 1992.

[Bregman, 1990] A.S. Bregman. *Auditory Scene Analysis.* MIT press, Cambridge, Massachusetts, 1990.

[Chirlian, 1994] P.M. Chirlian. *Signals and Filters.* van Nostrand Reinhold, 1994. ISBN 0–442–01324–8.

[Clarkson, 1993] P.M. Clarkson. *Optimal and Adaptive Signal Processing.* CRC Press, 1993.

[Cohen, 1992] L. Cohen. What Is a Multicomponent Signal? In *International Conference on Acoustics, Speech and Signal Processing*, volume V, pages 113–116, March 1992.

[Cohen, 1993] L. Cohen. The Scale Representation. *IEEE Transactions on Signal Processing*, 41:3275–3292, December 1993.

[Cohen, 1995] L. Cohen. *Time–Frequency Analysis.* Prentice Hall PTR, Englewood Cliffs, New Jersey 07632, 1995.

[Coifman and Wickerhauser, 1992] Ronald R. Coifman and Mladen Victor Wickerhauser. Wavelets and adapted waveform analysis. In John J. Benedetto and Michael Frazier, editors, *Wavelets: Mathematics and Applications*, Studies in Advanced Mathematics, pages 399–423. CRC Press, 1992.

[Cooke, 1993] M. Cooke. *Modelling Auditory Processing and Organisation.* Cambridge University Press, New York, 1993.

[Daubechies, 1990] I. Daubechies. The Wavelet Transform, Time–Frequency Localization and Signal Analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005, September 1990.

[Davenport, 1970] W.B. Davenport. *Probability and Random Processes.* McGraw–Hill, 1970.

[El-Maleh and Kabal, 1997] Kh. El-Maleh and P. Kabal. Comparison of Voice Activity Detection Algorithms for Wireless Personal Communication Systems. *Proc. IEEE Canadian Conference on Electrical and Computer Engineering*, pages 470–473, May 1997.

[Ellis, 1996] D.P.W. Ellis. *Prediction–Driven Computational Auditory Scene Analysis.* PhD thesis, Massachusetts Institute of Technology, 1996.

[Fliege and Zölzer, 1990] N.J. Fliege and U. Zölzer. Multi–Complementary Filter Bank: A New Concept with Aliasing–Free Subband Signal Processing and Perfect Reconstruction. *Proc. EUSIPCO'92*, pages 207–210, 1990.

[Fliege, 1991] N.J. Fliege. *Systemtheorie.* Teubner, Stuttgart, 1991.

[Fliege, 1994] N.J. Fliege. *Multirate Digital Signal Processing.* Wiley, 1994.

[Friedlander and Francos, 1995] B. Friedlander and J.M. Francos. Estimation of Amplitude and Phase Parameters of Multicomponent Signals. *IEEE Transactions on Signal Processing*, pages 917–925, April 1995.

[Friedman and Kandel, 1994] M Friedman and A. Kandel. *Fundamentals of Computer Numerical Analysis.* CRC Press, Inc., Boca Raton, Florida 33431, 1994.

[Ghitza, 1992] O. Ghitza. Auditory Nerve Representation as a Basis for Speech Processing. In S. Furui and M.M. Sondhi, editors, *Advances in Speech Signal Processing*, pages 453–485. Dekker, New York,Basel,Hong Kong, 1992.

[Gould, 1955] G. Gould. J.S. Bach: Goldberg Variations. CBS Compact Disc LMYK44868, 1955.

[Gröchenig, 1993] K. Gröchenig. Acceleration of the Frame Algorithm. *IEEE Transactions on Signal Processing*, 41(12):3331–3340, March 1993.

[Hall, 1980] D.E. Hall. *Musical Acoustics.* Wadsworth, Inc., 1980. ISBN 0–534–00758–9.

[Helstrom, 1995] C.W. Helstrom. *Elements of Signal Detection.* PTR Prentice–Hall, Englewood Cliffs. NJ 07632, 1995.

[Hlawatsch and Boudreaux-Bartels, 1992] F. Hlawatsch and G.F. Boudreaux-Bartels. Linear and Quadratic Time–Frequency Signal Representations. *IEEE SP Magazine*, pages 21–67, 1992.

[Irino and Patterson, 1997] T. Irino and R.D. Patterson. A Time–Domain, Level–Dependent Auditory Filter: The Gammachirp. *J. Acoust. Soc. Am.*, 101:412–419, January 1997.

[Irino, 1996] T. Irino. A 'Gammachirp' Function as an Optimal Auditory Filter with the Mellin Transform. In *International Conference on Acoustics, Speech and Signal Processing*, pages 981–984, May 1996.

[James *et al.*, 1994] B. James, B.D.O. Anderson, and R. C. Williamson. Conditional Mean and Maximum Likelihood Approaches to Multiharmonic Frequency Estimation. *IEEE Transactions on Signal Processing*, pages 1366–1375, 1994.

[Kammeyer, 1992] K.D. Kammeyer. *Nachrichtenübertragung.* Teubner, Stuttgart, 1992.

[Kay, 1988] S.M. Kay. *Modern Spectral Estimation.* Prentice Hall, Inc., Englewood Cliffs, New Jersey 07632, 1988.

[Kay, 1993] S.M. Kay. *Statistical Signal Processing.* Prentice Hall, Inc., Englewood Cliffs, New Jersey 07632, 1993.

[Kliewer, 1993] J. Kliewer. Analyse komplexer Signalstrukturen mit Hilfe von Wavelet–Transformationen. Diploma Thesis, Technical University of Hamburg–Harburg, Distributed Systems Department, 1993.

[Kronland-Martinet *et al.*, 1987] R. Kronland-Martinet, J. Morlet, and A. Grossmann. Analysis of Sound Patterns through Wavelet Transforms. *International Journal of Pattern Recognition and Artificial Intelligence*, 1:273–302, 1987.

[Kronland-Martinet, 1988] R. Kronland-Martinet. The Wavelet Transform for Analysis, Synthesis and Processing of Speech and Music Sounds. *Computer Music Journal*, 12(4):11–20, 1988.

[Loughlin *et al.*, 1992] P. J. Loughlin, J. W. Pitton, and L. E. Atlas. Proper Time–Frequency Energy Distributions and the Heisenberg Uncertainty Principle. In *IEEE Int. Symp. on Time–Frequency & Time–Scale Analysis*, pages 151–154, 1992.

[Lüke, 1985] H.D. Lüke. *Signalübertragung*. Springer–Verlag, 1985.

[Lyon, 1996] R.F. Lyon. The All–Pole Gammatone Filter and Auditory Models. Technical report, Apple Computer Inc., 1996.

[Maher, 1990] Robert C. Maher. Evaluation of a Method for Separating Digitized Duet Signals. *J. Audio Eng. Soc.*, 38(12):956–979, December 1990.

[Mallat and Zhong, 1992] S. Mallat and S. Zhong. Characterisation of Signals from Multiscale Edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(7):710–732, July 1992.

[Marple, 1987] S.L. Marple. *Digital Spectral Analysis with Applications*. Prentice Hall, Inc., Englewood Cliffs, New Jersey 07632, 1987.

[Martin and Padmanabhan, 1993] W.M. Martin and M. Padmanabhan. Using IIR Adaptive Filter Bank to Analyze Short Data Segments of Noisy Sinusoids. *IEEE Transactions on Signal Processing*, 41 (8):2583–2590, August 1993.

[Mayer-Lindenberg, 1995] F. Mayer-Lindenberg. A Parallel Computer Based on Simple DSP Modules. *Microprocessing and Microprogramming*, 41:301–314, 1995.

[Mayer-Lindenberg, 1997a] F. Mayer-Lindenberg. A Heterogeneous Parallel System Employing a Configurable Interconnection Network. In *Proceedings of the 9th IASTED International Conference – Parallel and Distributed Computing and Systems, Washington DC*, pages 13–16, 1997. ISBN 0-88986-240-0.

[Mayer-Lindenberg, 1997b] F. Mayer-Lindenberg. A Large Parallel System and its Programming Environment. In *Proceedings of the 8th Annual International Conference on Signal Processing Applications and Technology (ICSPAT), San Diego*, pages 1401–1405. Miller Freeman, 1997.

[McAulay and Quatieri, 1986] R.J. McAulay and T.F. Quatieri. Speech Analysis/Synthesis Based on a Sinusoidal Representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34 (4):744–754, August 1986.

[Meddis *et al.*, 1990] R. Meddis, M.J. Hewitt, and T.M. Shackleton. Non–Linearity in a Computational Model of the Response of the Basilar Membrane. In P. Dallos, C.D. Geisler, J.W. Matthews, M.A. Ruggero, and C.R. Steele, editors, *The Mechanics and Biophysics of Hearing*, Lecture Notes in Biomathematics, pages 403–410. Springer–Verlag, 1990.

[Mertins, 1996] A. Mertins. *Signaltheorie*. Teubner, Stuttgart, 1996.

[Meyer, 1994] Matthias Meyer. A Pilot Implementation of the Host–Engine Software Architecture for Parallel Digital Signal Processing. Diploma Thesis, Technical University of Hamburg–Harburg, Distributed Systems Department, 1994.

[Moorer, 1975] J.A. Moorer. *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer.* PhD thesis, Stanford University, 1975.

[Nakatani *et al.*, 1995a] T. Nakatani, T. Kawabata, and H.G. Okuno. A Computational Model of Sound Stream Segregation with Multi–Agent Paradigm. In *International Conference on Acoustics, Speech and Signal Processing*, pages 2671–2674, 1995.

[Nakatani *et al.*, 1995b] T. Nakatani, H.G. Okuno, and T. Kawabata. Residue–Driven Architecture for Computational Auditory Scene Analysis. In C.S. Mellish, editor, *14th International Joint Conference on Artificial Intelligence*, pages 165–172. Morgan Kaufman Publishers, 1995.

[Noll, 1997] P. Noll. MPEG Digital Audio Coding. *IEEE ASSP Magazine*, pages 59–81, September 1997.

[Oppenheim and Shafer, 1975] A.V. Oppenheim and R.W. Shafer. *Digital Signal Processing.* Prentice–Hall, 1975.

[Papoulis, 1962] A. Papoulis. *The Fourier Integral and its Applications.* McGraw–Hill, 1962.

[Papoulis, 1987] A. Papoulis. *Signal Analysis.* McGraw–Hill, 1987.

[Papoulis, 1990] A. Papoulis. *Probability and Statistics.* Prentice Hall, Inc., Englewood Cliffs, New Jersey 07632, 1990.

[Patterson *et al.*, 1992] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. Complex Sounds and Auditory Images. In Y. Cazals, L. Demany, and K. Horner, editors, *Auditory Physiology and Perception*, Advances in Biosciences, pages 429–443. Pergamon Press, 1992.

[Pearson, 1991] E.R.S. Pearson. *The Multiresolution Fourier Transform and its Application to Polyphonic Audio Analysis.* PhD thesis, University of Warwick, 1991.

[Picinbono, 1997] B. Picinbono. On Instantaneous Amplitude and Phase of Signals. *IEEE Transactions on Signal Processing*, 45 (3):552–560, March 1997.

[Popper, 1935] K. R. Popper. *Logik der Forschung.* Julius Springer Verlag, Wien, 1935.

[Ramalingam and Kumaresan, 1994] C.S. Ramalingam and R. Kumaresan. Voiced–Speech Analysis Based on the Residual Interfering Signal Canceler (RISC) Algorithm. In *ICASSP*, volume I, pages 473–476, 1994.

[Reekie and Meyer, 1994] John Reekie and Matthias Meyer. The Host–Engine Software Architecture for Parallel Digital Signal Processing. In *Proceedings of the Australasian Workshop on Parallel and Real–Time Systems*, Melbourne, Australia, July 1994. North–Holland.

[Rife and Boorstyn, 1974] D.C. Rife and R.R. Boorstyn. Single–Tone Parameter Estimation from Discrete–Time Observations. *IEEE Transactions on Information Theory*, 20 (5):591–598, September 1974.

[Rigden, 1976] J.S. Rigden. *Physics and the Sound of Music.* John Wiley & Sons, 1976. ISBN 0–471–02433–3.

[Rioul and Vetterli, 1991] O. Rioul and M. Vetterli. Wavelets and Signal Processing. *IEEE SP Magazine*, pages 14–38, October 1991.

[Scheirer, 1998] E.D. Scheirer. Using Musical Knowledge to Extract Expressive Performance Information from Audio Recordings. In D. F. Rosenthal and H. G. Okuno, editors, *Computational Auditory Scene Analysis*, chapter 24, pages 361–379. Erlbaum Associates, 1998. ISBN 0–8058–2283–6.

[Serra, 1989] X. Serra. *A System for Sound Analysis/ Transformation/ Synthesis Based on a Deterministic Plus Stochastic Decomposition.* PhD thesis, Stanford University, 1989.

[Shamma, 1993] S.A. Shamma. Common Principles in Auditory and Visual Processing. In P. et al. Rudomin, editor, *Neuroscience: From Neural Networks to Artificial Intelligence*, pages 189–205. Springer–Verlag, Berlin, Heidelberg, New York, 1993.

[Shensa, 1992] M.J. Shensa. The Discrete Wavelet Transform: Wedding the Á Trous and Mallat Algorithm. *IEEE Transactions on Signal Processing*, 40(10):2464–2482, October 1992.

[Simoncelli *et al.*, 1992] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, and D.J. Heeger. Shiftable Multiscale Transforms. *IEEE Transactions on Information Theory*, 38(2):587–607, March 1992.

[Slaney, 1988] M. Slaney. Lyon's Cochlear Model. Technical report, Apple Computer Inc., 1988.

[Slaney, 1993] M. Slaney. An Efficient Implementation of the Patterson–Holdsworth Auditory Filter Bank. Technical report, Apple Computer Inc., 1993.

[Slaney, 1998] M. Slaney. A Critique of Pure Audition. In D. F. Rosenthal and H. G. Okuno, editors, *Computational Auditory Scene Analysis*, chapter 3, pages 27–38. Erlbaum Associates, 1998. ISBN 0–8058–2283–6.

[Solbach and Wöhrmann, 1996] L. Solbach and R. Wöhrmann. Sound Onset Localization and Partial Tracking in Gaussian White Noise. In *Proceedings of the International Computer Music Conference, Hong Kong*, 1996. reprint available via ftp://ftp.ti6.tu–harburg.de/pub/paper.

[Solbach *et al.*, 1998] L. Solbach, R. Wöhrmann, and J. Kliewer. The Complex–Valued Continuous Wavelet Transform as a Preprocessor for Auditory Scene Analysis. In D. F. Rosenthal and H. G. Okuno, editors, *Computational Auditory Scene Analysis*, chapter 18, pages 273–291. Erlbaum Associates, 1998. ISBN 0–8058–2283–6.

[Stoica and Nehorai, 1988] P. Stoica and A. Nehorai. Statistical Analysis of Two Non–Linear Least–Squares Estimators of Sine Waves Parameters in the Colored Noise Case. *International Conference on Acoustics, Speech and Signal Processing*, pages 2408–2411, 1988.

[Tanguiane, 1993] A.S. Tanguiane. *Artificial Perception and Music Recognition (Lecture Notes in Artificial Intelligence, Vol 746).* Springer Verlag, 1993. ISBN 0387573941.

[Wan and Schneider, 1997] Ch. Wan and A.M. Schneider. Further Improvements in Digitizing Continuous–Time Filters. *IEEE Transactions on Signal Processing*, 45 (3):533–542, March 1997.

[Wang, 1994] A. Wang. *Instantaneous and Frequency Warped Signal Processing Techniques for Auditory Source Separation.* PhD thesis, Stanford University, 1994.

[Wilson *et al.*, 1992] R. Wilson, Calway A.D., and E.R.S. Pearson. A Generalized Wavelet Transform for Fourier Analysis: The Multiresolution Fourier Transform and Its Application to Image and Audio Signal Analysis. *IEEE Transactions on Information Theory*, 38(2):674–690, March 1992.

[Yen, 1987] N. Yen. Time and Frequency Representation of Acoustic Signals by Means of the Wigner Distribution Function: Implementation and Interpretation. *J. Acoust. Soc. Am.*, 81:1841–1850, June 1987.

[Zölzer, 1997] U. Zölzer. *Digital Audio Signal Processing.* Wiley, Chichester, 1997. ISBN 0-471-97226-6.

[Zwicker and Fastl, 1990] E. Zwicker and H. Fastl. *Psychoacoustics.* Springer, Berlin Heidelberg New York, 1990.