FRACTAL ADDITIVE SYNTHESIS: SPECTRAL MODELING OF SOUND FOR LOW RATE CODING OF QUALITY AUDIO

THÈSE N° 2711 (2003)

PRÉSENTÉE À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

SECTION DES SYSTÈMES DE COMMUNICATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Pietro POLOTTI

laurea in fisica, Univertità degli studi di Trieste, Italie
et de nationalité italienne

acceptée sur proposition du jury:

Dr G. Evangelista, Prof. M. Vetterli, directeurs de thèse
Dr M. Erne, rapporteur
Prof. M. Hasler, rapporteur
Prof. A. Sarti, rapporteur
Prof. U. Zoelzer, rapporteur

Lausanne, EPFL
2003

Contents

List of Figures

Abstract

Musical and audio signals in general form a major part of the large amount of data exchange taking place in our information-based society. Transmission of high quality audio signals through narrow-band channels, such as the Internet, requires refined methods for modeling and coding sound. The first important step is the development of new analysis techniques able to discriminate between sound components according to effective perceptual criteria. Our ultimate goal is to develop an optimal representation in a psychoacoustical sense, providing minimum rate and minimum "perceptual distortion" at the same time. One of the most challenging aspects of this task is the definition of a good model for the representation of the different components of sound. Musical and speech signals contain both deterministic and stochastic components. In voiced sounds the deterministic part provides the pitch and the global timbre: it is in a sense the fundamental structure of a sound and can be easily represented by means of a very restricted set of parameters. The stochastic part contains what we might call the "life of a sound", that is an ensemble of microfluctuations with respect to an electronic-like/non-evolving sound as well as noise due to the physical excitation system. The reproduction of the latter is of fundamental importance to perceive a sound like a natural one. We faced this challenge by developing a new sound analysis/synthesis method called Fractal Additive Synthesis (FAS).

The first step was the definition of a new class of wavelet transforms, namely the Harmonic-Band Wavelet Transform (HBWT). This transform is based on a cascade of Modified Discrete Cosine Transform (MDCT) and Wavelet Transforms (WT). By means of the HBWT, we are able to separate the stochastic from the deterministic components of sound and to treat them separately.

The second step was the definition of a model for the stochastic components. The spectra of voiced musical sound have non-zero energy in the sidebands of the spectral peaks. These sidebands contain information relative to the stochastic components. The effect of these components is that the waveform of what we call a pseudo-periodic signal, i.e. the stationary part of voiced sounds, changes slightly from period to period. Our work is based on the experimentally verified assumption that the energy distribution of a sideband of a voiced sound spectrum is mostly shaped like powers of the inverse of the distance from the closest partial. The power spectrum of these pseudo-periodic processes is then modeled by means of a superposition of modulated $1/f$ components, i.e., by means of what we define as a pseudo-periodic $1/f$–like process. The time-scale character of the wavelet transform is well adapted to the selfsimilar behavior of $1/f$ processes. The wavelet analysis of $1/f$ noise yields a set of very loosely correlated coefficients that in first approximation can be well modeled by white noise in the synthesis. The fractal properties of the $1/f$ noise also motivated

our choice of the name Fractal Additive Synthesis.

The next step was the definition of a model for the deterministic components of voiced sounds, consistent with the HBWT analysis/synthesis method. The model is from some point of view inspired by the sinusoidal models. The two models provide a complete method for the analysis and resynthesis of voiced sounds in the perspective of structured audio (SA) sound representations. For the stationary part of voiced sounds compression, ratios in the range of 10-15:1 are easily achievable.

Even better results in terms of data compression can be obtained by taking psychoacoustic criteria into consideration. A psychoacoustic based selection of perceptually relevant parameters was implemented and tested. Compression ratios of 20-30:1, depending on the musical instrument, were achieved.

An extension of the method based on a pitch-synchronous version of the HBWT with perfect reconstruction time-varying cosine-modulated filter banks was also studied. This makes the method able to handle, for instance, the slight pitch deviations or the vibrato of a musical tone or more relevant changes of pitch as in a glissando.

Finally, the method has been successfully extended to non-harmonic sounds by the introduction and definition of an optimization procedure for the design of non-perfect reconstruction cosine-modulated filter banks with inharmonic band subdivisions. These extensions make FAS more flexible and suitable to analyze, encode, process and resynthesize a large class of musical sounds.

The final result of this work is the development of a method for modeling in a flexible way both the stochastic and the deterministic parts of sounds at a very refined perceptual level and with a minimum amount of parameters controlling the synthesis process. In the context of SA the method provides a sound analysis/synthesis tool able to encode and to resynthesize sounds at low rate, while maintaining their natural timbre dynamics for high quality reproduction.

Sommario

I segnali audio, musicali e non, sono una parte consistente dell'enorme quantità di dati che vengono scambiati nella nostra società fondata sull'informazione. La trasmissione ad alta qualità di segnali attraverso canali di comunicazione a banda stretta quali *internet* richiede dei metodi sofisticati per la codifica e la modellazione del suono. Il primo passo importante è lo sviluppo di una nuova tecnica di analisi capace di discriminare tra le componenti del suono secondo dei criteri percettivi efficaci. La nostra meta finale è quella di sviluppare una rappresentazione ottimale dal punto di vista psicoacustico, che assicuri contemporaneamente un rate di dati minimo ed una minima "distorsione percettiva". Uno degli aspetti più impegnativi a questo fine è la definizione di un buon modello per la rappresentazione delle differenti componenti del suono. I segnali musicali e vocali contengono sia componenti deterministiche che stocastiche. Nei suoni intonati la parte deterministica fornisce l'altezza ed il timbro globale. E' in un certo senso la struttura fondamentale del suono e può essere facilmente rappresentata mediante un insieme molto ridotto di parametri. La parte stocastica contiene la "vita del suono", vale a dire un complesso di micro-fluttuazioni rispetto all'andamento di un suono elettronico privo di evoluzione, come pure il rumore dovuto all'eccitazione fisica del sistema. La riproduzione della componente stocastica è di fondamentale importanza al fine di percepire un suono come naturale. Abbiamo affrontato questo compito sviluppando un nuovo metodo di analisi e sintesi chiamato Sintesi Additiva Frattale (FAS).

Il primo passo è stato quello di definire una nuova classe di trasformate wavelet, vale a dire la *trasformata wavelet a bande armoniche* (HBWT). Questa trasformata è basata sulla successione di una *trasformata coseno discreta modificata* (MDCT) e di una *trasformata wavelet* (WT). Mediante la HBWT è possibile separare le componenti stocastiche di un suono da quelle deterministiche e ricostruirle perfettamente, in modo indipendente le une dalle altre.

Un secondo passo è stato quello di definire un modello per le componenti stocastiche del suono. Dal punto di vista del dominio della frequenza sappiamo che lo spettro dei suoni intonati ha energia non nulla nelle bande laterali dei picchi spettrali. Queste bande laterali contengono l'informazione relativa alle componenti stocastiche. L'effetto di queste componenti è che le forme d'onda di ciò che abbiamo chiamato un segnale pseudo-periodico, ovvero la parte stazionaria dei suoni intonati, cambia lentamente periodo per periodo. Il nostro lavoro è basato sull'assunto, sperimentalmente verificato, che la distribuzione di energia delle bande laterali dei suoni intonati ha per lo più un andamento uguale all'inverso di una potenza della distanza dalla parziale più vicina. Lo spettro di potenza di questi processi pseudo-periodici può essere pertanto modellato da una sovrapposizione di componenti $1/f$ modulate, ovvero tramite l'oggetto

matematico che abbiamo definito come processo pseudo-periodico di tipo $1/f$. Il carattere tempo-scala della trasformata wavelet è ben adattato alle proprietà di autosomiglianza dei processi $1/f$. L'analisi wavelet del rumore $1/f$ fornisce un insieme di coefficienti debolmente correlati tra loro, che nella sintesi possono essere modellati, in prima approssimazione, da rumore bianco. Le proprietà frattali del rumore $1/f$ sono anche il motivo della nostra scelta del nome Sintesi Additiva Frattale.

Un passo successivo è stato quello di definire un modello per le componenti deterministiche dei suoni intonati, consistente con il metodo di analisi/sintesi delle HBWT. Il modello è ispirato da un certo punto di vista ai modelli sinusoidali. I due modelli introdotti per le componenti stocastiche e deterministiche del suono nel loro insieme forniscono un metodo completo per l'analisi e la resintesi dei suoni intonati nell'ottica della rappresentazione dei suoni proposta nella definizione dell'Audio Strutturato (SA). Per la parte stazionaria dei suoni intonati rapporti di compressione dell'ordine di 10-15:1 sono facilmente ottenibili.

Risultati anche migliori in termini di compressione dati possono essere ottenuti prendendo in considerazione criteri psicoacustici. Una selezione di parametri rilevanti basata su criteri psicoacustici è stata implementata e valutata sperimentalmente. In questo modo si possono ottenere rapporti di compressione dell'ordine di 20-30:1, a seconda dello strumento musicale.

E' stata realizzata un'estensione del metodo basata su una versione pitch-sincrona delle HBWT, ottenuta mediante banchi di filtri coseno modulati tempo-varianti e a ricostruzione perfetta. Ciò rende il metodo in grado di trattare, per esempio, le piccole deviazioni di intonazione, il vibrato di una nota musicale o anche più rilevanti cambiamenti del pitch come in un glissando.

Infine, il metodo è stato esteso a suoni non armonici mediante l'introduzione e lo sviluppo di una procedura di ottimizzazione per il disegno di banchi di filtri coseno modulati a ricostruzione non perfetta, con una suddivisione in bande non armoniche. Queste estensioni rendono la FAS più flessibile ed adatta per l'analisi, la codifica, l'elaborazione e la resintesi di una vasta classe di suoni musicali.

Il risultato finale di questo lavoro è lo sviluppo di un metodo per modellare in modo flessibile sia le parti stocastiche che quelle deterministiche del suono ad un livello percettivo molto raffinato e con una minima quantità di parametri controllanti il processo di sintesi. Nel contesto dell'audio strutturato il metodo fornisce una tecnica di analisi e sintesi del suono capace di rappresentare parametricamente e di resintetizzare i suoni a basso rate, mantenendo la loro dinamica timbrica naturale al fine di ottenere una riproduzione di alta qualità.

Chapter 1

Introduction

In this introduction we illustrate the motivations that led to this work. We also try to provide an historical background to the ideas on which this work lies. For this purpose we propose some considerations about reciprocal influences between technological research and aesthetic research in particular for what concerns signal processing applied to musical sounds and the electroacoustic music experience of the last 50 years, respectively. Timbre is one of the most important constructive element of the music of the 20^{th} century. The main idea is that this fact is intimately related to the nowadays need for a timbral high quality reproduction in any kind and gendre of music.

In order to face the increasing demand of audio products both in terms of quantity and of timbral quality, we need to investigate deeper in the domains of signal processing and psychoucoustics at the same time, in order to find new representations for sounds, which are effective both in terms of audio quality reproduction and of data compression. As discussed in Section 1.1.2, our idea is that this can be done without indulging in solutions as those proposed by higher-level coding, which would reduce timbre back again to the role it had in 19^{th} century and earlier. Fractal Additive Synthesis (FAS) technique claims to be one of the possible method for low rate coding of high quality audio.

1.1 Towards a complete and intuitive spectral model of sounds

One of the main disciplines where timbre research and timbre perception investigation is carried out is electroacoustic music. One of the electroacoustic music composers main task consists of looking for new timbres and studying what are the perceptual relationships with other new or already known timbres. By combining and composing these new sounds in a musical sense, it is often possible to discover new relationships and meanings and to push our perceptual and cognitive border and experience of timbre beyond the already known.

Timbre is an incredibly complex physical and perceptual phenomenon. Several attempts to produce a convincing classification of timbre failed to find a satisfying and exhaustive system [70] [33] [12] [60]. Especially when noisy sounds are taken into account, classifying timbres and defining perceptual distances between timbres in a rigorous fashion is an extremely hard task. The history of

timbre classification for the analysis of electroascoustic music works began with Pierre Schaeffer in 1967 [76]. His main idea was to consider timbres as 'sound objects' ('l'objet sonore'), indivisible and unique entities. His classification is for sure incomplete and possibly not precise. However, it is interesting to notice how the problem stated by schaeffer is the same one that nowadays people involved in the MPEG-7[1] standard have to face in order to define proper timbre description tools for audio retrievial. In this context noise assumes a fundamental role as a component of any timbre, even of tuned sounds.

In the next section we sketch a short history of timbre with the main purpose of tracing the perceptual and aesthetical discovery of noise. The reason for this is that the FAS starting goal was the achievement of an effective and convincing model for the noisy components of sounds.

1.1.1 Timbre: an historical perspective

From Pythagoras through the Middle Ages, when music was part of the *Quadrivium*, music was included in the scientific part of human knowledge, together with arithmetic, geometry and astronomy. J. S. Bach is one of the most well-known example of what is considered the use of arithmetic in music. In that vision of music, timbre was considered as a tool and never as a goal until the very end of the 18^{th} century. As an extreme statement one could say that the different instruments had the purpose only of making the polyphony and the harmonic syntax as intelligible as possible. Noise entered music only through drums –recalling war– or theater effects as the simulations of thunders or other atmospheric events.

The experience of the big romantic orchestra of the 19^{th} century, finding its climax in Hector Berlioz and Richard Strauss' Treatise on Instrumentation [5], opens the doors to a new conception of timbre. The timbral possibilities offered by the instrument combinations of the big orchestra raise the role of timbre to that of a goal and no longer a mere tool. This concept became extreme in the Second Viennese School in the first half of the 20^{th} century. Schoenberg's and Webern's compositions developed what Schoenberg himself called the *Klangfarbenmelodie* (sound-colour-melody), i.e. the idea of using musical instrument timbres as the main syntactical elements in the compositional process.

Also during the first half of the 20^{th} century, the Italian futurist Luigi Russolo devoted his life to building instruments for generating and playing noises. He performed various concerts with his instruments, mainly in Paris. Stravinsky had planned to use Russolo's main creation: the *intonarumori* (noise-player) in *Les Noces*, but gave up for technical reasons. Edgar Varèse was also very interested in Russolo's work and did a presentation of Russolo's new *Rumorarmonio* (Noise-harmonium) instrument in 1929. Many other composers were attracted by the mechanical noise of the new sound environment of industrial towns. An example is Honneger's symphony *Pacific 231*, written in 1923, a work depicting a locomotive (the Pacific 231 of the trans-American railway) and featuring industrial sonorities. Edgard Varèse himself realized Russolo's dream of a compo-

[1] MPEG-7 is an ISO/IEC (International Organization for Standardization) standard developed by MPEG (Moving Picture Experts Group). MPEG-7, formally named "Multimedia Content Description Interface", is the description of multimedia content data that supports some degree of interpretation of the information meaning, which can be passed onto, or accessed by, a device or a computer code.

sition based only on noises, by writing *Ionisation* for 41 percussion instruments and two sirens (1931) [90]. By means of *Ionisation* Varèse freed music not only from harmony, as did Schoenberg, but also from pitch. *Ionisation* can be seen as an anticipation of the advent of electronic music. Some years later, John Cage used for the first time the loudspeaker as a "performer" in a concert: *Imaginary Landscape no. 1* for two variable-speed turntables, piano strings and cymbal (1939) was the first piece of what he called electric music.

Electroacoustic music as an autonomous gendre started with the French experience denominated by *Musique Concrète* in the Studios of Radio France in Paris with the works of Pierre Schaffer [77]. The main idea of the *Musique Concrète* is to use any recorded sound, typically noise, to generate a musical language and to create musical compositions. The products of the *Musique Concrète* of the fifties and of the following decades as well are music pieces recorded on tapes.

Later on contemporary music produced what is usually called *live electronics* [97]. The idea of *live electronics* is essentially to augment the possibilities of traditional instruments by means of real-time digital processing of sounds. Once more Cage produced the first pre-digital examples of *live electronics*: *Cartridge Music* for contact microphones and phonograph cartridges attached to various household objects (1960). The use of real-time processing of sound in electroacoustic music was consolidated by Karlheinz Stockhausen, who wrote in 1964 *Mixtur* for 5 orchestra groups, 4 sine-wave generators, 4 ring-modulators and *Mikrophonie I*, for 6 performers: 2 tam-tam players, 2 microphone "players", 2 filter "players". A new great period for *Live electronics* based on digital sound processing started years later in the Südwestrundfunk (SWR) Experimentalstudio in Freiburg with *Das Atmende Klarsein* [54] (1981), for bass flute, small chorus and live electronics by Luigi Nono. This was the first work of three years of experimentation culminating in *Prometeo, la tragedia dell'ascolto* (1984-85) [55] [41], [35]. One of the principal targets of Nono's research is the microscopic dimension of the timbre space. The noise is the interesting and uncontaminated part of instrumental sounds, where to look for new possibilities of expressivity. *Prometeo* is not only a research into timbre. In this work, Nono realizes also the ideal of a total artwork, joining music, poetry, use of lights, architectural acoustics and spatialized sound diffusion, all augmented by the employment of digital technology. The people who collaborated with Nono were the architect Renzo Piano, the conductor Claudio Abbado, the writer and philosopher Massimo Cacciari, the sound engineers Alvise Vidolin and Hans Peter Haller and the painter Emilio Vedova (see Figure 1.1).

What is the effect of all of this experience nowadays? Are the concept of "sound" as "timbrical print" of any pop ensemble or the huge widening of timbres employed in any gender of music, a consequence of this growth of timbrical sensitiveness? In our opinion the answer is yes. Often the success of a pop group is due to a "guessed sound" rather than to any melodic and/or harmonic or, even less, any compositional invention. In this sense the electroacoustic musician community can be thought as the avant-garde of what are the main direction of listening in the future and of what is the aesthetical research concerning a widening of timbre vocabulary.

Another example of anticipation coming from the electroacoustic music experience is the use of space in music and listening. Very early examples of this aspect are the *Gesang der Juenglinge* for tape solo (1956) by Karlheinz Stock-

Figure 1.1: Poster of the first representation of *Prometeo* in San Lorenzo church in Venice.

hausen and *Artikulation* for tape solo (1958) by György Ligeti, both produced in the *Westdeutscher Rundfunk* (WDR) studios in Cologne [43]. In this pieces, Stockhausen and Ligeti use diffusion systems composed by five and four loudspeakers, respectively, placed around the audience. The movements of sound from one loudspeaker to the other are notated in the score and become a compositional element. Today, surround systems such as the 5.1 system for surrounding effect in movies are a standard. Nowadays, the investigation into spatial perception of sound done by electroacoustic musicians goes much further, towards directions that could be interesting for virtual reality research in terms of synthesis of sound localization.

We believe that new aesthetical requirements and investigations can lead to a different concept of music potentially capable of influencing people's way of listening. In particular they determine an expectation for higher sound quality with a consequent need for enhancement of technical requirements for music representation, transmission and reproduction. It will be clear that the main point of FAS is to provide a convincing model for the noisy components of sounds. The goal is to provide high-quality reproduction, in which the resynthesized sounds maintain their naturalness. The noisy components turn out to be extremely important in order to preserve naturalness even in synthetic sounds. FAS provides an intelligent/signal-adapted spectral representation of sound where both high-quality timbre reproduction and low data rates are achieved.

1.1.2 Data compression. Perceptual versus symbolic coding

MPEG-4[2] includes two main approaches to audio music coding: the synthesis language called SAOL (Structured Audio Orchestra Language) for synthetic sounds and the Parametric Audio Coding (PAC) for natural sounds.

The SAOL language is used to define an "orchestra" made up of "instruments", which create and process control data. An instrument is a small network of signal processing primitives that might emulate specific sounds such as those of a natural acoustic instrument. The instruments are downloaded in the bitstream. MPEG-4 does not standardize "a single method" of synthesis. Any current or future sound-synthesis method can be described in SAOL, as wavetable, FM, additive, physical-modeling, and granular synthesis. Control of the synthesis is accomplished by downloading "scores" in the bitstream. A score is a time-sequenced set of commands that calls various instruments at specific times. The score description is downloaded in a language called Structured Audio Score Language (SASL). The SASL allows the musicians to gain finer control over the final synthesized sound. For synthesis processes that do not require a fine control, the established Musical Instrument Digital Interface (MIDI) protocol may also be used.

In the early 80's, when computer music was starting, computer synthetic reproductions of baroque music had a certain short-lived success. We might go so far as to say that Bach would be satisfied by a MIDI coding of his music,

[2]MPEG-4 is an ISO/IEC (International Organization for Standardization) standard developed by MPEG (Moving Picture Experts Group). MPEG-4 provides standardized technological elements, enabling the integration of the production, distribution and content access paradigms of fields as digital television, interactive graphics applications and interactive multimedia in the world wide web. More information about MPEG-4 can be found at MPEG's home page: http://mpeg.telecomitalialab.com.

where sounds are reduced to pitch, (rough) dynamics, duration and timbre (i.e. synthetic or sampled instruments chosen within a preset bank). Nevertheless it is not possible to put timbre back to the age of Bach or Mozart by means of a more or less sophisticated MIDI-like coding technique. It is symptomatic to observe that musicians, after an initial enthusiasm, are mainly escaping synthetic sounds in favor of digital processing of real-life sounds, whose timbral richness cannot be renounced.

PAC uses the Harmonic and Individual Lines plus Noise (HILN) technique to code audio signals at bit rates of 4 kbit/s and above using a parametric representation of the audio signal. The basic idea of this technique is similar but simpler to FAS. The sound is decomposed into audio objects, which are represented by model parameters. In the HILN coder models for sinusoids, harmonic tones, and noise are provided. A perceptual model is employed to select those objects that are most important for the perceived quality of the signal. For example, the frequency and amplitude parameters are quantized according to the Just Noticeable Differences (JND).

Like HILN, FAS is a parametric audio coding technique in a strict perceptual sense. No alteration of the perceptual content of the sound is previewed or allowed. The indivisible 'objet sonore' (sound object) of Schaeffer, certainly not representable by means of MIDI-like codes, is preserved. No orchestra piece and no solo instrument piece as well can be coded in a satisfactory way by means of a SAOL-SASL approach: roughly speaking, Miles Davis' trumpet is unique and cannot be reduced to a synthetic trumpet played by a sequencer. Our task is to find a coding approach looking into the intimate structure of the signal and into the intimate functioning of our auditory system. This has to be the result of the cooperation of signal processing and psychoacoustics. Any backward-sighted score-like coding is constrained by the unavoidable and culturally-related limitations of any music notation system.

1.1.3 Sound synthesis and high level parameter representation

In the early 50's in the *Westdeutscher Rundfunk* (WDR) studios in Cologne, Karlheinz Stockhausen and other composers started producing a new kind of music synthetically generated by means of analog instruments as oscillators and noise or impulse generators. The music composed in the WDR studios reflected the ideal of the total control belonging to part of the contemporary music of that time.

Later came digital music. The father of digital or computer music is Max Mathews [50] [51], director the *Acoustical and Behavioral Research Center* at *Bell Laboratories* from 1962 to 1985. About the birth of computer music Mathews writes: "Computer performance of music was born in 1957 when an IBM 704 in NYC played a 17 second composition on the Music I program, which I wrote. The timbres and notes were not inspiring, but the technical breakthrough is still reverberating. Music I led me to Music II through V...". Music V is the first program for software synthesis widely employed by electroacoustic musicians until the early 90's, when it was substituted by the more user-friendly Csound.

However, the real great success of sound synthesis came with Frequency Modulation (FM) [10]. FM has also been the greatest commercial success of electronic music, well exploited by Yamaha. Its synthesizers are still in the

house of any musician who has been concerned with electronic music during the 80's and 90's. The FM technique was invented at Stanford University in the late 60's by John Chowning. The main reason for such a great success is certainly the extreme effectiveness of its control interface both in terms of simplicity and perceptual intelligibility. Only two intuitive parameters are used: one controls the spectral density, the other the harmonicity/inharmonicity factor. The drawback of this simplicity is of course that no detailed control on the timbre is possible. FM is a kind of "marvelous kaleidoscope" where the beautiful pictures that are generated cannot be adapted at one's own will; they have just to be taken as they are. Nevertheless, the lesson coming from FM concerning the success of its perceptual metaphor is extremely important.

The importance of an appealing interface is also a goal of our method. In the last chapter of this thesis we present a real-time implementation of FAS, which is equipped with a very intuitive interface.

We will spend only a few more words about another powerful family of synthesis algorithms, i.e. physical models. In physical modeling the perceptual meaning of the parameters and the naturalness of sound results are also very well achieved objectives. On the other side, they are anchored to the "physicality" of the model in the same way as the traditional luthery is. From a coding point of view, it would be necessary to provide a physical model of each particular instrument (not just a generic clarinet, for example) and of the player behavior as well.

1.2 Outline of the thesis

The structure of the thesis is the following. Chapter 2 provides an overview of the whole method without entering into technical details. Chapter 3 goes into some details by introducing the mathematical formalism of the pseudo-periodic $1/f$-like model. The discussion concerns the fractal properties of the $1/f$ noise and its relationships with the time-scale representation provided by the wavelet transform. In Chapter 4, we show how it is possible to reduce the pure pseudo-periodic $1/f$ spectral model to a more concrete and perceptually oriented modeling technique, which is FAS. FAS provides an effective coding method for high quality audio at low rate by means of a set of perceptually meaningful parameters. Chapter 5 introduces some important extensions of the method in order to include a larger class of sounds and instruments that can be encoded by means of FAS. Finally, in Chapter 6, we draw our conclusions.

Chapter 2

Fractal Additive Synthesis, a Method for Sound Analysis and Synthesis

One of the most challenging aspects of sound analysis and representation is the definition of a useful model for the noisy part of sounds. We need a faithful representation of those components of sound whose spectra lie outside the frequency support of the partials. We subdivide a sound into its deterministic and stochastic components. The deterministic part of sounds provides the global timbre of a sound; it is in a sense the fundamental structure of the sound. The stochastic part contains the "life of a sound", that is all the microfluctuations with respect to an electronic-like/purely deterministic sound including the noise due to the physical excitation system. The main idea of our method is that these microfluctuations with respect to a pure deterministic behavior can be reconstructed from the power spectrum.

This chapter provides a complete overview of the whole Fractal Additive Syntehsis (FAS) method, leaving the formalism and the details to the following chapters. The starting point is the experimental evidence that the spectra of voiced sounds, i.e. sounds with a well defined pitch are formed by harmonic partial peaks, whose sidebands have an approximately $1/f$ behavior around the peak itself (see Figure 2.1), i.e. a pseudo-periodic $1/f$-like behavior. We then define a well-suited analysis and resynthesis method based on the Harmonic-Band Wavelet Transforms (HBWT). Thanks to the mathematical properties of the HBWT, the synthesis of signals with pseudo-periodic $1/f$ –like power spectra is straightforward. These spectra are very good approximations of those of real-life voiced sounds. In first approximation, the only thing we need to do is to control the energies of white noise coefficients, according to a restricted number of parameters derived from the analysis of real sounds.

The following step is a parametric model for representing the synthesis coefficients in case of voiced sounds with stable pitch. Finally, we extend the method to the case of voiced sounds with variable pitch and to the case of sounds with non-harmonic spectra, such as percussive sounds.

The first type of parameters come from a further insight into the HWBT analysis, which reveals the existence of a small but non-zero correlation be-

tween the coefficients. An autoregressive (AR) analysis and resynthesis model, employing white noise as excitation and reproducing the above-mentioned loose correlation, generalizes the white noise coefficient model. We also take into account the scale-dependent time evolution of the resynthesis parameters. This provides an efficient parametric model for the stochastic part of sound. Finally, a model for the harmonic components is developed, inspired by the sinusoidal modeling techniques. The idea is to model a complex version of the coefficients corresponding to the deterministic components by means of a polynomial interpolation of their amplitudes and phases.

The already mentioned FAS extension to the case of voiced sounds with time-varying pitch provides a model able to deal with the slight detunings that occur in natural voiced-sounds, as well as to deal, for instance, with vibrato effects. In order to do this we design a perfect reconstruction time-varying filter bank whose number of channels is tuned to the variation of the pitch of the analyzed voiced-sound. This filter design technique and its adaptation to the fractal additive scheme are the main subjects of the first part of the fifth chapter. Finally the extension of the technique to inharmonic sounds is illustrated. A inharmonic filter design procedure is defined in order to apply the same principles of the method for voiced sounds to sounds produced by percussion instruments as gongs, tympani or tubular bells, as well as to sounds with expanded quasi-harmonic spectrum as piano sounds.

This method as a whole can be seen both as a data compression technique and as a musical tool for sound synthesis and processing able to provide synthetic sounds with a natural timbre dynamic.

In Section 2.1 and 2.2 we briefly review traditional wavelets, in order to introduce the Harmonic-Band Wavelets and the pseudo-periodic $1/f$-like model, respectively. In Section 2.3 and 2.4, we give a complete overview of the FAS method from the methodological and experimental point of view.

2.1 Wavelets and harmonic-band wavelets

The wavelet transform has been widely employed in sound analysis and processing [39], [17], [21], [40], [42]. The main idea of the transform is to provide a bidimensional representation of signals in order to overcome the intrinsic limitations of the Fourier transform [69], [13], [37], [11], [46] ,[47]. The wavelet transform or expansion provides in fact a time-scale representation of digital signals. Other bidimensional signal representations have been proposed, starting with the time-frequency representation provided by the Short Time Fourier Transform (STFT), the phase vocoder and the Wigner distribution [59].

Since time and frequency are conjugate variables, they obey the uncertainty principle:

$$\Delta\omega \cdot \Delta t \geq 1/2 \tag{2.1}$$

This means that it is not possible to have simultaneously an arbitrarily high resolution both in time and frequency. The higher the resolution of our signal representation in the time-domain the lower the resolution in the frequency-scale domain and vice-versa. Let us make precise what we mean by "scale". This concept is fundamental in wavelet theory and is intrinsically related to the human perceptual system, not only for what concerns the auditory system but also for the visual system. Scaling is represented by contraction and expansion

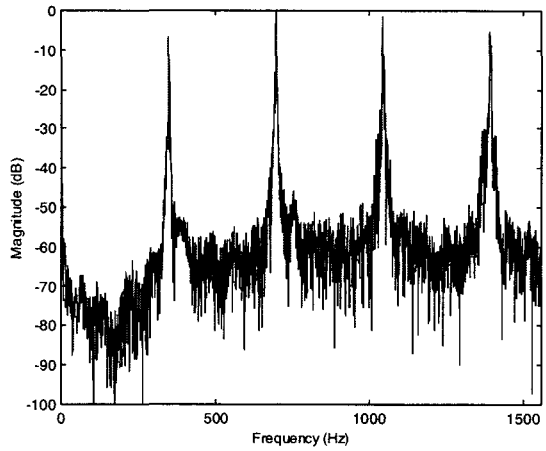


Figure 2.1: Magnitude Fourier Transform (FT) of a trumpet

in the domain of the function, which correspond to expansion and contraction, respectively, in the frequency-domain.

The wavelet transform leads to a multi resolution analysis (MRA), where the indetermination product $\Delta\omega \cdot \Delta t$ of the analysis functions is invariant by scaling and time shift. This is a peculiar type of time-frequency analysis. In the wavelet case, the time-frequency plane is subdivided into rectangles, corresponding to a logarithmic segmentation of the frequency axis and of the time intervals at various scales. In other words, the wavelet-based representation provides a time and frequency domain subdivision scheme imitating the "logarithmic" human hearing system by means of a logarithmic tiling of the time-frequency plane (see Figure 2.2). According to the representation of Figure 2.2, we have a more detailed frequency localization in the low frequency area but a coarser time resolution, while a sharper time resolution but a coarser frequency resolution is achieved in the high frequency area.

2.1.1 WT and multiresolution analysis

The wavelet representation is obtained by projecting a signal over the set of wavelets $\psi_{n,m}(t)$, where the first index n represents the scale and the second index m represents time-shift. When we compute the wavelet transform of a sound, we actually stop our analysis at a finite scale level. At each scale, the residual signal corresponds to the residual low-pass band. In other words, we look at the signal at different "zoom" scales, corresponding to different frequency contents, i.e. to different detail levels. We can associate an approximation subspace to each resolution scale, i.e. to each residual low-pass frequency band (see Figure 2.2):

$$\{V_n\}_{n \in N} \qquad with\ \mathbf{V}_n \subset \mathbf{V}_{n-1} \tag{2.2}$$

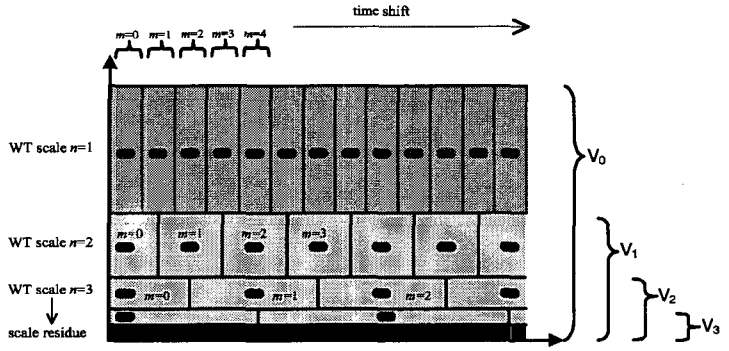


Figure 2.2: Time-Fequency Plane Tesselation of the ordinary wavelet transform. Each dot corresponds to a wavelet coefficient. At each scale n the sampling rate is divided by two.

The wavelet components are nothing but the difference between the components of two successive subspaces.

In Figure 2.3 we show an example of wavelet at three different scales, both in the time domain and in the frequency domain. The relationship between time and frequency resolution is evident. Each wavelet can be thought of as a "grain" [27], [71], representing a certain frequency band "presence" (energy) at a specific time-interval. These grains, multiplied by the analysis coefficients and overlap-added at the same sampling rate of the original sound, return the whole original signal.

We observe that we are dealing with the particular case of dyadic wavelets. Thus, since the frequency band of each residual component decreases by a factor two, we can decrease the sampling rate (downsample) by a factor of 2 as well. This is the reason for the denomination "multirate analysis". The coefficients of the different wavelet scale levels have a different sampling rate. It is like having a system whose parts work at different speeds, in order to optimize the effort. The "effort" in this case corresponds to the amount of data. This corresponds also to the idea of "critical sampling", i.e. the idea of employing the minimum possible number of samples in order to be able to perfectly reconstruct the original signal. In this way, the total amount of data of the wavelet representation remains approximately the same as that of the time-domain sound representation.

The dyadic wavelets can be simply generated by a pair of Quadrature Mirror Filters (QMF) $H_0(\omega)$, $H_1(\omega)$. These filters satisfy the condition of power complementarity (see Figure 2.4):

$$|H_0(\omega)|^2 + |H_1(\omega)|^2 = 2$$

This is one of the conditions, actually the most intuitive one, which grant the perfect reconstruction of the system. In this type of analysis we are always able to backtrack, i.e. the transform operates under perfect reconstruction constraint. In other words, a time-limited signal can be reconstructed by means of

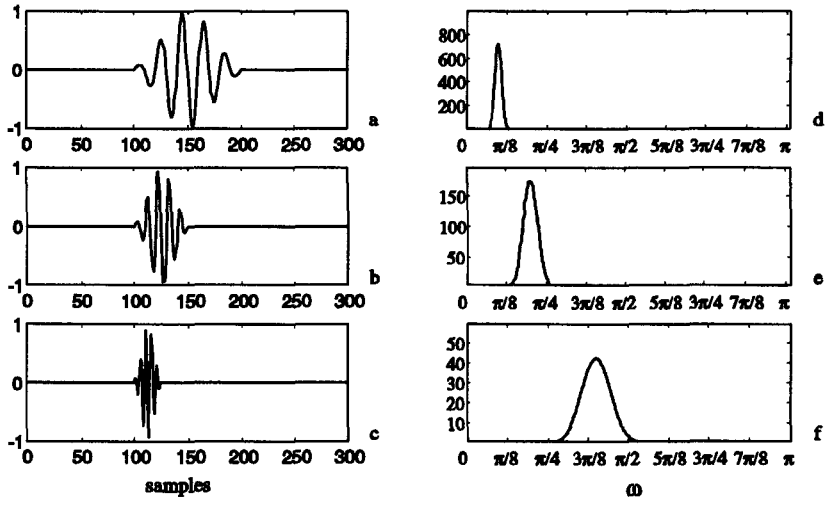


Figure 2.3: A wavelet function at three different scales in the time-domain (a , b and c) and in the frequency-domain (d, e and f).

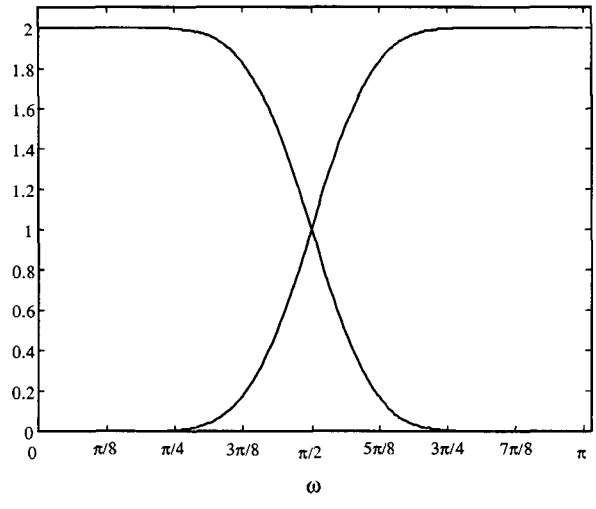


Figure 2.4: Magnitude FT of a pair of Quadrature Mirror Filters (QMF).

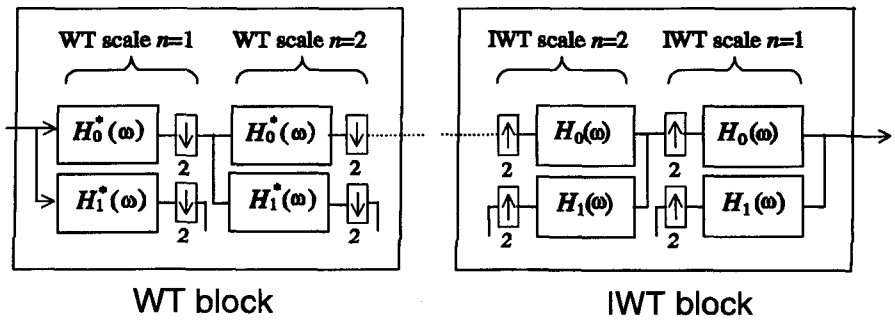


Figure 2.5: Scheme of a two-channel critically sampled filter-bank, implementing the wavelet transform. The filters H_0 are low-pass QMF filters implementing the scale function. They perform multiresolution analysis, i.e. they separate the subspaces V_n. At each scale level n we obtain the scale residue of the signal, i.e., the projection of the signal onto the subspace V_n. The filters H_1 are high-pass QMF filters implementing the wavelet projection onto the spaces $W_n = V_{n-1} - V_n$ at the different scale levels.

a discrete and finite grid of wavelet transform values, i.e. the dots in Figure 2.2. We are able to disassemble a signal into different resolution components and to assemble it back. The challenge is to find decompositions that are perceptually and musically meaningful. This will be achieved by means of the harmonic-band wavelets. The wavelet decomposition of the signal can be computed by means of the filter bank of Figure 4. We need only two prototype filters. The boxes containing a downward and an upward arrow stand for downsampling and upsampling, respectively. The filter $H_1(\omega)$ with impulse response $h_1(n)$ implements the projection of the signal onto the spaces $W_n = V_{n-1} - V_n$ at different scales n, while the filter $H_0(\omega)$ with impulse response $h_0(n)$ implements the projection of the signal onto the spaces V_n.

Another intuitive image of the wavelet transform consists in looking at the sequence of the scale coefficients as a "time-shrunk" (downsampled) version of the low-pass filtered signal. As the scale n increases, the residue represents the "trend" of the signal. According to this perspective, signal decomposition of sounds in wavelet "grains" can also be thought of as the separation of the global behavior from the fluctuations with respect to this global behavior at different scale levels. At a certain scale, this global behavior becomes the local mean of the signal. This point of view will be important in the discussion of the Harmonic-Band Wavelets, i.e. a frequency-periodic version of the ordinary wavelets. The frequency resolution of the dyadic wavelet transform is one octave. This introduces severe limitations in their use for musical sound processing. Improvements of the frequency resolution up to arbitrary bandwidth tiling by means of frequency warping have been proposed [21], [20], [38], [7]. Nevertheless the results of that work do not satisfy our requirements. For voiced sounds, one would like to tune the characteristics of the transform to the pitch of the signal. A first approach to the problem has been introduced in [19], [18] with the Pitch-Synchronous Wavelet Transform (PSWT). In that case the "global behavior" of a voiced sound, i.e. a sound with a detectable pitch

is its average period. The PSWT components represent the fluctuations of the sound with respect to its average period. By introducing the Harmonic-Band Wavelet Transform (HBWT) we gain considerable flexibility. With respect to the PSWT, the HBWT allows one to analyze and resynthesize each harmonic separately. With respect to ordinary wavelets, the PSWT and the HBWT provide a much more meaningful representation of voiced sounds.

The HBWT as well as the PSWT realize a periodic version of the frequency-domain subdivision of ordinary wavelets (see Figure 2.6). This is obtained by means of the modulation and demodulation scheme described in the next chapter. Here it is sufficient to know that it is possible to tune the frequency domain subdivision of the HBWT to the pitch of any given voiced sound by choosing the proper number of bands, i.e. of channels of the HBWT filter bank. More precisely, each frequency range corresponding to a single harmonic component and its two $1/f$-like spectral sidebands is analyzed by means of two channels of a HBWT (see Figure 2.6). In the ordinary wavelet representation the higher scales (corresponding to the low frequencies) represent the slow changes of a signal with respect to the "average" of the signal (0 in the case of an audio signal), i.e. with respect to a constant. The lower scales (high frequencies) represent the changes with respect to the local mean at different rates. In the HBWT representation of voiced sounds the "local means" are the average waveforms of each of the harmonics, while the lower scales (the bands away from the harmonics) contain the information concerning the fluctuations with respect to the average waveforms at different rates. In this way we are able to separate the harmonic part of voiced sounds from the noisy components containing the "timbre dynamics" of the sound.

2.1.2 Harmonic-Band Wavelet Transform (HBWT)

The HBWT are implemented by means of a P-channel filter bank, followed by P wavelet filter banks (see Figure 2.7). The number of channels P is tuned to the average pitch of the voiced sound to be analyzed and/or synthesized. Each filter of the P-channel filter bank is a bandpass filter separating a single sideband of the corresponding harmonics. The outputs of the P-channel filter bank are then downsampled P times and separately wavelet transformed. The P-order downsampling is possible since the bandwidth of each output is π/P. The resulting coefficients at the multirate time shift m are represented by the dots of Figure 2.6, channel by channel (i.e. sideband by sideband) and wavelet scale by wavelet scale (i.e. subband by subband). In Figure 2.6 the sidebands correspond to the index p, while the subbands correspond to the index n of Figure 2.3. In the synthesis structure (see Figure 2.8) each inverse wavelet transform (IWT) reconstructs the single sideband. The subsequent P-order upsampling shrinks back the spectrum of the reconstructed signal to the proper sideband bandwidth, providing a periodic spectral version of the sideband. Finally the P-channel filter bank selects the proper sideband frequency range for each channel. The sum of the outputs is the perfectly reconstructed signal.

In order to implement the P-channel filter bank, we employed a Type IV cosine modulated basis or Modified Discrete Cosine Transform (MDCT) [53], [89].

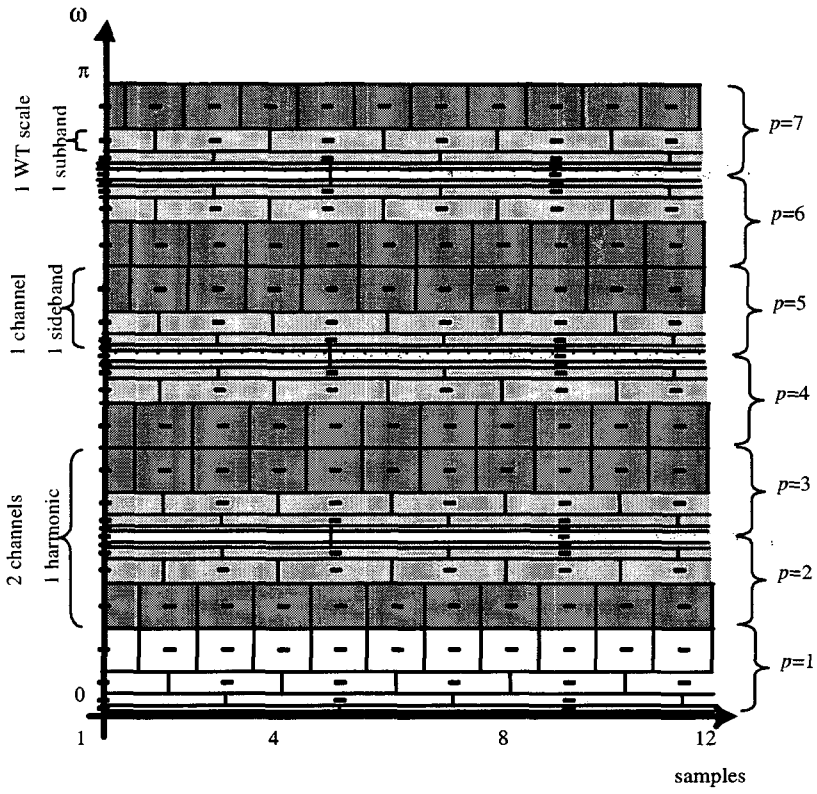


Figure 2.6: Time-Fequency Plane Tesselation of the HBWT. The tesselation is a frequency-periodic version of Figure 1. The number of periods corresponds to the number of band-pass filters (that is, the number of channels) trapping a single sideband of one harmonic denoted by the index p.

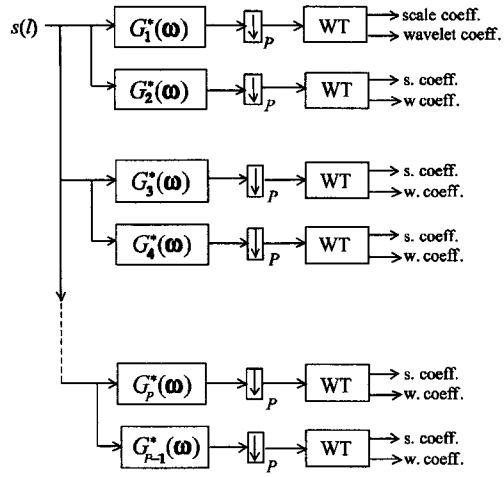


Figure 2.7: HBWT analysis filter bank. The filters G_p implement the MDCT, while the WT blocks represent a wavelet transformation implemented as in Figure 2.5.

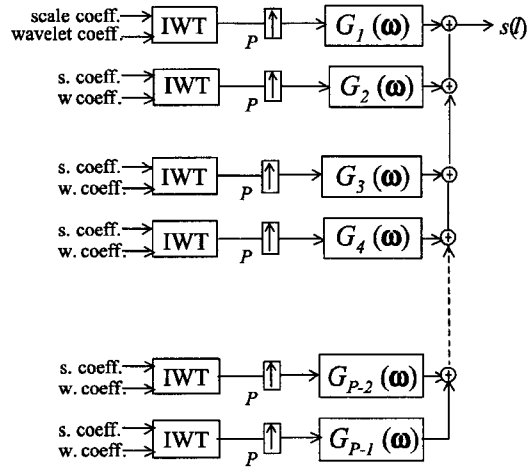


Figure 2.8: HBWT synthesis filter bank. The same notation of the previous figure holds. The blocks IWT represent the inverse WT.

The DTFT of the HBWT is given by:

$$\Theta_{n,m,p}(\omega) = \Psi_{n,m}(P\omega)G_p(\omega)$$

where $G_p(\omega)$ is the frequency response of the p^{th} filter of the P-channel filter bank, $\Psi_{n,m}(P\omega)$ is the upsampled version of the Fourier transform of the wavelet function at scale n and time-shift m (see Figure 2.2) and $\Theta_{n,m,p}(\omega)$ is the Fourier transform of the HBWT p^{th} sideband at n^{th} scale (n^{th} subband) and time shift m (see Figure 2.6). The magnitude frequency response of the HBWT filter bank is given in Figure 2.14.

The HBWT is discussed in more details in Sections 3.3 and 3.5. For an overview on Wavelet Transforms and Cosine Modulated Filter Banks the reader is referred to the existing literature [13], [47], [89], [96].

2.2 The pseudo-periodic $1/f$–like model

The main reason for defining and studying the pseudo-periodic $1/f$ -like noise model is the experimental evidence revealing an approximate pseudo-periodic $1/f$ behavior of the spectral harmonic sidebands of voiced sounds in music (see Figure 2.1). In the next chapter, we give a rigorous definition of the pseudo-periodic $1/f$-like noise in terms of a cosine modulation and demodulation scheme. For the present discussion it is sufficient to think about pseudo-periodic $1/f$ signals as signals with spectral harmonic peaks, whose sidebands have approximately a $1/|f - f_k|$ behavior, where f_k is the frequency of the k^{th} partial. The main idea is to adapt the spectrum of a synthesized pseudo-periodic $1/f$-like signal to that of a real-life sound. The synthesis process is controlled by means of a restricted set of parameters defining the $1/f$ shape of each harmonic sideband of the synthetic spectrum. The analysis scheme necessary for the extraction of the resynthesis parameters and the synthesis scheme are implemented by a HBWT filter bank and its inverse, respectively. This model has the advantage of being extremely concise. The lower limit of a single parameter controlling the spectral shape of the corresponding harmonic sideband would be an extremely good result from a data compression point of view. Actually, some refinements are necessary in order to reach a good quality in sound reproduction at the cost of an increased number of parameters. These refinements are the subject of the next section. In this section we describe the pure pseudo-periodic $1/f$-like model.

2.2.1 The WT and the $1/f$ noise

As a first step we consider the analysis and synthesis method of simple $1/f$ noise by means of wavelets introduced in the previous section. The main idea is to exploit the scaling properties of both wavelets and $1/f$ signals. In the previous section we have seen how dyadic wavelets are based on a dyadic scale law, i.e. at different scale the wavelet functions are similar. This dyadic scale law corresponds in the frequency domain to an octave band subdivision of the frequency axis. The same conclusions can be drawn from the analysis of $1/f$ signals, when we consider a logarithmic subdivision of the frequency axis (for instance an octave band subdivision as in dyadic wavelets). In first approximation, the $1/f$ spectrum has constant energy within each octave (see Figure

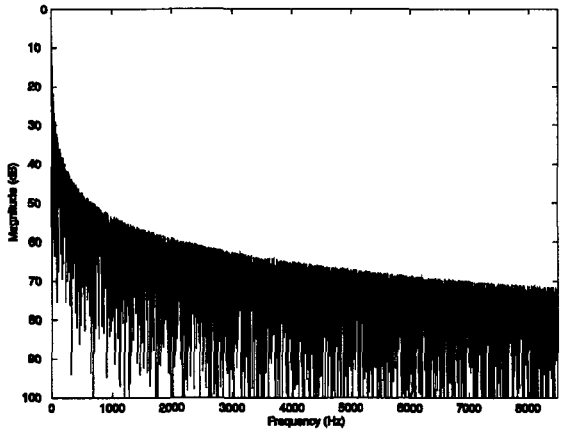


Figure 2.9: Magnitude FT of a $1/f$ noise

2.9). From an intuitive point of view, what happens at low frequencies happens in the same way at higher frequencies with exponentially decreasing amplitude. Roughly speaking, it is in this, which the fractal properties of the $1/f$ noise consist. Combining these characteristics with the time-scale analysis provided by wavelets one obtains a well suited tool for the analysis and synthesis of $1/f$ noise. We are able to model a $1/f$ spectrum by means of a superposition of properly scaled wavelets (Figure 2.10 and 2.11). The main point of the model is that, having a set of wavelet filters, we can just "feed" them with white noise coefficients. The scaling factor of the wavelet bands, i.e. the $1/f$ slope, will be determined by the energies of the white noise coefficients.

The fractal nature of $1/f$ signals and the idea of modeling the harmonic sidebands of voiced sounds by means of $1/f$ "spectral segments" justify the denomination "fractal synthesis". The term "additive" is due to the fact that we process each $1/f$ sideband separately and then we sum all of them together. With respect to the ordinary additive synthesis we not only add sinusoidal components but also noisy components. It is possible to show (see the next chapter) that a discrete-time signal synthesized by means of a wavelet filter bank, employing zero-mean white noise coefficients with properly scaled energy, has an average power spectrum of the following type:

$$\overline{S}_N(\omega) = \sigma^2 \sum_{n=1}^{N} 2^{n\gamma} \left|\Psi_{n,0}(\omega)\right|^2 + 2^{N\gamma} \left|\Phi_{N,0}(\omega)\right|^2, \tag{2.3}$$

where $\Psi_{n,0}(\omega)$ represents the Fourier transform of the wavelet function, $\Phi_{N,0}(\omega)$ the Fourier transform of the corresponding scaling function, γ is a parameter controlling the slope of the $1/f$ spectrum and σ is a parameter controlling the overall energy. The second index m was set arbitrarily to 0, since both $\left|\Psi_{n,0}(\omega)\right|^2$ and $\left|\Phi_{N,0}(\omega)\right|^2$ are invariant by time shift. The spectrum in (2.3) is a multilevel approximation of a $1/f$ behavior as depicted in Figure 2.10. The amplitude at each scale is controlled by the factor $\sigma^2 2^{n\gamma}$, i.e. by the parameters

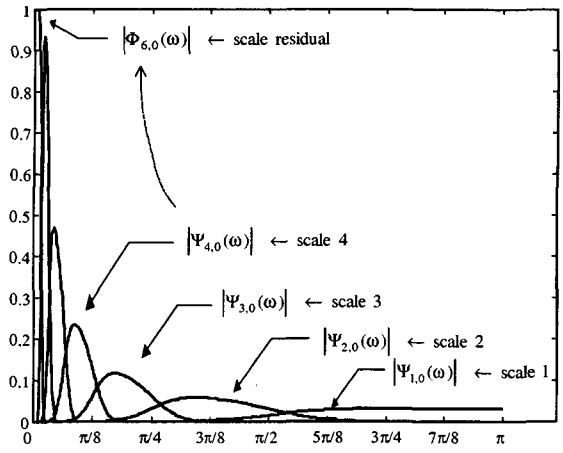


Figure 2.10: Magnitude FT of the filters implementing the DWT.

γ and σ. Notice that the approximation of the $1/f$ spectrum synthesized by Daubechies wavelets [13] is very accurate.

The synthesis of $1/f$ noise by means of the WT is illustrated in a more formal way in Section 3.2.

2.2.2 The HBWT and the pseudo-periodic $1/f$-like noise

We want to extend this method to the pseudo-periodic case, i.e. to voiced sounds whose spectrum is $\sum_k 1/|f - f_k|$-like, where $f_k = k/P$ is the k^{th} harmonic peak, $k = 1, 2, ..., \lfloor P/2 \rfloor - 1$ and P is the sound average pitch. As already mentioned, we want to separate each sideband of the harmonics by means of wavelets. This is the purpose of the P-channel filter bank, where the number of channels P corresponds to the average pitch of the sound. Each filter $g_p(l)$, $p = 0, 1, .., P-1$ has a nominal bandwidth of $\Delta\omega = \pi/P$ or, equivalently, $\Delta f = 1/2P$ and its central frequency is tuned to one of the sidebands of the harmonics.

In other words, the k^{th} harmonic is processed by means of the pair of filters corresponding to the indexes $p = 2k - 1$ and $p = 2k$. In Figure 2.12, we show a single harmonic component of a voiced sound, while Figure 2.13 represents the analytical "frequency grid" of the $2N + 2$ HBWT subbands spanning the band of width $2\pi/P$, which correspond to the harmonics and their two sidebands. Actually, what the cosine modulated P-channel filters perform is not only a bandpass filtering but also a base-band shift of the resulting signal, according to the demodulation scheme illustrated in the next chapter. The base-band signal can then be downsampled P times. After downsampling, we obtain a $1/f$-like signal. This signal can be processed by means of a wavelet filter bank, according to the method for the analysis and synthesis of simple $1/f$ noise described before. According to the results for the simple $1/f$ noise, we can

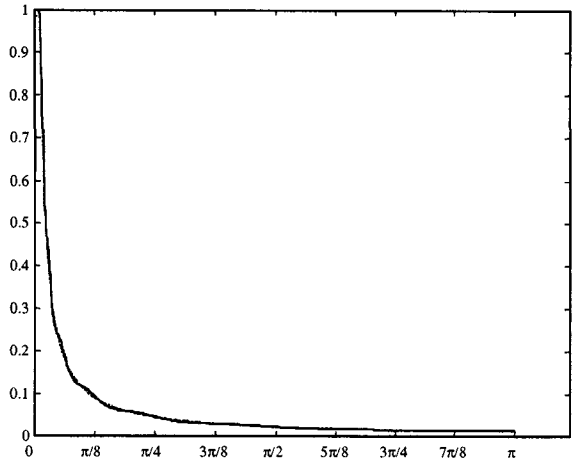


Figure 2.11: Synthesis of $1/f$ noise. a) Magnitude Fourier Transforms of Daubechies Wavelets. b) $1/f$-like noise synthesized by means of Daubechies Wavelets (solid line) compared to the ideal $1/f$ behavior (dashed line).

thus adopt sets of white noise coefficients to reproduce the $1/f$-like behavior of each harmonic sideband. The average spectrum of a synthetic pseudo-periodic $1/f$-like signal is given by:

$$\overline{S}(\omega) = \frac{1}{P}\sum_{p=1}^{P}\sigma_p^2 \left|G_{p,0}(\omega)\right|^2 \left(\sum_{n=1}^{N} 2^{n\gamma_p}\left|\Psi_{n,0}(P\omega)\right|^2 + 2^{N\gamma_p}\left|\Phi_{N,0}(P\omega)\right|^2\right), \tag{2.4}$$

where $G_{p,0}(\omega)$ is the frequency response of the p^{th} filter of the P-channel filter bank, $\Psi_{n,0}(P\omega)$ is the Fourier transform of the upsampled version of a wavelet function at scale n and $\Phi_{N,0}(P\omega)$ is the Fourier transform of the upsampled version of the corresponding scale residual function. Each sideband is then characterized by the parameters σ_p and γ_p of equation (2.4). The first parameter controls the overall spectral energy of the sideband and the second one the shape of the $1/f$ slope, i.e. the energies of the white noise coefficients at the different scales. A larger parameter corresponds to a narrower sideband. In the following section we will see more in detail how these parameters can be extracted from the HBWT analysis. Equation (2.4) is a "periodic version" of equation (2.3). The spectrum resulting from (2.4) is shown in Figure 2.15.

Our goal is to employ the inverse HBWT as a synthesis tool for modeling a pseudo-periodic spectrum of a real-life voiced sound of the type shown in Figure 2.1. The whole method is schematized in Figure 2.16, where one can see how the amplitudes of the HBWT subbands are scaled according to the analysis of a real-life sound. The resulting "spectral mask" is then fed by means of white noise coefficients, in order to provide a sound whose spectrum is similar to that of the input. In this way, we obtain an efficient scheme for reproducing the noisy

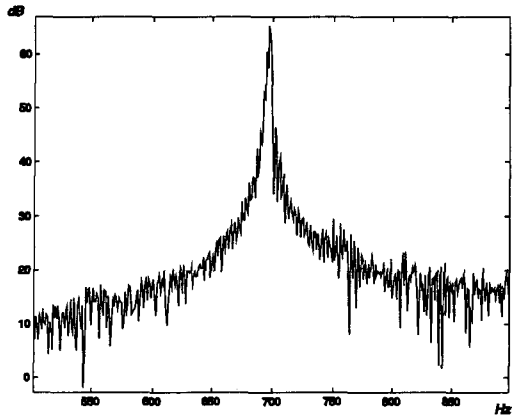


Figure 2.12: Magnitude FT of a single harmonic of a trumpet

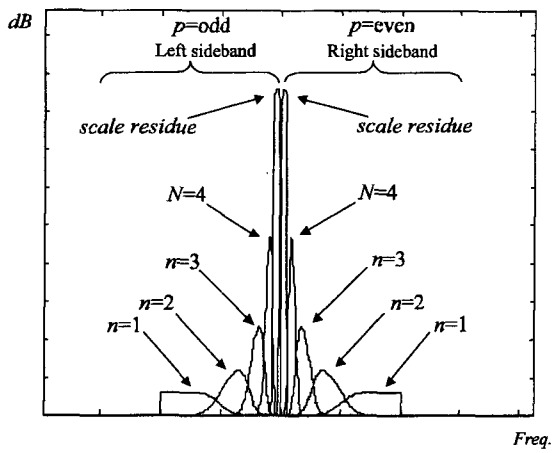


Figure 2.13: Magnitude Fourier transforms of the HBWT subband decomposition of a single harmonic. Left and right sidebands.

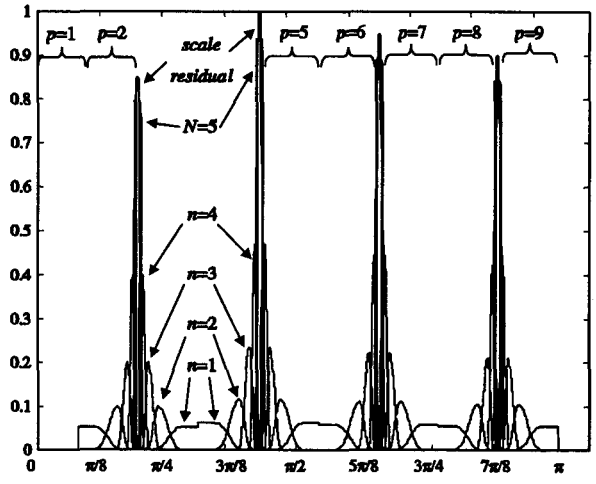


Figure 2.14: Magnitude Fourier transform of the harmonic-band wavelet.

sidebands of the spectrum of voiced sounds in music. By means of very few parameters, compared to the amount of audio data, we are able to generate the necessary synthesis coefficients, i.e. white noise with properly scaled energies. As already said, the parameters σ_p and γ_p are derived from the HBWT analysis itself.

In Section 4.2, we present in more detail the experimental evaluation of the parameters γ_p by means of the results of linear regression tests on the energies of the HBWT analysis coefficients of the subbands. According to these experimental results, the $\sum_k 1/|f - f_k|$-like spectral shape assumption is justified. The pseudo-periodic $1/f$-like model is discussed in a more exhaustive way in Sections 3.3 and 3.4.From an acoustic point of view the method requires further refinements. The white noise resynthesis coefficient assumption is too strict and not satisfactory. The next section discusses how it is possible to overcome these limitations.

2.3 Fractal additive analysis and synthesis method

Based on the pseudo-periodic $1/f$-like scheme, we want to build a complete, flexible and acoustically efficient analysis-based synthesis method. The whole method is limited to the stationary part of sound and is particularly well suited for long, sustained sounds. One of the principal aims of the method is to reproduce the natural dynamics of the timbre of a sustained sound, avoiding the static character of sustained synthetic sounds. Currently, one of the most successful synthesis methods is the wavetable synthesis, based on overlap-add techniques of one-period-waveforms of the instrument to synthesize. The synthetic result in the case of long sustained sounds misses the naturalness of timbre that we

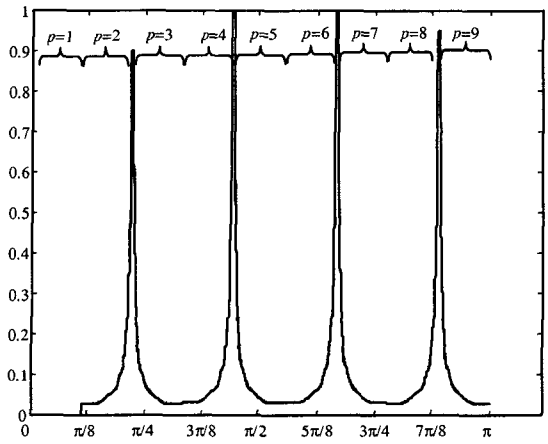


Figure 2.15: Spectrum of an ideal pseudo-periodic $1/f$ noise

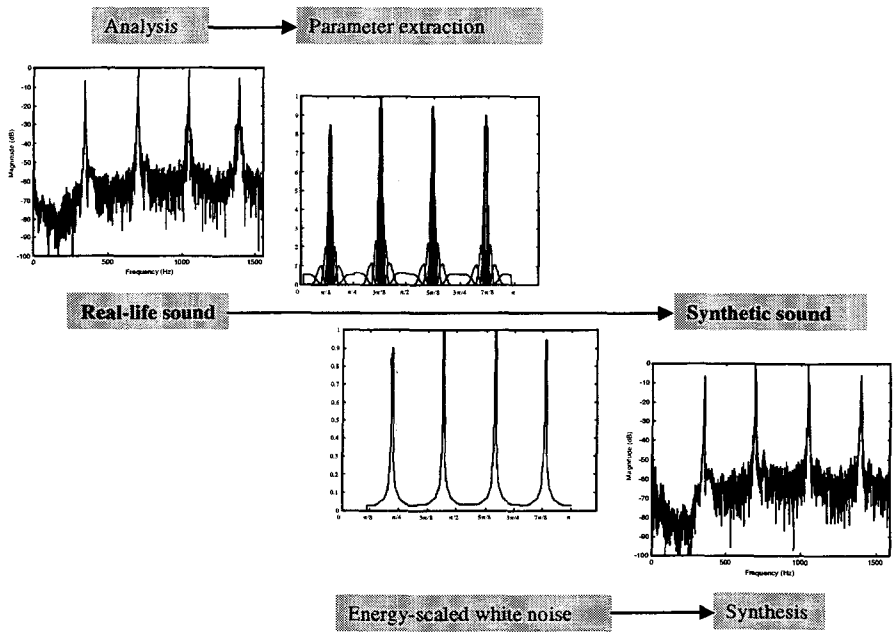


Figure 2.16: The complete analysis and synthesis process from a real-life sound to the final synthetic sound in a spectral representation.

claim for our method.

The transient, that is the sound attack, is preserved as it is. A pitch detector can be adopted in order to define the extension of the transient. Only the portion of sound where a steady pitch is detected is processed according to the FAS model.

2.3.1 FAS stochastic model

We decompose the stationary part of a sound in its harmonic-band time-frequency components. We distinguish these components into three groups. The first group corresponds to the deterministic components of the sound (see Figure 2.6 the narrow white subbands). These components cannot be reproduced by means of noisy coefficients. The model for the deterministic components will be described later. The second group corresponds to the portion of the spectrum close to the harmonics, which contains the microfluctuations with respect to pure periodicity (see Figure 2.6 the light gray subbands). The behavior of these components is well approximated by the $\sum_k 1/|f - f_k|$ model. We can either use one parameter γ per sideband that controls the energy of all of its subbands, or estimate independent parameters controlling the energy of each HBWT subband separately. In the first case, we need to perform a linear regression and we minimize the total number of parameters. In the latter case we are able to provide a better approximation of the spectrum at the cost of an increased number of parameters. The third group of components includes the first subbands of the HBWT decomposition (see the dark gray subbands in Figure 2.6). According to the analysis results, it is possible to see how these subbands, which lie far away from the harmonics, contain the most significant and unmasked information concerning the additional noise due to the excitation systems. These noises include, for example, breath noise or reed buzz in wind instruments and bow noise in string instruments. Usually their energy is larger than that expected according to the $\sum_k 1/|f - f_k|$ model.

At this point, it is necessary to define the parameters of the model, i.e. the parameters that represent the analysis coefficients and control the generation of the synthesis coefficients (see Figure 2.17 and 2.18). As already mentioned, the simple white noise coefficient approximation is not satisfactory from an acoustical point of view. We obtain something that sounds as properly energy-scaled white noise. In fact, we know that the HBWT analysis coefficients are not completely uncorrelated. A non-zero autocorrelation is detectable within the coefficients of each scale of each channel, while no relevant cross correlation exists between coefficients of different scales and different channel. Thus, it was necessary to improve the method by introducing an LPC (Linear Predictive Coding) analysis of the HBWT analysis coefficients in order to detect and then reproduce the existing autocorrelation. The autoregressive (AR) filters so obtained are used to color the white noise used as input to the resynthesis filter bank, thus reproducing the time-correlation within the subbands. This is fundamental to make the noisy synthetic subbands similar to the real ones from an acoustical point of view. Additionally in the analysis part (Figure 2.17) we compute the variance of the coefficients of each subband, in order to estimate their energy. The estimate is performed over windowed sets of coefficients and a time envelope is extracted from the energy values. The FAS stochastic model

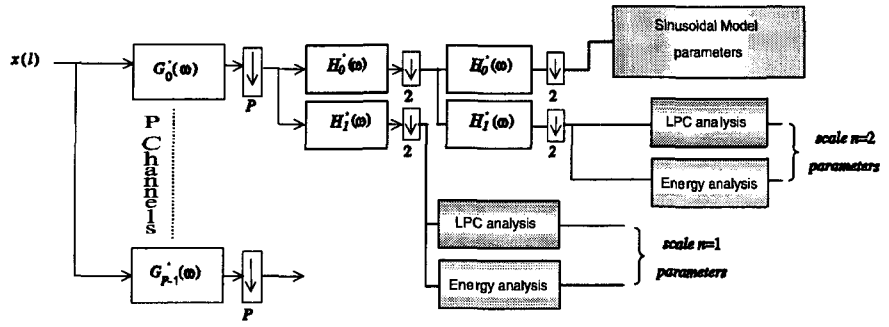


Figure 2.17: HBWT analysis-based parameter extraction for the fractal additive resynthesis.

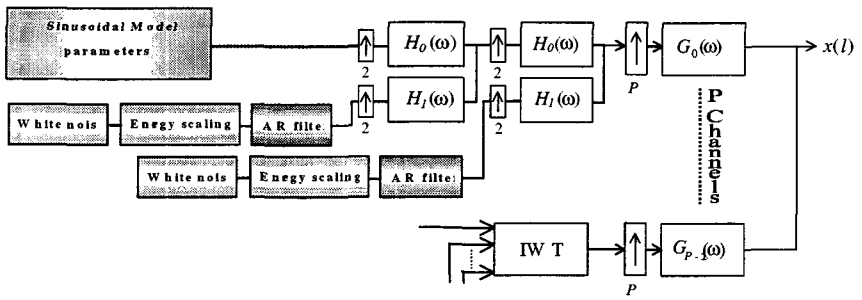


Figure 2.18: Fractal additive synthesis scheme.

is presented more in details in Section 4.3.

2.3.2 FAS deterministic model

Our model for the deterministic components recalls somehow the sinusoidal modeling techniques even if it has different implications. In the sinusoidal approach, one models the amplitude and the phase of the partial sinusoidal components. In the FAS the role of the sinusoidal partials is played by the two MDCT sidebands of the partial spectral peak. For the resynthesis, it is necessary to model the detuning of the real life sound partial with respect to a fixed pitch P. In order to do this, we consider a complex combination of the two sets of HBWT analysis coefficients relative to each peak and we model the amplitude and the phase of the so obtained complex coefficients. From Figure 2.19 and 2.20, it is easy to see that the curves drawn by the amplitudes and the phases of the complex sinusoidal coefficients form somehow smooth and regular curves. In particular, the phases are almost linear, where the slope of the linear curves is related to the fact that the pitch of the sound is in general not an integer division of the sampling frequency. The amplitude curves present some ripples, which are nevertheless irrelevant from a perceptual point of view. Therefore, the piecewise linear approximation shown in 2.19 is more than sufficient in order

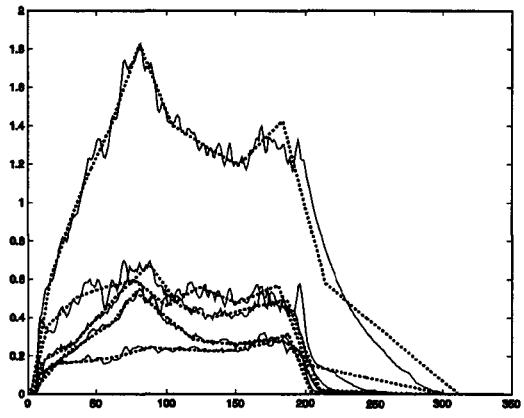


Figure 2.19: Polynomial interpolation of the amplitude envelopes of the first 5 harmonics of a violin sound (D3)

to reproduce the synthesis coefficients for a high quality reproduction of a real life sound.

The high-level parametric representation of the synthesis coefficients makes our method also an interesting tool for sound synthesis and sound processing in terms of digital audio effects. In the latter case, the resynthesis coefficients are generated by means of a modulation of the parameters obtained from the analysis of the processed sound.

The FAS deterministic model is presented more in details in Section 4.4.

2.3.3 FAS extensions

The case of sounds with variable pitch is the next problem one needs to face. Typically, a vibrato sound is a good example in this sense. Its pitch varies like in Figure 2.21. By means of a pitch detector one obtains the pitch values of the kind shown in Figure 2.22. $P(r)$ is now the time-varying pitch and thus the time-varying number of channels of the multiband filter bank. The design of a $P(r)$-channel filter bank as the one shown in Figure 2.23 is not a straightforward task. The overall structure of the wavelet transformation and the coefficient parametric modeling do not change at all. The only matter is the design of a filter bank able to change the number of channels period by period, while maintaining the PR constraints and the smoothness of the coefficients curves as those of Figure 2.19 and 2.20. An example of that is introduced and discussed in Section 5.1.

Finally, an extension of the method to the inharmonic case is taken into consideration. So far the HBWT model has been confined to the harmonic spectrum case. The time-frequency plane tiling was strictly harmonic. This is a major limitation and makes the method unusable for a large class of sounds, for instance all the sounds produced by percussion instruments. The spectra of many of these instruments show relevant peaks centered on non-harmonically distributed frequencies (see Figure 2.24). These peaks are the partials or deterministic components of the sound and can be sinusoidally modeled. The

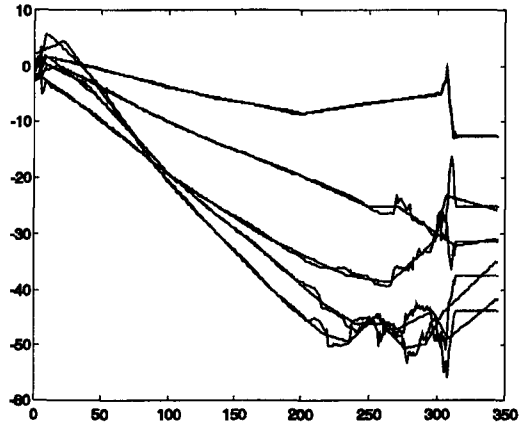


Figure 2.20: Polynomial interpolation of the phase envelopes of the first 5 harmonics of a violin sound (D3)

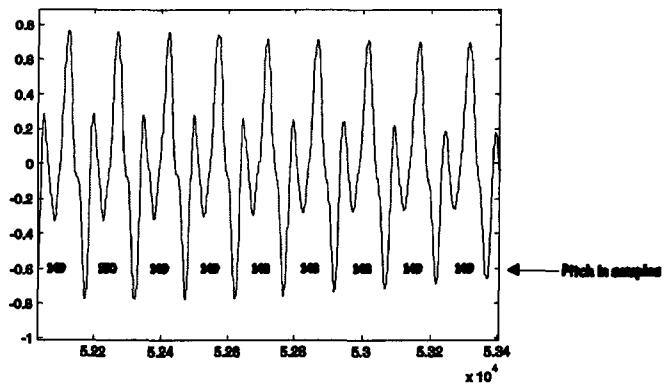


Figure 2.21: 9 periods of a flute note with pitch variable from 148 to 150 samples

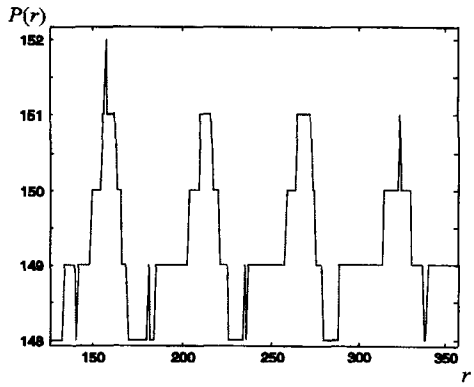


Figure 2.22: Varying pitch of a flute note with vibrato. r indexes the periods.

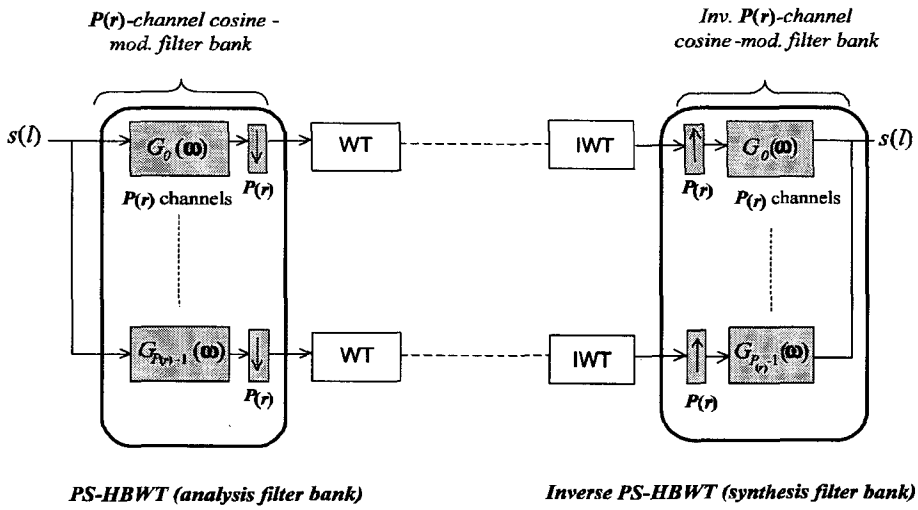


Figure 2.23: PS-HBWT analysis and synthesis filter bandks. The index r denotes the sequence of periods.

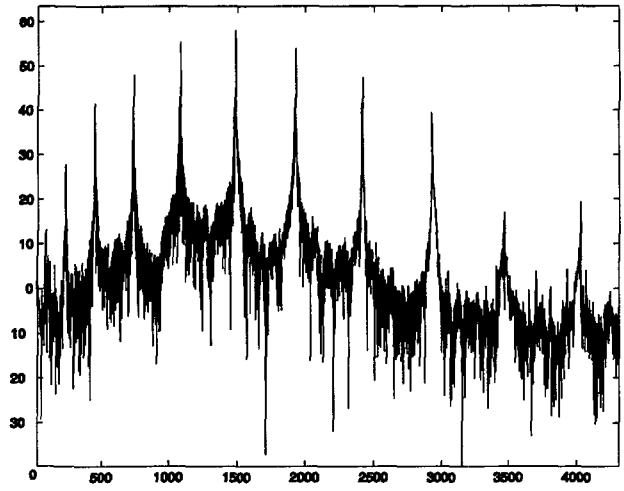


Figure 2.24: Magnitude FT of a tubular bell sound

partials also show an approximately $1/f$ spectral behavior around the peaks as in the harmonic case. These $1/f$-shaped spectral sidebands are treated as stochastic components, i.e. the same stochastic model used in the harmonic case is employed. It is therefore useful to find a way to extend the FAS method to sounds with spectra of the kind of Figure 2.24. The main problem is to provide a more flexible analysis/synthesis structure extending the FAS model to inharmonic sounds. In order to do this, we abandon the perfect reconstruction (PR) structure provided by the HBWT and resort to a non-PR scheme able to deal with aperiodic spectra like the one in Figure 2.24. A non-PR structure leads to aliasing problems and artifacts in the resynthesis. These artifacts are minimized by the filter design procedure and optimization. Figure 2.25 shows how by reconstructing both the n^{th} partial and the aliasing due to the downsampling of order P_n and subtracting it to the partial residue one can keep track of the aliasing through the following partial analysis steps. In this way, we obtain a reduction of the aliasing. At the limit of the overall residue tending to zero, the scheme of Figure 2.25 is PR. This part is discussed more in detail in Section 5.2.

2.4 Experimental results

FAS can be seen both as a method for audio coding and data compression and as a sound processing tool for sound synthesis and processing. These subjects, introduced in the next two subsections, are discussed in a more exhaustive way in Chapters 4 and 5.

2.4.1 An audio coding tool

The experimental results change significantly according to the instrument we analyze and resynthesize. The first step is the choice of the wavelet scale at

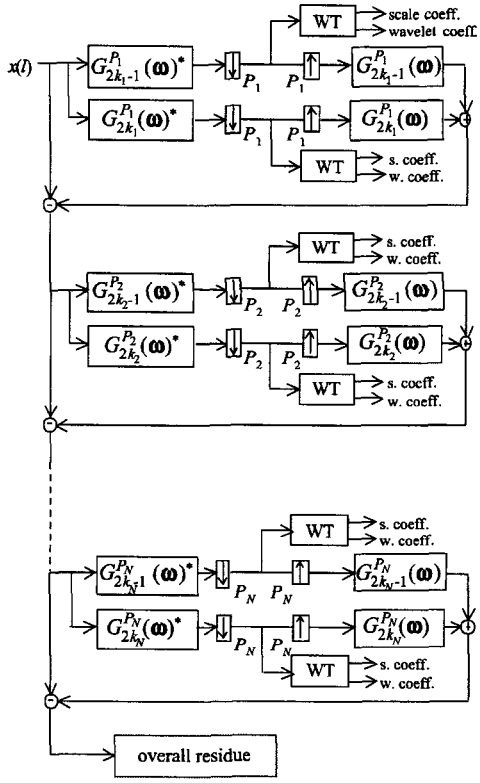


Figure 2.25: Inharmonic analysis filter bank. For each partial $n = 1, .., N$ of an inharmonic sound we find a "hypothetical pitch" P_n, which could "fit" the partial itself. From the P_n-channel filter bank we select only the filters corresponding to the channels $2k-1$ and $2k$, where k is the index of the harmonic of the P_n bands coinciding with the partial n. The outputs of the filters $G_{2k-1}(\omega)^*$ and $G_{2k}(\omega)^*$ undergo a wavelet transformation as in the harmonic case. The reconstruction of each partial by means of the filters $G_{2k-1}(\omega)$ and $G_{2k}(\omega)$ and the subtraction from the residue signal allow us to keep track of the aliasing, with the purpose of reducing it.

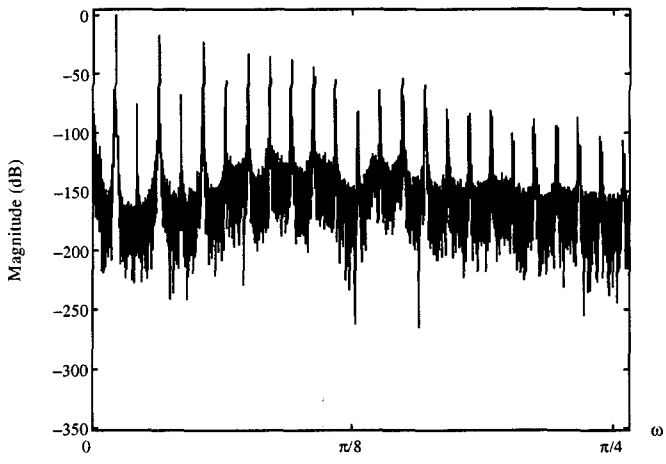


Figure 2.26: Magnitude Fourier transform of a real-life clarinet note.

which we stop the analysis. This defines which portion of the spectrum we resynthesize by means of the stochastic model and which portion by means of the deterministic-sinusoidal model. Normally 2 or 3 scales are reasonable in order to preserve the main time characteristic of the sound as the harmonics and their time envelopes. The second step is to define the extension of the transients; the length of the attack varies largely according to the instrument, the pitch and the stabilization speed of the sound.

We tested our algorithm with different instrument sounds: a violin, a viola, a cello, a flute, an oboe, a bassoon, a clarinet, a trumpet, a French horn and a trombone. The two different degrees of approximation, i.e. a) a single parameter γ per sideband, b) independent subbands give appreciably different acoustical results. The last case provides very good results in terms of sound reproduction. The AR filters employed are of the 10^{th} order for the second subband. The order diminishes with the order of the subbands. The energy values of the subbands are updated every 20 coefficients. We performed a perfect reconstruction of each subband separately in order to compare them with the synthetic ones. In Figure 2.26 and 2.27 the magnitude FT of a real-life clarinet and the magnitude of the synthetic clarinet, respectively, are reported.

In order to use this method as a data compression tool, psychoacoustic criteria play a fundamental role. By means of the computation of masking effects, we are able to discard a significant percentage of the analysis coefficients. This leads to compression performances of the order of 25:1 for the stationary part of a monophonic voiced sound. For more details and numerical results, see Section 4.5.

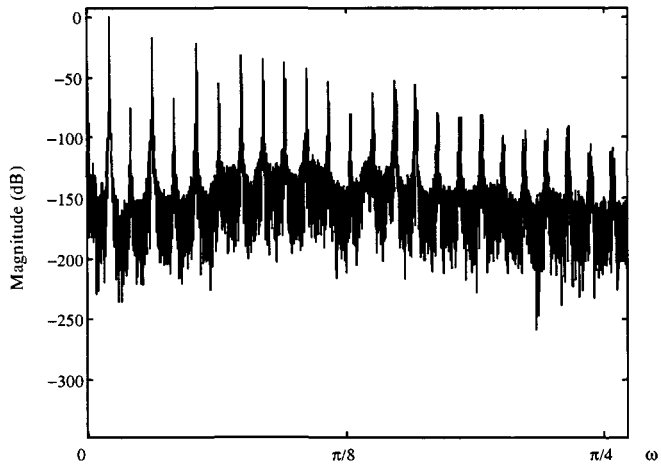


Figure 2.27: Magnitude FT of a synthetic version of the clarinet of Figure 2.26 obtained by means of the FAS.

2.4.2 A sound design tool

From the previous discussion, it is also evident how the HBWT perfect reconstruction scheme can be used as a powerful sound analysis tool. It is possible to separate the harmonic components from the noisy components at different scales. We can in fact employ any subset of the HBWT analysis coefficients as input to the IHBWT filter bank. We can for instance reconstruct a single wavelet-band n of all the channels p. We can extract one specific subband of one specific sideband (fixed n and p), as well as any other arbitrary combination of subbands. Also, we can separate the whole noisy component of a single harmonic sideband. Each component can be processed and analyzed separately. We are for instance able to separate the lip impulses of a brass-instrument player, as in the case of a trombone.

From another perspective, the method is a powerful tool for sound hybridization. From a single real-life sound we can in fact extract a wide gamut of new sounds, according to the already described possibilities of combining the HBWT subbands. In such a way timbre hybridization is straightforward: we can realize any "mixture" of subbands coming from the analysis of different instruments. A very simple example can be obtained by combining the reconstructed harmonic components of one instrument with the noise sidebands of another one. For instance an oboe with the noise of the bow of a string instrument or a violin with the noise of the breath of a flute. This can be successfully employed as a new cross-synthesis technique. We obtained interesting results in combining the subbands of a horn, a trumpet, a bassoon, a clarinet and an oboe. See also Section 4.1.

Finally, the method is also a new synthesis method on its own. Our final

goal is to realize a real-time system. A prototype has been realized in Pure data (Pd) [66], [65], a software environment for real time digital audio processing (see Section 5.3). An ordinary additive synthesis system is easily implementable in real-time by means of an amplitude controller per harmonic partial. By means of a set of "sliders" one is able to control dynamically the energy of the corresponding sinusoidal components. As already mentioned, FAS is a form of additive synthesis, where one adds both sinusoidal and noisy components together. Thus, in the case of FAS, the number of sliders per harmonic partial increases. We can decide to have for instance 2 or 3 sliders per harmonic partial, according to the degree of accuracy one wants to reach. By means of 2 sliders we can control the partial amplitude and the slope of both the $1/|f - f_k|$-like sidebands, i.e. of all of the noisy components of the harmonics at once. If one wants to control the two sideband slopes independently, 3 sliders would be necessary. Finally one could employ a number of sliders equal to the number of subbands plus one (7 for 3 scale levels) in order to control the stochastic components and the deterministic components of each partial energy separately. As illustrated in the last chapter, this is the solution adopted in our prototype. In order to generate the colored noisy coefficients, the system would also need to implement a bank of $2N$ AR filters per harmonic partial. The AR filters organized in presets, come from the analysis of real-life musical instruments. The interpolated time envelopes of both the stochastic and deterministic coefficients are also drawn from the analysis and stocked as presets and editable in a graphical way. A random generator provides the raw white noise coefficients. The implemented prototype is described in more detail in Section 5.3.

Some audio examples are retrievable at the address:
http://lcavwww.epfl.ch/~pietro/audio_examples.

2.5 Summary

In this chapter, we have given an overview of FAS. FAS is a method to deal with the different components of voiced sounds, focusing our attention on their noisy components, which are important for maintaining a real-life "color" in sounds. Our method provides a convincing noise model for voiced sounds in music with very good experimental results. The performance of the method as a data compression tool, even if confined to the stationary part of voiced sounds, are satisfactory. We have pointed out that this method is also an analysis tool able to separate the harmonic components from the noisy components of sound. Alternately, it can be viewed as an independent synthesis method.

An improvement of the method was achieved by devising a pitch synchronous version, i.e. a time varying version of the filter banks. This frees the method from the limitations of a fixed number of channels, which restricted the set of sounds that one can analyze to those with a well-defined and stable pitch. By means of this extension, the method can be applied, for instance, to vibrato sounds.

Another extension was obtained by means of an inharmonic version of the P-channel filter bank. This allows one to define the frequency range of the sidebands, i.e. of the partials, in an arbitrary way. We are able to employ the method also in the case of nearly inharmonic sounds, such as in the low register of the piano, or in the case of inharmonic sounds as those produced by

percussion instruments.

Chapter 3

The Pseudo-Periodic 1/f Noise

In this chapter, we detail the pseudo-periodic $1/f$ model and provide a formal definition of both the pseudo-periodic $1/f$ noise and the Harmonic-Band Wavelet Transform (HBWT). $1/f$ noise and long term correlated stochastic processes inspired the first steps of this research for a new and effective model for noise in musical sound that led to the FAS method.

The starting point is a powerful method for the synthesis of $1/f$ stochastic processes by means of orthonormal wavelet bases introduced in [99]. The main idea is that, in order to obtain a good approximation of a given $1/f$ stochastic process, it is possible to adopt collections of mutually uncorrelated zero-mean processes with proper scale-dependent energy as wavelet synthesis coefficients. A single parameter is sufficient to control the slope of the $1/f$ -shaped power spectrum of the synthetic signal. This parameter determines the variances, i.e. the energies of the synthesis coefficients for each different wavelet subband. Based on this result, we will introduce a scheme for the analysis and synthesis of pseudo-periodic signals. We need:

a) to define in a formal way a pseudo-periodic signal model, i.e. the pseudo-periodic $1/f$-like noise

b) to define an appropriate mathematical tool for the analysis and the synthesis of this type of signals, i.e. the HBWT.

The theoretical investigation of this chapter is aimed at solving both problems at once by introducing a general cosine modulation and demodulation scheme. Thanks to this scheme we are able to provide a rigorous definition of the pseudo-periodic $1/f$-like noise, as well as to define consistently the multichannel filter bank basis functions, which form with the ordinary WT one of the two ingredients of the HBWT. The introduction of a multirate filter bank as a signal vectoring operator allows one to extend the class of Multiplexed Wavelet Transforms (MWT) [19] to the HBWT. The HBWT provides a new class of wavelet transforms well-suited for separating and analyzing the harmonics of sounds with a detectable pitch. As it will be discussed in detail later in this chapter, harmonic separation is performed by means of a cosine modulated filter bank. Each output of the filter bank is then analyzed by means of a wavelet filter bank. Compared to the MWT, the HBWT allows one to process each

sideband of each partial independently.

The chapter is organized as follows. In section 3.1, we revisit the properties and the characteristics of the $1/f$ noise. In Section 3.2, we briefly review the synthesis of $1/f$ processes by means of the WT. In Section 3.3, we define the pseudo-periodic $1/f$ noise process by means of a harmonic-band modulation and demodulation scheme. In Section 3.4, we illustrate the theoretical result on which the FAS is based. The proof of the main theorem is reported in Appendix A. In Section 3.5, we introduce the discrete-time HBWT and their properties. We also describe an operational scheme for the analysis and synthesis of the pseudo-periodic $1/f$ noise. Section 3.6 illustrates the results obtained by means of a refinement of the method, which adopts a Frequency Warped version of the Wavelet Transform (FWWT).

3.1 $1/f$ noise

As illustrated in the previous chapter, the power spectra of musical signals of the kind of Figure 3.2 contain peaks centered on the harmonics, whose shape is influenced by the long-term correlation of the stochastic fluctuations with respect to a pure periodic behavior. From a perceptual point of view, these chaotic but correlated microfluctuations are relevant if one needs to emulate naturalness and timbre dynamics in sounds with a detectable pitch. $1/f$ processes arise in many physical and biological systems as well as in man-made phenomena such as variations in traffic flow, economic data, network traffic as well as in music [98] [102]. These processes are significantly correlated at large time lags. In this perspective, $1/f$ noise plays somehow the role of the "main character" in our work.

This section is devoted to a short review of the principal points characterizing the $1/f$ noise. As a general introduction to the problem one can state that the $1/f$ noise is a random and non-stationary process, whose name refers to the behavior of its average "power spectrum".

$$\overline{S}(f) = \frac{const}{|f|^{\gamma}}, \tag{3.1}$$

where f is the frequency and γ is a parameter with $0 < \gamma < 2$. Usually $\gamma \simeq 1$. The "power spectrum" in (3.1), strictly speaking, is not defined. As an anticipation, we could say that the meaning of the (3.1) lays on the fact that, even if the $1/f$ noise is non-stationary, its increments are stationary. This main problem and the other two main characteristics of the $1/f$ noise, i.e. its long term correlation and its fractal nature are shortly recapitulated in the following subsections.

3.1.1 The $1/f$ noise: a non stationary random process.

The $1/f$ noise appears as the spectrum of the fluctuations of the parameters of many physical systems [36]. It was detected first as an excess of low frequencies in the vacuum tubes and much later in the semiconductors. Bernamont in 1937 [6] and McWorther in 1955 [101] developed models for the $1/f$ noise for the vacuum tubes and for the semiconductors, respectively. From the first half of

the 50s the $1/f$ noise was observed as fluctuation of the parameters of many other physical systems.

Soon, the non stationariety of the $1/f$ noise or, in other words, the divergency of the integral of the spectrum (3.1) appeared as a problem. Different theoretical approaches were attempted in order to solve the problem. A first proposal for the definition of the spectrum of the $1/f$ processes was given by Mandelbrot. He suggested that the $1/f$ noise could be considered as a non stationary process. In this hypothesis the variance is time dependent and the autocorrelation $R(t_1, t_2)$ depends on both t_1 and t_2 explicitly. In this perspective, it becomes fundamental to consider that the process can be observed only for a finite time (that means a finite variance) and that one needs to know the state of the system at a certain time $t_0 < t$, where t is the present time. The experimental evidence is in open contradiction with Mandelbrot's hypothesis. In fact the power spectrum of the $1/f$ noise can be measured with no assumptions on its stationariety and without knowing its initial state.

In the time domain, without going into details of models such as the fractional Brownian motion (fBm) [49] [28] [26] or the discrete fractional Gaussian noise (dfGn) [4] [16], we can say at an intuitive level that one can consider the $1/f$ noise as a non stationary process, whose increments are stationary. In the frequency domain this corresponds to say that the $1/f$ noise is a stochastic process that appears stationary when filtered by means of an ideal bandpass filter. A more detailed characterization and definition of the $1/f$ noise in the frequency domain is illustrated in Section 3.1.4

3.1.2 The " memory" of the $1/f$ noise

A second fundamental characteristic of the $1/f$ noise is that of being a stochastic process "less random" than other types of noise. In fact, the $1/f$ noise presents a long term correlation among events, i.e. it has an evolutionary character. In other words, the behavior at a certain time is strongly influenced by the previous history of the process. The influence of the events at increasing time lag decays in a much slower way with respect to the exponential decays corresponding to the differential equations usually associated to the models of physical systems. It is relevant how the $1/f$ noise appears to be also effective for modeling the fluctuations of the notes and the dynamics of many genders of music, i.e. of a product of the human mind.

By memory of a stochastic process we mean the decay rate of its autocorrelation: the lower the decay rate, the more the present events are influenced by the past behavior of the process itself. White noise does not have memory of the past: its autocorrelation $R(t_1, t_2)$ is 0 for $t_1 \neq t_2$. Many processes are not white, but their autocorrelation decays quickly. On the contrary $1/f$ processes have a very long memory, that is the decay of the autocorrelation function is very slow, of the order of $1/t^n$ or even logarithmic but never exponential. The nearer γ is to 1 the bigger is the influence of the past. For γ approaching 0 or 2 the autocorrelation decay with time lag becomes faster. Assuming that the process is modeled by a linear system and that its past history is entirely represented by the present values of its state variables, how many variables are necessary for a system when its fluctuations have a power spectrum $1/f$? i.e. how many numbers are necessary in order to describe the influence of the past on the present? For the white noise ($\gamma = 0$) the answer is 0; for the Brownian

motion ($\gamma = 2$) is 1: the initial position; for γ =1 the answer changes radically. We can estimate it, for instance, by determining how many variables are necessary for a linear fit with an error of $\pm 5\%$. The present behavior is influenced in a homogeneous way by each of these variables and each one represents a trend of the data at different time scales. The idea that the information describing the past is summed up and stored as trends at different time scales is particularly appealing. It seems close to the fashion in which human beings record information as parts of consistent models through more levels rather than as separated and uncorrelated pieces. Also noticeable the fact that music parameters present a statistical behavior similar to the $1/f$ noise and that randomly chosen notes sound more musical, if their spectral density is $1/f$-like. This suggests a relationship between the structure of the $1/f$ noise and the way in which we perceive and remember. Finally, the influence of memory as an experience of the past is implicit in the development of anything produced by humans. This could be an appealing interpretation of the recurrency of a $1/f$ spectral behavior also in the analysis of economic or sociological data.

3.1.3 Fractal properties of the $1/f$ noise

The third fundamental characteristic of the $1/f$ noise is that of being statistically selfsimilar, i.e. the statistic of $1/f$ noise is scale invariant. Intuitively this is the junction element with the WT and its multiresolution properties. More precisely, by statistically selfsimilar processes one denotes all the random processes $x(t)$ that satisfy the relationship:

$$x(t) \stackrel{P}{=} a^{-H} x(at), \tag{3.2}$$

where $\stackrel{P}{=}$ denotes identity in a statistical way, a is any real number and H is the homogeneity degree of the process. If $x(t)$ is Wide Sense Stationary (WSS) and its power spectrum is of the kind of 3.1, it follows that:

$$R_x(\tau) = a^{-2H} R_x(a\tau) \qquad \forall a \in \mathbb{R}$$

Fractal geometry is recurrent in nature. Fractal waveforms characterize, for instance, geographic contours, earthquake distributions, turbulent flows and many others [48] [3]. According to the relation (3.2) $1/f$ processes are fractal signals. The fractal dimension allows one to measure the density of a fractal object within a space $C(X)$, where $C(X)$ is the space of the compact subsets of the metric space X defined on some particular metric. If $A \in C(X)$ is a compact set, $N(A, \epsilon) = M$ is defined as the minimum number of closed spheres of radius ϵ necessary to cover A. One can define a quantity D in the following way [3]:

$$D = \lim_{\epsilon \to 0} \frac{\ln(N(A, \epsilon))}{\ln(1/\epsilon)} \tag{3.3}$$

where the term D, when it exists, is called the fractal dimension of A and $D_x - 1 < D < D_x$, where D_x is the dimension of the space X.

As an example, the fractional Brownian motion self-similar parameter lies in the interval $0 < H < 1$ (corresponding to $1 < \gamma < 3$ [99]) and its fractal dimension is:

$$D = 2 - H = \frac{5}{2} - \frac{\gamma}{2}, \tag{3.4}$$

which intuitively provides a measure of its indentation [2], [99] [100]. Equations (3.3) and (3.4) provide one of the possible way to deduce the value of the parameter γ that is one of the two main parameters necessary for the synthesis of pseudo-periodic signals.

3.1.4 Frequency domain $1/f$ noise characterization

The fundamental notion for a characterization of the $1/f$ processes in the frequency domain lies on empirical considerations: a $1/f$ process is a statistically selfsimilar process, which is stationary when filtered by means of ideal bandpass filters. Since the spectral measures of physical processes can be only referred to a frequency range related to finite times of observations and to a finite resolution, this seems to be the most natural way to define and discriminate the $1/f$ noise from other statistically self-similar processes. More precisely the following definition is given in [99]:

Definition 3.1 *A wide-sense statistically self-similar zero mean random process $x(t)$ shall be said to be a $1/f$ process if there exist ω_0 and ω_1 satisfying $0 < \omega_0 < \omega_1 < \infty$ such that when $x(t)$ is filtered by an ideal bandpass filter with frequency response*

$$B(\omega) = \begin{cases} 1, & \omega_0 < |\omega| < \omega_1 \\ 0, & otherwise \end{cases}$$

the resulting process $y(t)$ is wide-sense stationary and has finite variance.

Also in [99] one can find the following result:

Proposition 3.1 *A $1/f$ process $x(t)$, when filtered by an ideal bandpass filter with frequency response*

$$B(\omega) = \begin{cases} 1, & \omega_L < |\omega| < \omega_U \\ 0, & otherwise \end{cases}$$

with $0 < \omega_L < \omega_U < \infty$, yields a random process $y(t)$ wide-sense stationary, with finite variance and having power spectrum, for some $\sigma_x^2 > 0$

$$S_y(\omega) = \begin{cases} \sigma_x^2 / |\omega|^\gamma, & \omega_L < |\omega| < \omega_U \\ 0, & otherwise \end{cases},$$

where the spectral exponent γ is related to the selfsimilarity parameter H according to the following relationship $\gamma = 2H + 1$.

3.2 $1/f$ noise analysis and synthesis by means of WT

Theorem 3 in [99] shows that given an orthonormal wavelet basis $\psi_{n,m}(t)$ and a collection of mutually uncorrelated zero-mean synthesis coefficients $x_n(m)$ we obtain a process

$$x(t) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} x_n(m) \psi_{n,m}(t)$$

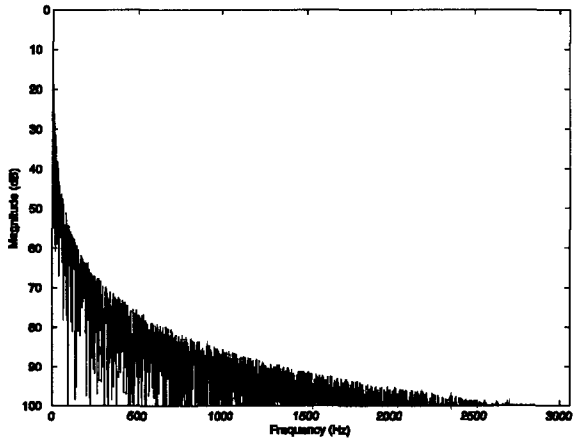


Figure 3.1: Magnitude FT of a $1/f$ stochastic process.

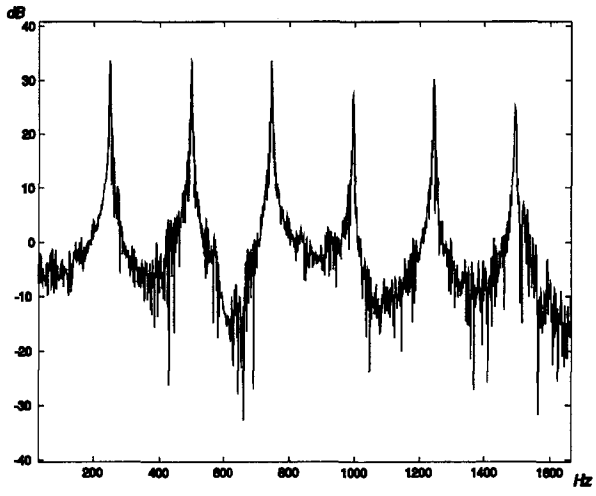


Figure 3.2: First 6 harmonics of a violin note (B2).

that is nearly $1/f$, i.e. its time-averaged power spectrum

$$\bar{S}_x(\omega) = \sigma^2 \sum_{n=-\infty}^{\infty} 2^{\gamma n} \left|\Psi\left(2^n\omega\right)\right|^2 \tag{3.5}$$

satisfies the relationship:

$$\frac{\sigma_L^2}{|\omega|^\gamma} \leq \bar{S}_x(\omega) \leq \frac{\sigma_U^2}{|\omega|^\gamma}$$

for some $0 < 2\sigma_L \leq 2\sigma_U < \infty$, i.e. the average power spectrum of $x(t)$ is upper and lower bounded by an $1/f$ spectrum. Furthermore the following self-similarity relationship holds for any integer k:

$$|\omega|^\gamma \bar{S}_x(\omega) = \left|2^k\omega\right|^\gamma \bar{S}_x(2^k\omega) \tag{3.6}$$

In order to extend this result to pseudo-periodic signals we will introduce a new set of multiwavelets. These multiwavelets are associated to a continuous-time filter bank with an infinite number of channels, whose outputs are down-sampled and analyzed by means of the Discrete-Time Wavelet Transform (DTWT). A discrete-time counterpart of the previous result is thus necessary. It is easy to show that the discrete-time synthesis process.

$$x(l) = \sum_{n=1}^{N} \sum_{m=-\infty}^{\infty} b_n(m)\psi_{n,m}(l) + \sum_{m=-\infty}^{\infty} a_N(m)\phi_{N,m}(l), \tag{3.7}$$

where $\phi_{N,0}(l)$ is the scaling sequence relative to the DTWT $\psi_{n,0}(l)$, is wide-sense cyclostationary (WSCS) of period $2N$ with average power spectrum

$$\overline{S}_N(\omega) = \sum_{n=1}^{N} 2^{n\gamma} \frac{\left|\Psi_{n,0}\left(\omega\right)\right|^2}{2^n} + 2^{N\gamma} \frac{\left|\Phi_{N,0}\left(\omega\right)\right|^2}{2^N} \tag{3.8}$$

Here $\Psi_{n,0}\left(\omega\right)$ represents the DTFT of the wavelet sequence $\psi_{n,0}(l)$ and $\Phi_{n,0}\left(\omega\right)$ the DTFT of the corresponding scaling sequence $\phi_{n,0}(l)$.

Let $H_1(\omega)$ and $H_0(\omega)$ be the frequency responses of the QMF filters used to generate the discrete-time dyadic wavelets $\psi_{n,0}(l)$. They satisfy the relationships (see also Figure 2.4):

$$\left|H_1\left(\omega\right)\right|^2 + \left|H_0\left(\omega\right)\right|^2 = 2$$

and

$$H_0\left(\omega\right) H_1^*\left(\omega+\pi\right) + H_0^*\left(\omega+\pi\right) H_1\left(\omega\right) = 0.$$

From this and the recursive definition of the Discrete-Time Wavelet Transform (DTWT), we have:

$$\frac{\left|\Psi_{n,0}\left(\omega\right)\right|^2}{2^n} = \frac{\left|H_1\left(2^{n-1}\omega\right)\Phi_{n-1,0}\left(\omega\right)\right|^2}{2^n} = \frac{\left|H_1\left(2^{n-1}\omega\right)\right|^2}{2} \prod_{r=0}^{n-2} \frac{\left|H_0\left(2^r\omega\right)\right|^2}{2} \tag{3.9}$$

The spectrum in (3.8) is a multilevel approximation of a $1/f$ behavior as depicted in Figure 3.3. The accuracy of the approximation depends on the flatness

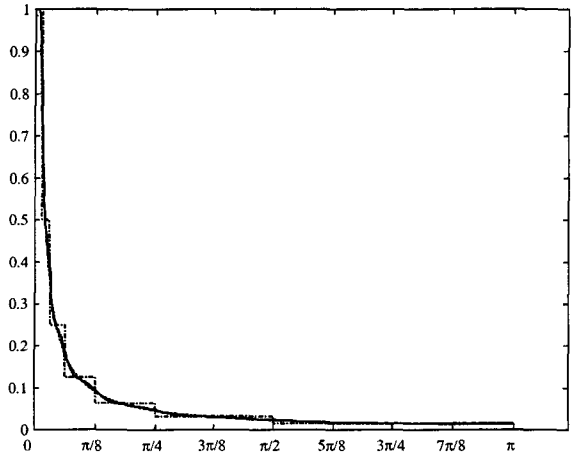


Figure 3.3: $1/f$ -like noise: ideal spectral behavior (dashed line), synthesized by means of Daubechies wavelet (solid line) and synthesized by ideal bandpass wavelets (dashed-dotted line).

and on the order of the filters. In the case where H_0 and H_1 are ideal filters, i.e. for

$$H_0(\omega) = \begin{cases} \sqrt{2} & if \ \frac{-\pi}{2} < \omega < \frac{\pi}{2} \\ 0 & otherwise \end{cases}$$

and

$$H_1(\omega) = \sqrt{2} - H_0(\omega)$$

we obtain from (3.9):

$$\frac{|\Psi_{n,0}(\omega)|^2}{2^n} = \begin{cases} 1 & for \ \ \omega \in \left[\left(1 - \frac{2^n-1}{2^n}\right)\pi, \left(1 - \frac{2^{n-1}-1}{2^{n-1}}\right)\pi\right] \\ 0 & otherwise \end{cases}$$

which yields a synthesized average power spectrum:

$$\overline{S}_N(\omega) = 2^{N\gamma}\chi_{\left[0,\left(1-\frac{2^N-1}{2^N}\right)\pi\right]}(\omega) + \sum_{n=1}^{N} 2^{n\gamma}\chi_{\left[\left(1-\frac{2^n-1}{2^n}\right)\pi,\left(1-\frac{2^{n-1}-1}{2^{n-1}}\right)\pi\right]}(\omega), \qquad \omega > 0$$

where

$$\chi_{[0,1]}(\omega) = \begin{cases} 1 & for \ \ 0 \le \omega \le 1 \\ 0 & otherwise \end{cases}$$

This corresponds to the staircase function shown in Figure 3.3.

While staircase approximation using octave band ideal filters has a pure demonstrative value, the approximation obtained by means of easily implementable Daubechies' filters is a very accurate one, as shown in Figure 3.3.

The self-similarity relation (3.6) does not carry over to discrete-time since the invariance for scale is only approximately true in that case.

3.3 Modulation scheme and pseudo-periodic $1/f$ model

In this section we consider a general modulation and demodulation scheme that leads to a useful representation of pseudo-periodic processes. Based on this scheme, we provide a definition of pseudo-periodic $1/f$-like noise suitable for the synthesis and the analysis of voiced sounds [62] [63].

3.3.1 Harmonic-band modulation and demodulation

The frequency spectra of pseudo-periodic signals are characterized by harmonically spaced peaks at frequencies $\widehat{\omega}_k = \frac{2\pi k}{T_P}$, where T_P is the average period of the signal. In order to separate the contribution of each of the harmonic bands, one can devise a set of ideal narrow-band filters of bandwidth $\Delta\omega = \frac{\pi}{T_P}$ each fitting a single sideband of the harmonics (see Figure 3.4). The magnitude of the Fourier transform of these filters is given by:

$$G_p(\omega) = \begin{cases} \chi_{\left[\frac{p\pi}{T_P}, \frac{(p+1)\pi}{T_P}\right[}(\omega) & p \geq 0 \\ \chi_{\left]\frac{p\pi}{T_P}, \frac{(p+1)\pi}{T_P}\right]}(\omega) & p < 0 \end{cases},$$

where $p = 0, \pm 1, \pm 2, ...$, and

$$\chi_{[A,B[}(\omega) = \begin{cases} 1 & if \quad A \leq \omega < B \\ 0 & otherwise \end{cases}$$

is the characteristic function of the interval [A,B[. In our notation, the positive frequency right sideband $R^+$of the k^{th}harmonic corresponds to the band indexed by $p = 2|k|$. Its negative frequency companion, which we still denote as the right sideband R^-, is the band indexed by $p = -2|k| - 1$. Similarly, positive and negative left sidebands, L^+ and L^-, are indexed, respectively, by $p = 2|k| - 1$ and $p = -2|k|$. Notice that for the d.c. component ($k = 0$) the bands R^- and L^-, respectively, coincide with the bands L^+ and R^+. The outputs of these filters may be baseband shifted, according to a suitable demodulation scheme. In dealing with real signals, it is convenient to combine positive and negative frequencies in such a way that the resulting component signal is still real. This is achieved by demodulating the output of each filter by the frequency of the corresponding harmonics, i.e. by multiplying the band p signal by

$$\frac{1}{2} e^{-j\left(\left\lceil \frac{p}{2} \right\rceil \frac{2\pi}{T_P} t + \beta_p\right)},$$

where $\beta_p = \beta_{-p-1}$ are otherwise arbitrary phase factors. We then add together the outputs of the demodulated R^+ and R^-, and those of the demodulated L^+ and L^-. This results in the demodulation scheme reported in Figure **??**. Considering couples of positive and symmetric negative bands, demodulation may be described as the projection $\langle K_p(t, \bullet), x(\bullet) \rangle$, $p = 0, 1, ...$, of a signal $x(t)$, where $\mathbf{K}_p$ is a set of real linear operators with kernels

$$K_p(t, \tau) = \frac{1}{T_P} \cos\left(\frac{t - (-1)^p (2p+1)\tau}{2T_P} \pi + \beta_p\right) \operatorname{sinc}\left(\frac{t - \tau}{2T_P}\right), \qquad p = 0, 1, ... \tag{3.10}$$

where the sinc function represents ideal lowpass filtering, properly baseband demodulated by the cosine term. The operators described by the kernels (3.10)

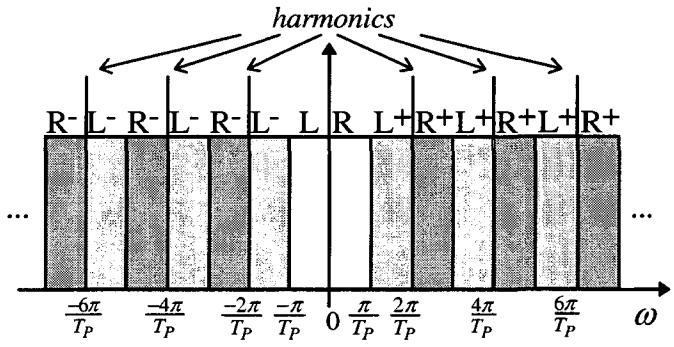


Figure 3.4: Harmonic sideband allocation.

perform a harmonic cosine demodulation to baseband of the signal subband with frequency support in

$$W_p \equiv \left]\frac{-(p+1)\pi}{T_P}, \frac{-p\pi}{T_P}\right] \cup \left[\frac{p\pi}{T_P}, \frac{(p+1)\pi}{T_P}\right[. \tag{3.11}$$

The presence of the constant phase factors β_p allows us to generalize the cosine demodulation scheme to other schemes, such as sine demodulation. We denote by $\mathbf{V}_p$ the $\mathbf{L}^2$ subspace of signals bandlimited to W_p . The operator K_p defines an isomorphism $\mathbf{V}_p \leftrightarrow \mathbf{V}_0$, where $\mathbf{V}_0$ is the space of bandlimited baseband signals, with frequency support in $\left]-\frac{\pi}{T_P}, \frac{\pi}{T_P}\right[$. In fact, one can verify that K_p is invertible, with inverse kernel $K_p^{-1}(t,\tau) = K_p(\tau,t) = K_p^{\dagger}(t,\tau)$, where the symbol $\dagger$ denotes the adjoint. Hence K_p is unitary. Conversely the operators K_p^{-1} perform a harmonic cosine modulation, repositioning the demodulated subband to the domain given by (3.11). It should be noted that, unless $p = 0$, domain and range space of the operator are different. Thus, K_p and K_p^{-1} do not commute, rather the domain and range space of $K_p^{-1}K_p$ is $\mathbf{V}_p$, while the domain and range space of $K_pK_p^{-1}$ is $\mathbf{V}_0$. Also, the identity operator in $\mathbf{V}_p$ has kernel $I_p(t,\tau) = \frac{1}{T_P}\cos\left(\frac{(2p+1)\pi(t-\tau)}{2T_P}\right)\mathrm{sinc}\left(\frac{t-\tau}{2T_P}\right)$, which, for $p = 0$, corresponds to $I_0(t,\tau) = \frac{1}{T_P}\mathrm{sinc}\left(\frac{t-\tau}{T_P}\right)$

Harmonic cosine modulation and demodulation is the main ingredient of our representation and formal definition of pseudo-periodic signals.

3.3.2 Pseudo-periodic $1/f$-like noise: a rigorous definition

We model acoustic pseudo-periodic signals with fundamental frequency $f_0 = \omega_0/2\pi$ by means of a superposition of cosine modulated bandlimited $1/f$ processes. Each process contributes to a single side band of one of the harmonics of a pseudo-periodic signal. Each one of these $1/f$ processes is characterized by two parameters σ and γ (see equation (3.5)). The parameter σ controls the global energy of the process, while the parameter γ controls the slope of the spectral curve. In the pseudo-periodic case we denote each harmonic partial by means

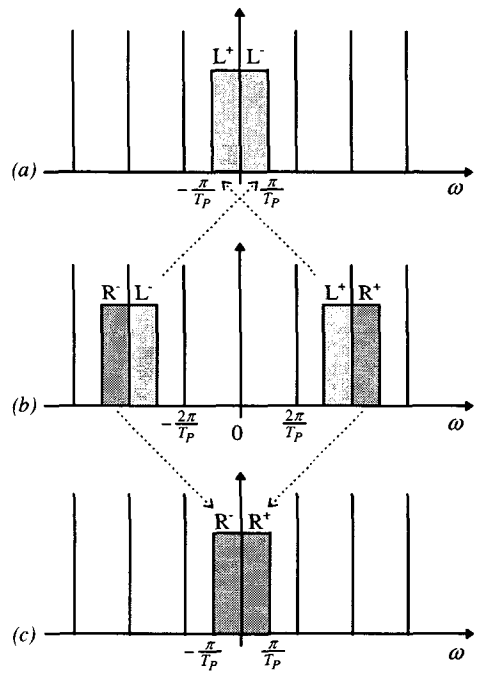


Figure 3.5: Baseband shift of harmonic sidebands: (*b*) sidebands of the 2^{nd} harmonics; (*a*) demodulation of the left sidebands; (*c*) demodulation of the right sidebands.

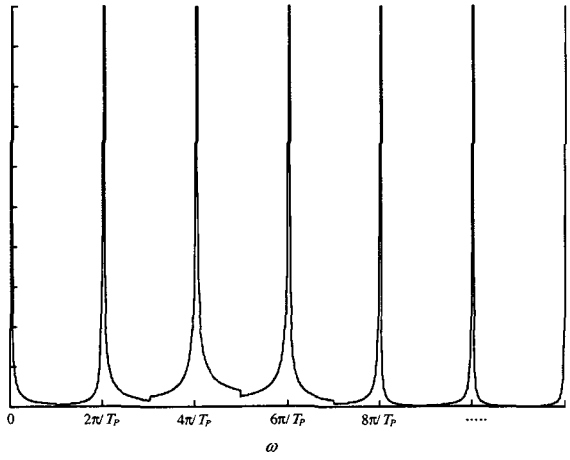


Figure 3.6: Pseudo-periodic $1/f$-like power spectrum. Each modulated $1/f$ process has bandwidth π/P. All the processes have same σ but different γ.

of the index k and we distinguish between the left and right sideband by means of the indexes L and R, respectively. We obtain a set of parameters $\sigma^2_{k,R}$ and $\sigma^2_{k,L}$,corresponding to the amplitudes of the side bands of the harmonics k and a set of parameters $\gamma_{k,R}$ and $\gamma_{k,L}$ controlling the slope of their $1/f$-like spectra.

An example of ideal pseudo-periodic $1/f$ spectrum is shown in Figure 3.6. In this case the parameter σ is the same for each process, while the parameter γ, i.e. the slope changes from process to process. The modulating frequencies are chosen to be harmonically related. The bandwidth B of each process equals half the harmonic spacing, i.e. $B = \omega_0/2 = \pi/T_P$.

In other words, the average spectrum of the model process has the following form[1]:

$$S(\omega) = \sum_k \frac{\sigma^2_{k,R}}{|\omega - k\omega_0|^{\gamma_{k,R}}} \chi_{[k\omega_0,\ (k+1/2)\omega_0[}(\omega) + \frac{\sigma^2_{k,L}}{|\omega - k\omega_0|^{\gamma_{k,L}}} \chi_{[(k-1/2)\omega_0,\ k\omega_0[}(\omega), \qquad \omega \geq 0 \tag{3.12}$$

We can provide a formal definition of pseudo-periodic $1/f$ –like noise that extends that of the $1/f$ noise given in [6]. In fact, if each sideband of the harmonics is baseband shifted by means of cosine demodulation, as described in the previous Section, the resulting process is $1/f$, bandlimited to $[-\omega_0/2, \omega_0/2]$. This is equivalent to say that, by passing the demodulated component processes through an ideal bandpass filter:

$$G^{(\epsilon)}(\omega) = \chi_{[-\omega_0/2, -\epsilon]}(\omega) + \chi_{[\epsilon, \omega_0/2]}(\omega)$$

[1] A notation remark: in the following equation and for the rest of this work a sum with unspecified boundaries denotes that the index runs from -∞ to +∞.

with ϵ arbitrarily small, one obtains a finite-variance wide-sense stationary process. This is actually the main idea of the definition of $1/f$ noise in Section 3.1. Therefore, we can provide the following:

Definition 3.2 *A stochastic process $x(t)$ is said to be a $1/f$-like pseudo-periodic noise if there exists a $T_P > 0$ such that when $x(t)$ is operated by K_p in (3.10) it yields a collection of processes*

$$x_p(t) = \int_{-\infty}^{\infty} K_p(t,\tau)x(\tau)d\tau, \qquad p = 0, 1, \ldots , \tag{3.13}$$

which, when filtered through $H^{(\epsilon)}(\omega)$, with $\omega_0 = \frac{2\pi}{T_P}$, become wide-sense stationary and bandlimited processes with power spectrum

$$S_{x_p}(\omega) = \begin{cases} \sigma_p^2/\,|\omega|^{\gamma_p} & if \quad \epsilon < |\omega| < \omega_0/2 \\ 0 & otherwise \end{cases} \tag{3.14}$$

for some γ_p and σ_p.

The operations involved in (3.13) are equivalent to filtering the single sidebands of each of the harmonics, separately for the positive and negative frequencies, and properly baseband shifting the result. The phase factors β_p in (3.10) are arbitrary. It can be shown that the power spectrum $S_{w_p}(\omega)$ does not depend on the choice of β_p. Similarly, any signal generated by harmonic modulation of a $1/f$ baseband process with arbitrary phase yields a $1/f$ process when demodulated by means of (3.13). This is true even if the phase factors do not coincide. Therefore, our definition is consistent.

Comparing (3.14) with the model spectrum in (3.12), we can make the following associations:

$$\gamma_{p_{odd}} = \gamma_{2k-1} = \gamma_{k,L}, \qquad \gamma_{p_{even}} = \gamma_{2k} = \gamma_{k,R}$$

and

$$\sigma_{p_{odd}} = \sigma_{2k-1} = \sigma_{k,L}, \qquad \sigma_{p_{even}} = \sigma_{2k} = \sigma_{k,R}$$

Since the resulting processes $x_p(t)$ in *Definition 3.2* are bandlimited to $[-\omega_0/2, \omega_0/2]$, they can be sampled with sampling rate $\frac{\omega_0}{2\pi} = \frac{1}{T_P}$. It can be shown that the operations in (3.13) followed by sampling at a rate $1/T_P$ are equivalent to the projection of $x(t)$ on the set of functions $\{g_{p,r}(t)\}_{p=0,1,\ldots;r\in Z}$, defined as follows:

$$g_{p,r}(t) = g_{p,0}(t - rT_P) \tag{3.15}$$

with

$$g_{p,0}(t) = \frac{1}{\sqrt{T_P}} \cos\left(\frac{2p+1}{2T_P}\pi t\right) \operatorname{sinc}\left(\frac{t}{2T_P}\right) \tag{3.16}$$

The set in (3.16) is easily shown to form an orthonormal basis. The functions $g_{p,0}(t)$ are the impulse responses of ideal bandpass filters, with passband (3.11), that is, the $\operatorname{sinc}\left(\frac{t}{2T_P}\right)$ ideal lowpass filter with passband $\left]-\frac{\pi}{T_P}, \frac{\pi}{T_P}\right[$, modulated to the band (3.11) by the cosine function. It is clear that the coefficient obtained by projecting a signal $x(t)$ on a basis element $g_{p,r}(t)$, where r corresponds to the time rT_P and p corresponds to the band (3.11), is just the sample at time rT_P of the component of $x(t)$ bandlimited to (3.11), i.e. $\langle x, g_{p,r}\rangle = \sqrt{T_P}x_p(rT_P)$.

3.4 Synthesis of pseudo-periodic $1/f$ noise by means of HBWT

In order to extend the result of Section 3.2 to the synthesis of $1/f$ pseudo-periodic processes by means of wavelet bases, and to introduce their discrete-time counterpart, we need the following:

Lemma 3.2 *A stochastic process $x(t)$ defined as follows:*

$$x(t) = \sum_{p=0}^{\infty} \sum_{r=-\infty}^{\infty} \nu_p(r) g_{p,r}(t) \tag{3.17}$$

where the $g_{p,r}(t)$ are given in (3.15) and $\{\nu_p(r)\}$ are jointly stationary discrete-time stochastic processes, i.e. $R_{\nu_p,\nu_{p'}}(r,r') = R_{\nu_p,\nu_{p'}}(r-r')$, is wide sense cyclostationary (WSCS) with period T_P.

Proof. We have

$$R_x(t+kT_P, t'+kT_P) = \\ = E\left\{\left(\sum_{p=0}^{\infty}\sum_{r=-\infty}^{\infty} \nu_p(r) g_{p,r}(t+kT_P)\right)\left(\sum_{p'=0}^{\infty}\sum_{r'=-\infty}^{\infty} \nu_{p'}(r') g_{p',r'}(t'+kT_P)\right)\right\}$$

which, by making the substitutions $r'-r=l$ and $k-r=r''$ and using (3.15) becomes

$$R_x(t+kT_P, t'+kT_P) = \\ = \sum_{p,p'=0}^{\infty}\sum_{l,r''=-\infty}^{\infty} g_{p,0}(t-r''T_P) g_{p',0}(t'-(l+r'')T_P) R_{\nu_p,\nu_{p'}}(l) = R_x(t,t')$$

which does not depend on r. ■

We now introduce a continuous time multiwavelet basis forming the Harmonic-Band Wavelet set, i.e. the HBWT:

$$\xi_{n,m,p}(t) = \sum_r \psi_{n,m}(r) g_{p,r}(t) \tag{19} \tag{3.18}$$

where $n \in N$, $m \in Z$, $p = 0, 1, ...P$, $P \in N$ and $\{\psi_{n,m}(r)\}_{n\in N, m\in Z}$ is an ordinary discrete-time wavelet basis while $g_{p,r}(t)$ are defined in (3.15).

The Fourier transform of the HBWT corresponds to comb versions of ordinary wavelets, filtered by the filterbank with frequency responses $G_{p,0}(\omega)$:

$$\Xi_{n,m,p}(\omega) = \Psi_{n,m}(T_P\omega) G_{p,0}(\omega) \tag{3.19}$$

In (3.19) $G_{p,0}(\omega)$ is the Fourier transform of $g_{p,0}(t)$ and $\Psi_{n,m}(T_P\omega)$ is the Fourier transform of a comb wavelet [19]. This means that we have infinite comb wavelets, one for each p, and that the action of filtering is essentially equivalent to selecting a single sideband of the harmonics. What we have obtained is to wavelet transform each single sideband independently.

Furthermore, the harmonic-band wavelets (3.15) satisfy the following shift property:

$$\xi_{n,m,p}(t+2^N rT_P) = \xi_{n,m-2^{N-n}r,p}(t) \tag{3.20}$$

Consider the case where the $\{\nu_p(r)\}$ in (3.17) are WSCS processes with period $2^N P$ (see also (3.7)) defined as follows:

$$\nu_p(r) = \sum_{n=1}^{N} \sum_{m=-\infty}^{\infty} \beta_p^{n/2} \nu_p^n(m) \psi_{n,m}(r)$$

where the $\nu_p^n(m)$ are unit variance mutually uncorrelated coefficients, while $\beta_p = \sigma_p^2 2^{\gamma_p}$ are scale dependent energy factors. We can prove the following:

Lemma 3.3 *A stochastic process $x(t)$ such that*

$$x(t) = \sum_{p=0}^{\infty} \sum_{r=-\infty}^{\infty} \nu_p(r) g_{p,r}(t) = \sum_{p=0}^{\infty} \sum_{n=1}^{N} \sum_{m=-\infty}^{\infty} \beta_p^{n/2} \nu_p^n(m) \xi_{n,m,p}(t),$$

where the $\xi_{n,m,p}(t)$ are defined in (3.18), is cyclostationary with period $2^N T_P$.

Proof. We have:

$$R_x\left(t+2^N rT_P, t'+2^N rT_P\right) =$$

$$= E\left\{ \left(\sum_{p=0}^{\infty} \sum_{n=1}^{N} \sum_{m=-\infty}^{\infty} \beta_p^{n/2} \nu_p^n(m) \xi_{n,m,p}(t+2^N rT_P) \right) \right.$$
$$\left. \left(\sum_{p'=0}^{\infty} \sum_{n'=1}^{N} \sum_{m'=-\infty}^{\infty} \beta_{p'}^{n'/2} \nu_{p'}^{n'}(m') \xi_{n',m',p'}(t'+2^N rT_P) \right) \right\}. \tag{3.21}$$

By using (3.15), the shift property (3.20) and the fact that $R_{\nu_{n,p},\nu_{n',p'}}(m,m') = \beta_p^n \delta_{n,n';p,p';m,m'}$, equation (3.21) becomes:

$$R_x\left(t+2^N rT_P, t'+2^N rT_P\right) = \sum_{p=0}^{\infty} \sum_{n=1}^{N} \sum_{m=-\infty}^{\infty} \beta_p^n \xi_{n,m-2^{N-n}r,p}(t) \xi_{n,m-2^{N-n}r,p}(t').$$

Finally, by the substitution $m' = m - 2^{N-n}r$ we obtain

$$R_x\left(t+2^N rT_P, t'+2^N rT_P\right) = \sum_{p=0}^{\infty} \sum_{n=1}^{N} \sum_{m'=-\infty}^{\infty} \beta_p^n \xi_{n,m,p}(t) \xi_{n,m,p}(t') = R_x(t,t'),$$

which is independent on r. ■

The same result holds for the scale residue in (3.7).

We are then able to derive the following result for the synthesis:

Proposition 3.4 *Consider an orthonormal set of functions* $\{g_{p,r}(t)\}_{p=0,1,\ldots;\ r\in Z}$, *as defined in (3.15) and a collection of jointly uncorrelated sets of coefficients* $\{\nu_p(r)\}_{p=0,1,\ldots}$, *related to (3.7) by the following relation:*

$$\nu_p(r) = \frac{1}{\sqrt{T_P}}x(r) = \frac{1}{\sqrt{T_P}}\left(\sum_{n=1}^{N}\sum_{m=-\infty}^{\infty} b_{p,n}(m)\psi_{n,m}(r) + \sum_{m=-\infty}^{\infty} a_{p,N}(m)\phi_{N,m}(r)\right) \tag{3.22}$$

where $\{b_{p,n}(m)\}$ *and* $\{a_{p,N}(m)\}$ *are jointly uncorrelated WSS white noise processes with variances* $Var\{b_{p,n}(m)\} = \sigma_p^2 2^{n\gamma_p}$ *and* $Var\{a_{p,n}(m)\} = \sigma_p^2 2^{N\gamma_p}$. *Then the random process*

$$s(t) = \sum_{p=0}^{\infty}\sum_{r=-\infty}^{\infty} \nu_p(r) g_{p,r}(t)$$

has an average power spectrum of the form:

$$\overline{S}(\omega) = \frac{1}{T_P}\sum_{p=0}^{\infty}\sigma_p^2 \left|G_{p,0}(\omega)\right|^2 \left(\sum_{n=1}^{N} 2^{n\gamma_p}\left|\Psi_{n,0}(\omega T_P)\right|^2 + 2^{N\gamma_p}\left|\Phi_{n,0}(\omega T_P)\right|^2\right) \tag{3.23}$$

For the proof see Appendix at the end of the chapter.

In the ideal case $\left|G_{p,0}(\omega)\right|^2 = \left(\chi_{\left]\frac{-(p+1)\pi}{T_P}, \frac{-p\pi}{T_P}\right]}(\omega) + \chi_{\left[\frac{p\pi}{T_P}, \frac{(p+1)\pi}{T_P}\right[}(\omega)\right)$ and (3.23) is approximately $1/f$ near each harmonic $k\frac{2\pi}{T_P}$ with $k = \left\lfloor\frac{p+1}{2}\right\rfloor$, $p = 0, 1, \ldots$. That is for

$$(2k-1)\frac{\pi}{T_P} \le \omega \le 2k\frac{\pi}{T_P} \qquad \text{if } p \text{ is odd (right sideband)}$$

or

$$2k\frac{\pi}{T_P} \le \omega \le (2k+1)\frac{\pi}{T_P} \qquad \text{if } p \text{ is even (left sideband)}$$

we have

$$\frac{\sigma_{L,p}^2}{\left|\omega - 2k\frac{\pi}{T_P}\right|^{\gamma_p}} \le \overline{S}_N(\omega) \le \frac{\sigma_{U,p}^2}{\left|\omega - 2k\frac{\pi}{T_P}\right|^{\gamma_p}}$$

for some $0 < 2\sigma_{L,p} \le 2\sigma_{U,p} < \infty$.

It follows from Proposition 3.4 that one can synthesize a signal with an approximately pseudo-periodic $1/f$ –like behavior, as shown in Figure 3.7. The inverse Harmonic-Band Wavelet Transform with random coefficients is used as a synthesis scheme for pseudo-periodic $1/f$–like noise. We are able to simulate a real-life pseudo-periodic signal with arbitrary pitch P. The parameters necessary to define the behavior of each sideband are only two: σ_p and γ_p. They control respectively the amplitude and the slope of the sideband spectrum.

3.5 Discrete-time harmonic-band wavelets

The discrete-time counterpart of (3.15) is the basis associated with an ideal P band filter bank, where P is the length in samples of the period of the pseudo-periodic signal. In order to obtain an efficient scheme for the analysis and synthesis of pseudo-periodic $1/f$ noise we consider an approximation of the

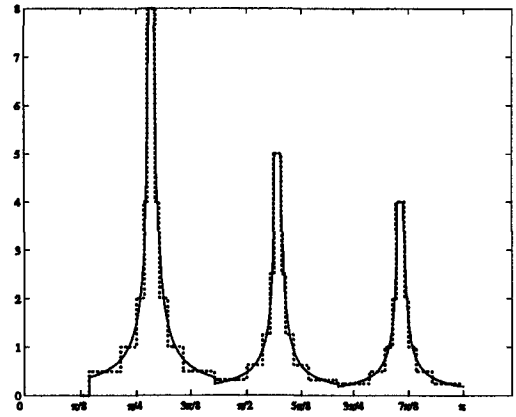


Figure 3.7: Synthesized pseudo-periodic $1/f$ -like noise: three harmonics with different σ_p and γ_p. a) solid line: ideal spectrum behavior b) dotted line: synthesis by means of ideal filter banks.

ideal filter bank by a perfect reconstruction structure [89]. In particular, we consider the Modified Discrete Cosine Transform basis (MDCT):

$$g_{p,r}(l) = g_{p,0}(l - rP) \qquad p = 0, ..., P-1;\ r \in Z, \tag{3.24}$$

with

$$g_{p,0}(l) = w(l) \cos\left(\frac{2p+1}{4P}(2l - P + 1)\pi\right), \tag{3.25}$$

where the length $2P$ lowpass prototype impulse response $w(l)$ satisfies the symmetry conditions given in [53]. That is:

$$w(l) = w(2P - l - 1) \qquad \text{for} \qquad l = 0, ..., 2P-1 \tag{3.26}$$

$$w^2(l) + w^2(P - l - 1) = 2 \qquad \text{for} \qquad l = 0, ..., P-1 \tag{3.27}$$

$$w(l) = 0 \qquad \text{for} \qquad l < 0,\ l > 2P - 1$$

The magnitude FT of the Discrete-Time Harmonic-Band Wavelet (DT-HBWT) is shown in Figure 3.9.

In order to synthesize the samples of $1/f$ -like processes $x_p(l)$ we adopt the discrete-time version of the method illustrated in Section 3.2. The overall structure is realized by introducing the DT-HBWT. The synthesis of $1/f$ pseudo-periodic noise is achieved by using white noise coefficients. The DT-HBW are defined by:

$$\xi_{n,m,p}(l) = \sum_r \psi_{n,m}(r) g_{p,r}(l) \qquad n = 1, 2, ...N;\ m \in Z;\ p = 0, 1, ..., P-1 \tag{3.28}$$

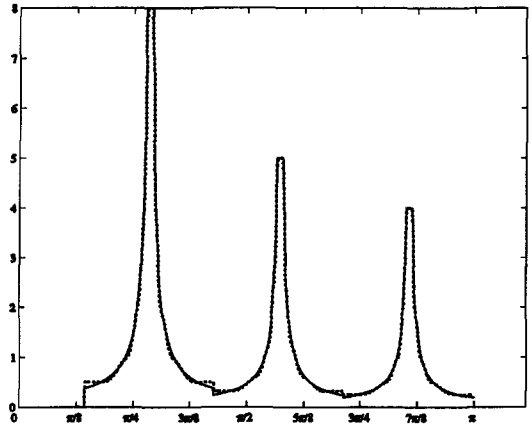


Figure 3.8: Synthesized pseudo-periodic $1/f$ -like noise: three harmonics with different σ_p and γ_p. a) solid line: ideal spectrum behavior b) dotted line: synthesis by means of MDCT and Daubechies Wavelet.

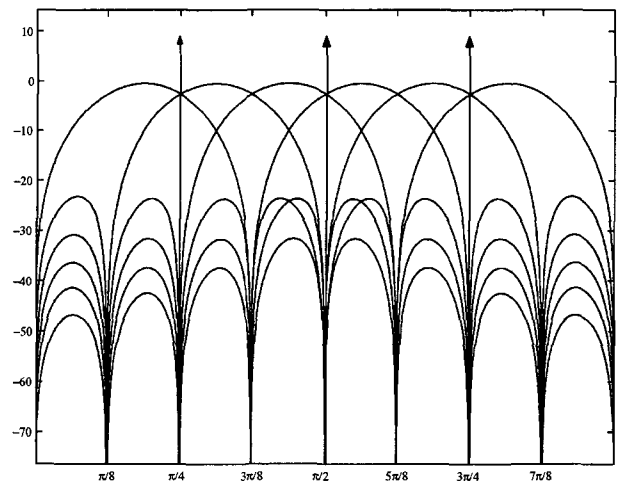


Figure 3.9: Magnitude Fourier transform of a 8 channel MDCT. Only the channels 2-7 are shown and the position of the related three harmonic peaks.

where $\psi_{n,m}(r)$ are discrete-time ordinary wavelets and $g_{p,r}(l)$ are the MDCT functions (3.24). The corresponding scale residue function is given by:

$$\zeta_{N,m,p}(l) = \sum_r \varphi_{N,m}(r) g_{p,r}(l) \qquad m \in Z;\ p = 0, 1, ..., P-1 \quad (3.29)$$

The conditions of orthogonality and completeness of the DT-HBW

$$\sum_{l=-\infty}^{\infty} \xi_{n,m,p}(l)\xi_{n,m,p'}(l) =$$

$$= \sum_{l=-\infty}^{\infty} \left(\sum_r \psi_{n,m}(r) g_{p,r}(l) \sum_{r'} \psi_{n,m}(r') g_{p',r'}(l) \right) = \delta_{r,r'}\delta_{p,p'}$$

and

$$\sum_p \xi_{n,m,p}(l)\xi_{n,m,p}(l') == \sum_p \sum_r \left(\psi_{n,m}(r) g_{p,r}(l) \psi_{n,m}(r) g_{p,r}(l') \right) = \delta_{l,l'},$$

respectively, follow from the orthonormality and completeness of the MDCT and the WT. Thus any signal $s(l) \in l^2$ can be expanded on a DT-HBW set according to the:

$$s(l) = \sum_{p=1}^{P} \left(\sum_{n=1}^{N} \sum_m b_{p,n}[m]\xi_{n,m,p}(l) + \sum_m a_{p,N}[m]\zeta_{N,m,p}(l) \right), \quad (3.30)$$

where the $b_{p,n}[m]$'s and the $a_{p,N}[m]$'s are the DT-HBW expansion coefficients and the corresponding harmonic-band scale residue coefficients at scale N, respectively. The DTFT of the basis elements (3.28) are shown in Figure 2.14. A structure for computing the DT-HBWT and its inverse is shown in Figure 3.10 and 3.11, respectively.

In the analysis structure, the signal is sent to a P-channel filter bank separating the sidebands. In view of perfect reconstruction, the output can then be downsampled by P. Each P-downsampled signal is then wavelet transformed. Signal reconstruction is achieved by separately inverse wavelet transforming the HBWT analysis coefficients and passing these sequences through the inverse P-channel filter bank with upsampling factor P. Upsampling moves the spectrum of each subsignal back to its proper subband.

The HBWT generalizes the Multiplexed Wavelet Transform (MWT), recently introduced in [19], to which they revert when $g_{p,0}(l) = \delta(l)$, where $\delta(l)$ is the unit pulse sequence. Similarly to the MWT, the HBWT is useful for separating, for each harmonics, the sinusoidal behavior (scaling component) from transients and noise (wavelet components). The idea of our model is to employ the theoretical result of Proposition 3.4 in order to have an efficient scheme for synthetically reproducing the noisy sidebands of the spectrum of voiced sounds in music. Experimental results confirm the validity of the $1/f$ model for the sidebands in a wide class of musical pseudo-periodic signals. Thanks to the result of Proposition 3.4 we can easily reproduce the synthesis coefficients in our scheme by means of white noise or weakly correlated noise. The coefficient energy is controlled by few parameters (2 per each sideband) drawn from the analysis scheme of the signal. In the next chapter we illustrate the experimental results, validating the existence of a pseudo-periodic $1/f$-like behavior in many voiced sounds.

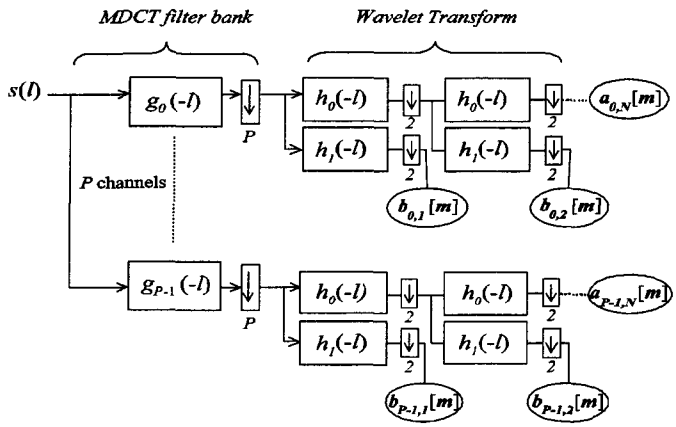


Figure 3.10: HBWT implementing scheme.

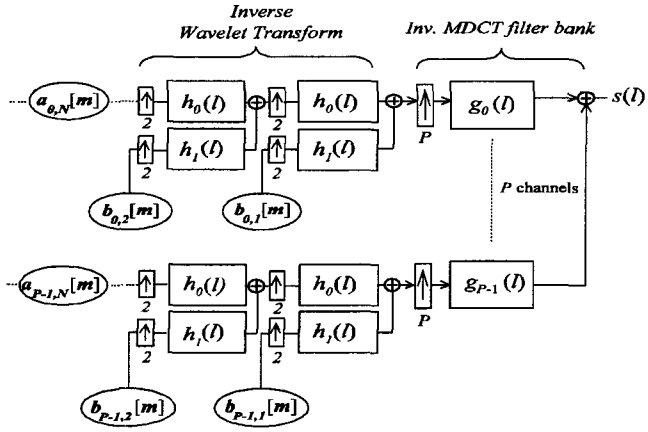


Figure 3.11: Inverse HBWT implementing scheme.

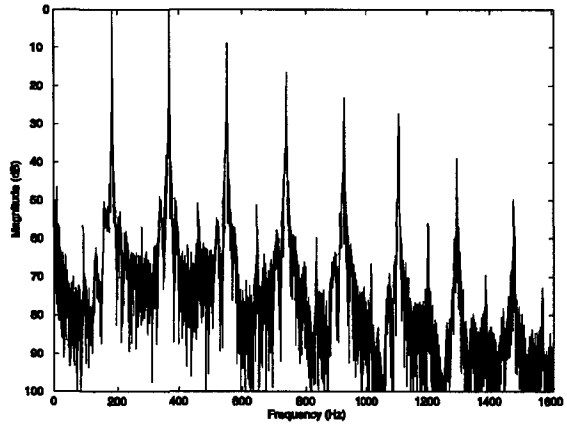


Figure 3.12: Magnitude FT of a French horn. Relevant spectral peaks different from the harmonic ones are detectable.

3.6 A refined spectral design via frequency-warped WT

By means of the HBWT, introduced in Section 3.4, the result concerning the synthesis of $1/f$-like processes by means of the WT was extended to the pseudo-periodic $1/f$-like case. We showed that it is possible to synthesize signals with pseudo-periodic $1/f$-like power spectra by employing white noise coefficients, with wavelet-band dependent energy. Our goal is to adapt the spectrum of the synthetic signal to that of a real-life pseudo-periodic sound.

Different levels of approximation can be achieved in the synthesis of noise components of pseudo-periodic sounds. A simple refinement consists in employing as resynthesis parameters the individual variances of the HBW subbands. This can be seen as a first acoustical refinement of the FAS method towards an acoustical refinement. This means to set the method free from the rigid $1/f$-like pseudo-periodic model increasing the cost for a higher quality sound reproduction. As illustrated in the next chapter, in order to do this we need to replace the parameters $2^{\frac{n}{2}\gamma_p}$ with a set of parameters $\sigma_{p,n}$ corresponding to the energies of the single n^{th} subband of the p^{th} sideband. Another refinement of the technique consists in setting our method free from the strict constraints of the $1/f$ model in order to obtain a better approximation of the spectrum shape. This can be achieved by employing the Frequency-Warped Wavelet Transform (FWWT), recently introduced in [21], [20]. We obtain an arbitrary segmentation of the frequency axis, i.e. of the wavelet analysis and synthesis bands. In this way we can reproduce the deviations of real spectra with respect to the strict pseudo-periodic $1/f$ -like model [61]. An example is shown in Figure 3.12, where relevant non-harmonic peaks are present in the spectrum of a French horn.

In this section we briefly review the FWWT. Then, we introduce the Harmonic-Band Frequency-Warped Wavelet Transform (HB-FWWT). The experimental

results are discussed in Chapter 4.

3.6.1 Frequency warping and Laguerre transform

The term warping denotes an operation of "distortion" of a function through some mapping of its domain. For signals represented in the frequency domain this mapping is represented by the notation:

$$\Omega = \theta(\omega)$$

where ω is the original frequency domain and θ is the frequency mapping. The frequency warped version of a signal $s(l)$ can be written in the frequency domain as:

$$S_w(\omega) = S(\theta(\omega)) = S(\Omega)$$

In the following paragraphs we consider the particular case of frequency warping via the Laguerre Transform (LT). The LT is the main element for defining the FWWT [20]. A frequency-warped version of the HBWT is then straightforward.

The Laguerre sequence is given by the sum:

$$\lambda_{r,d}(l) = \sqrt{1-d^2} \sum_{m=0}^{\min(r,l)} (-)^{m+r} \frac{(l+r-m)!}{m!\,(l-m)!\,(r-m)!} \cdot d^{r+l-2m},$$

whose z transform is:

$$\Lambda_{r,d}(z) = \sqrt{1-d^2} \frac{(z^{-1}-d)^r}{(1-dz^{-1})^{r+1}}. \tag{3.31}$$

The functions in (3.31) satisfy the following recursive relation:

$$\Lambda_{r,d}(z) = A(z)\Lambda_{r-1,d}(z) = A(z)^r \Lambda_{0,d}(z),$$

where

$$A(z) = \frac{z^{-1}-d}{1-dz^{-1}}$$

is the system function of a stable and causal allpass filter. On the unit circle $A(e^{j\omega}) = e^{-j\theta(\omega)}$, with

$$\theta(\omega) = -\arg A(e^{j\omega}) = \omega + 2\tan^{-1}\left(\frac{d\sin\omega}{1-d\cos\omega}\right). \tag{3.32}$$

The function (3.32) is the frequency warping mapping generated by the LT.

Since the Laguerre sequences are infinite length, the LT has to be approximated to a finite number of coefficients. It is possible to estimate [21] a minimum number M of Laguerre coefficients providing a good accuracy in the representation of a signal $s(l)$ of L samples:

$$M \geqslant \frac{L(1+|d|)}{1-|d|}. \tag{3.33}$$

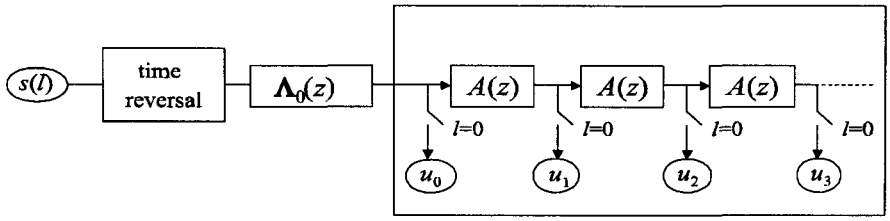


Figure 3.13: LT implementation scheme. The u_k form the LT of the signal $s(l)$.

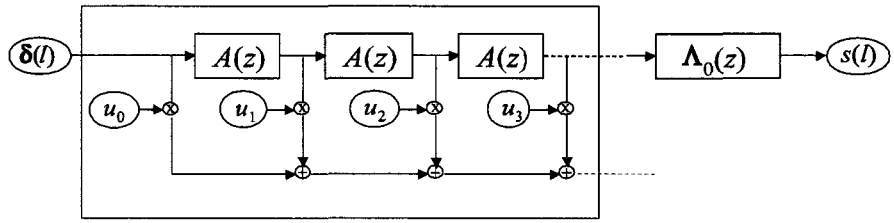


Figure 3.14: Inverse LT implementation scheme.

When (3.33) is strictly satisfied, the maximum error is less than the 5% of the maximum absolute value of the signal:

$$e_{\max} < 0.05 \cdot |\max(s(l))|$$

Figure 3.13 and 3.14 show the implementation structure for the LT and its inverse, respectively. The parts of the scheme included in the rectangles in the two figures form a switched dispersive delay line and a tapped dispersive delay line, which will be denoted, respectively, by the symbols of Figure 3.15 a) and b)

3.6.2 Frequency warped wavelets

Wavelet transforms are almost a useful tool for multiresolution analysis. Their most appealing feature is related to the non-uniform octave band subdivision of the space-frequency or time-frequency spaces in image and sound processing, respectively. The octave band subdivision as well as the principle of scale covariance seem to be successful from a perceptual point of view for both our hearing and visual system.

Figure 3.15: Graphic symbols for *a*) switched dispersive delay lines and *b*) tapped dispersive delay lines.

What we are trying to do here is to model the spectra of pseudo-periodic signals. The harmonic peaks of these spectra have an approximately $1/f$-like behavior, well fitting the power-of-2 observation perspective of the wavelet transform. Nevertheless this model can be improved and/or corroborated, if one makes the power-of-2 law more flexible and adaptable to real-life deviations from the model itself. This can be obtained by introducing in the FAS scheme the FWWT, i.e. a WT with an arbitrary non-uniform subdivision of the frequency axis replacing the octave-band subdivision. In the following paragraphs we give a concise review of the FWWT.

The frequency warped wavelets $\hat{\psi}_{n,m}(l)$ and their corresponding frequency warped scale sequences $\hat{\varphi}_{n,m}(l)$ obey, respectively, the following recursive relation:

$$\hat{\psi}_{n,m}(l) = \sum_{k=0}^{\infty} g_{n,m}(k)\hat{\varphi}_{n-1,k}(l)$$

and

$$\hat{\varphi}_{n,m}(l) = \sum_{k=0}^{\infty} h_{n,m}(k)\hat{\varphi}_{n-1,k}(l)$$

where the $g_{n,m}$ and the $h_{n,m}$. are some auxiliary sequences given by:

$$g_{n,m}(k) = \sum_{r=0}^{\infty} \lambda_{n,r}(k)h_1(r-2m)$$

and

$$h_{n,m}(k) = \sum_{r=0}^{\infty} \lambda_{n,r}(k)h_0(r-2m),$$

where the symbol $\lambda_{n,r}$ denotes a Laguerre sequence of order r associated to the n^{th} wavelet scale. The ordinary quadrature mirror filters h_1 and h_0 in this case play the role of coefficients of the Laguerre expansion of the functions $g_{n,m}$ and $h_{n,m}$ respectively.

The frequency warped wavelets form orthonormal and complete sets. For any $s(l) \in l^2$ $(\mathbf{N} \cup \{0\})$ it is possible to write:

$$s(l) = \sum_{n=1}^{N} \sum_{m} b_{n,m}\hat{\psi}_{n,m}(l) + \sum_{m} a_{N,m}\hat{\phi}_{N,m}(l)$$

Frequency warping or frequency axis deformation is obtained by means of the LT and is controlled subband n by subband n by the parameters d_n according to the following recurrence:

$$d_1 = \tan\left[(\pi - 2\omega_1)/4\right]$$

and

$$d_n = \tan\left[\frac{\pi}{4} - \Omega_{n-1}(\omega_n)\right],$$

where the ω_n are the arbitrary cut-off frequencies by which we subdivide the frequency axis, with $\omega_1 > \omega_2 > \ldots > \omega_n$ and the frequency mapping $\Omega_n(\omega)$ is given by [20]:

$$\Omega_n(\omega) = \theta_n(2\theta_{n-1}(\ldots 2\theta_2(\theta_1(\omega))\ldots))$$

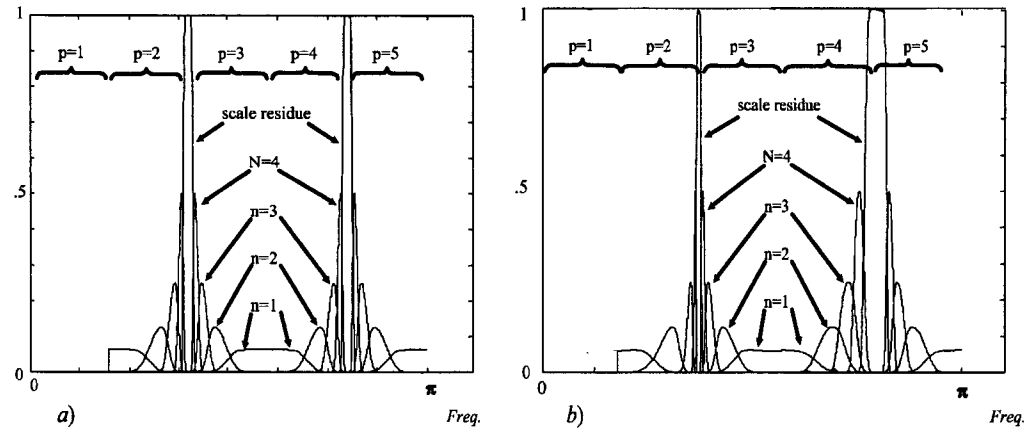


Figure 3.16: Magnitude frequency response of a filter bank implementing the ordinary HBWT, two harmonics (a), compared with the case of a HB-FWWT (b).

where, according to the (3.32), the $\theta_i(\omega)$ are:

$$\theta_i(\omega) = \omega + 2\tan^{-1}\left(\frac{d_i \sin\omega}{1 - d_i \cos\omega}\right)$$

for $i = 1, 2, ..., n$.

3.6.3 Harmonic-Band Frequency-Warped WT (HB-FWWT)

The DT-HBW introduced in Section 3.5 form an orthonormal and complete set in l^2. Computation of the DT-HBWT is achieved by means of a P-channel filter bank based on the MDCT cascaded by a WT of each channel. Any signal $s(l) \in l^2$ can be expanded on a DT-HBW set $\{\xi_{n,m,p}(l), \zeta_{N,m,p}(l)\}$ according to the (3.30), where the $\xi_{n,m,p}(l)$ and the $\zeta_{N,m,p}(l)$ are given in (3.28) and (3.29), respectively, with $p = 0, 1, ..., P-1$, $n = 1, ..., N$, $m \in \mathbb{Z}$. P is the number of channels and p is the channel index. In order to obtain a Frequency Warped version of the HBWT (HB-FWWT) we need simply to substitute ψ and ϕ in 3.22 with their warped version $\hat{\psi}$ and $\hat{\phi}$:

$$\nu_p(r) = \left(\sum_{n=1}^{N}\sum_{m} \hat{b}_{p,n}[m]\hat{\psi}_{n,m}(r) + \sum_{m} \hat{a}_{p,N}[m]\hat{\phi}_{N,m}(r)\right)$$

The great advantage is that each subband of each sideband can be adjusted by an optimization procedure, in order to fit any real-life spectrum of the kind, for instance, of Figure 3.12. In Figure 3.16b we show an example of how the frequency spectrum of Figure 3.16a can be modified by means of the FWWT. The position and bandwidth of each subband can be independently set by means of properly chosen parameters $\{d_{n,p}\}$ for each p and n.

At the same time we obtain a finer tool for verifying the pseudo-periodic $1/f$-like model on data. In fact by means of the HBWT analysis we obtain for each sideband p a set of parameters corresponding to the energies of the

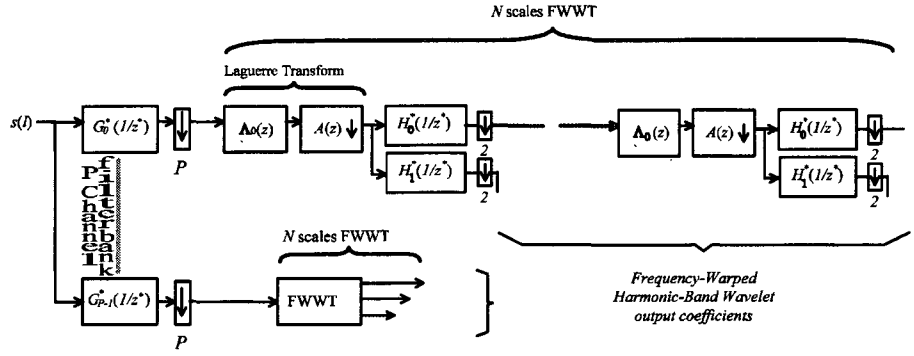


Figure 3.17: HB-FWWT analysis filter banks. $A \downarrow$ is a switched dispersive delay line implemented by a cascade of all-pass filters (see Figure 3.15a).

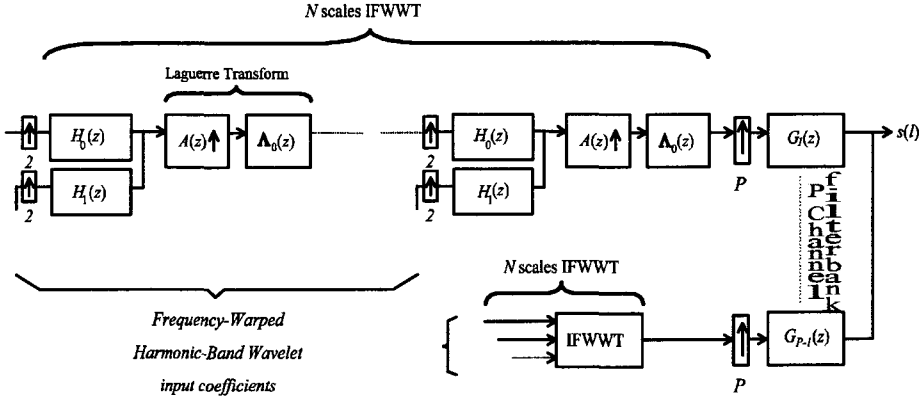


Figure 3.18: HB-FWWT synthesis filter bank. $A \uparrow$ is a tapped dispersive delay line implemented by a cascade of all-pass filters (see Figure 3.15b).

subbands. Each parameter is a point of the hypothetical $1/f$-like spectrum of the sideband p. By subdividing the latter with a finer resolution we have more points at our disposal, by which we can test the validity of the pseudo-periodic $1/f$-like model. The experimental results confirm the validity of the pseudo-periodic $1/f$-like model and are discussed in the next chapter. In Figure 3.17 and 3.18 we show the filter bank scheme, implementing the HB-FWWT and its inverse, respectively.

From a coding point of view, using the HB-FWWT with three scale levels implies a growth of the number of parameters of approximately a factor 3, but we still obtain a very good coding rate. Also, as it will be illustrated in the next chapter, many improvements can be obtained in terms of coding rate by means of parameters and coefficient modeling and the introduction of psychoacoustic criteria in the method.

3.7 Summary

In this chapter we introduced a new method for sound synthesis that allows us to control and reproduce the microfluctuations present in real life voiced sounds. This method is a sort of additive synthesis where one adds not only the harmonics but also modulated $1/f$ signals. We defined a new class of stochastic processes, i.e. the pseudo-periodic $1/f$ -like noise. We introduced a new type of multiwavelet transform useful for the representation of these processes, the Harmonic-Band Wavelet Transform (HBWT). We devised an efficient analysis/synthesis scheme able to generate pseudo-periodic $1/f$ -like noise.

The claim of this method is that it allows for reproduction of the stochastic fluctuations in sounds by means of a very restricted number of parameters. A better spectral modeling can be obtained by introducing energy parameters independently for each wavelet scale and by means of a Frequency Warping version of the HBWT (HB-FWWT).

In order to both improve the acousitcal results and at the same time make the method interesting in terms of coding rate, a further insight into the behavior of the HBWT coefficients has to be achieved. This is the subject of the next two chapters, where the results of HBWT coefficient modeling for voiced sounds with stable pitch P are discussed and then extended to voiced sounds with variable pitch and to inharmonic sounds. As we will see, these results make the Fractal Additive Synthesis (FAS) appealing both for audio coding and for sound synthesis and processing.

Appendix

We prove *Proposition* 3.4 of Section 3.4. We form

$$s(t) = \sum_{p=0}^{\infty} \sum_{r=-\infty}^{\infty} \nu_p(r) g_{p,r}(t)$$

where

$$g_{p,r}(t) = g_{p,0}(t - rT_P)$$

with

$$g_{p,0}(t) = \frac{1}{\sqrt{T_P}} \cos\left(\frac{2p+1}{2T_P}\pi t\right) \operatorname{sinc}\left(\frac{t}{2T_P}\right)$$

and

$$\nu_p(r) = \frac{1}{\sqrt{T_P}} \left(\sum_{n=1}^{N} \sum_{m=-\infty}^{\infty} b_{p,n}[m]\, \psi_{n,m}(r) + \sum_{m=-\infty}^{\infty} a_{p,N}[m]\, \phi_{N,m}(r) \right)$$

From *Lemma* 3.3 we know that $s(t)$ is $2^N T_P$-ciclostationary. Then the time-average power spectrum of the process $s(t)$ is:

$$\overline{S}(\omega) = \int_{-\infty}^{\infty} \overline{R_s}(\tau) e^{-j\omega\tau} d\tau = \int_{-\infty}^{\infty} \frac{d\tau}{2^N T_P} e^{-j\omega\tau} \int_{-\frac{2^N T_P}{2}}^{\frac{2^N T_P}{2}} R_s(t, t+\tau) dt,$$

which can be written as follows:

$$\overline{S}(\omega) =$$

$$= \frac{1}{2^N T_P} \int_{-\frac{2^N T_P}{2}}^{\frac{2^N T_P}{2}} dt \int_{-\infty}^{\infty} d\tau \sum_{p,p'=0}^{\infty} \sum_{k,k'=-\infty}^{\infty} L_{k,k';r,r'}(t,\tau) R_{\nu_p}(r, r'+2^N(k'-k)) e^{-j\omega\tau},$$

where

$$L_{k,k';r,r'}(t,\tau) = \sum_{r,r'=0}^{2^N-1} g_{p,0}(t - rT_P - 2^N k T_P) g_{p',0}(t+\tau - r'T_P - 2^N k' T_P).$$

The trick of the proof is to exploit the 2^N WSCS of the $\nu_p(m)$ proved in *Lemma 3.2*, in order to transform the finite integral over t into an integral over $(-\infty, \infty)$ equal to $G^*_{p,0}(\omega)$, i.e. the complex conjugate of the Fourier transform of $g_{p,0}(t)$. After routine calculation we obtain:

$$\begin{aligned} \overline{S}(\omega) &= \frac{1}{2^N T_P} \sum_{p=0}^{\infty} |G_{p,0}(\omega)|^2 \sum_{k=-\infty}^{\infty} \sum_{r=0}^{2^N-1} R_{\nu_p}(r, r+k) e^{-jk\omega T_P} = \\ &= \frac{1}{T_P} \sum_{p=0}^{\infty} |G_{p,0}(\omega)|^2 \overline{S}_p(\omega T_P), \end{aligned}$$

where $\overline{S}_p(\omega\, T_P)$ is the time-average power spectrum of the nearly $1/f$ processes bandlimited and modulated to the band (3.11). The result in (3.8) concludes our proof.

Chapter 4

Encoding the Sound

The main goal of this chapter is the definition of a method for the extraction of a reduced set of parameters describing the behavior of the HBWT analysis coefficients of voiced sounds within the frame of the $1/f$ pseudo-periodic model. We show how the output coefficients $b_{p,n}[m]$ of the filters $h_1(l)$ in Figure 3.10, can be efficiently modeled in terms of energy scaled and filtered white noise coefficients. The energy scaling envelopes parameters and the AR filter coefficients provide the FAS parametric representation of the stochastic components of voiced sounds.

A second model for the output of the last of the filters $h_0(l)$ of each channel p in Figure 3.10, i.e., for the Harmonic-Band (HB) scale coefficients $a_{p,N}[m]$ corresponding to the deterministic part of voiced sounds is then defined. What the HB-scale coefficients provide is a pseudo-sinusoidal model and only slight corrections are necessary in case the pitch is not perfectly tuned with the MDCT filter bank. Due to the smoothness of the curves generated by these coefficients a polynomial interpolation is well suited to provide a parametric representation of the coefficients themselves. In this way we obtain a full method for the resynthesis of a real-life voiced sounds. The resynthesis process is driven by a set of perceptually meaningful parameters, deduced from the analysis of the real-life sound itself.

Also, we have studied the potentialities of the method in terms of data compression by considering a psychoacoustic approach. Section 4.5 discusses an evaluation of the method in terms of psychoacoustic criteria.

In the next two sections, before introducing the two models, some applicative and experimental considerations on the HBWT and the pseudo-periodic $1/f$ model are presented. In particular in Section 4.1 we consider the possibility of partial reconstruction and resynthesis of sounds offered by the HBWT. In other words the HBWT can be viewed also as a method for timbre hybridization via cross-synthesis and sound morphing. Section 4.2 provides an evaluation of the experimental consistency of the pure pseudo-periodic $1/f$ model without any further coefficient modeling consideration except for the result of Proposition 3.4.

4.1 HBWT as a sound decomposition tool

We start with a very simple consideration about the analysis possibilities of the HBWT in terms of "sound decomposition tool" with interesting analytic, perceptual and musical applications. Due to the peculiar frequency domain subdivision shown in Figure 4.1, allows one to perform a partial reconstruction of the signal. We can extract one of the N subbands from all the harmonics (a single wavelet-band n from all of the channels) or one specific subband of one specific sideband, (fixed n and p) as well as any other arbitrary combination of subbands. This provides several possibilities in terms of sound processing results. More precisely we can define the n^{th} noise subband as:

$$s_n(l) = \sum_{p=1}^{P} \sum_{m} b_{p,n}[m]\xi_{n,m,p}(l) \tag{4.1}$$

and the n^{th} noise subband of the p^{th} sideband as

$$s_{p,n}(l) = \sum_{m} b_{p,n}[m]\xi_{n,m,p}(l).$$

Also, we can separate the noisy components of a single harmonic sideband, i.e., the p^{th} noise sideband

$$s_p(l) = \sum_{n=1}^{N} \sum_{m} b_{p,n}[m]\xi_{n,m,p}(l).$$

In this way timbre hybridization is straightforward: we can realize any "mixture" of subbands coming from the analysis of different instruments. A very simple example can be obtained combining the reconstructed harmonic component

$$s_{har}(l) = \sum_{p=1}^{P} \sum_{m} a_{p,N}[m]\zeta_{n,m,p}(l)$$

of one instrument with the noise sidebands $s_n(l)$ of another one. This can be successfully employed as a new sound morphing technique [56], [24]. We obtained interesting results by an hybridization of the subbands of a trumpet with a bassoon and an oboe with a viola.

4.2 The pseudo-periodic $1/f$ abstract model

The results of Chapter 3 provide a method for the synthesis of the stochastic microfluctuations of the steady part of sounds. By means of HBWT and the pseudo-periodic $1/f$ model a separation of sounds in harmonic peaks and stochastic components is straightforward. The reconstructed harmonic components sound clearly poor and unnatural to our ear. The reproduction of the harshness of the stochastic components is essential to provide sound with a convincing, natural "flavor". The HBW subbands of each harmonic peak are well suited to represent the stochastic fluctuations with respect to the harmonic components.

The synthesis technique for the pseudo-periodic $1/f$–like model requires the estimation of three parameters per each harmonic partial k: σ_k, $\gamma_{k,R}$, $\gamma_{k,L}$. The

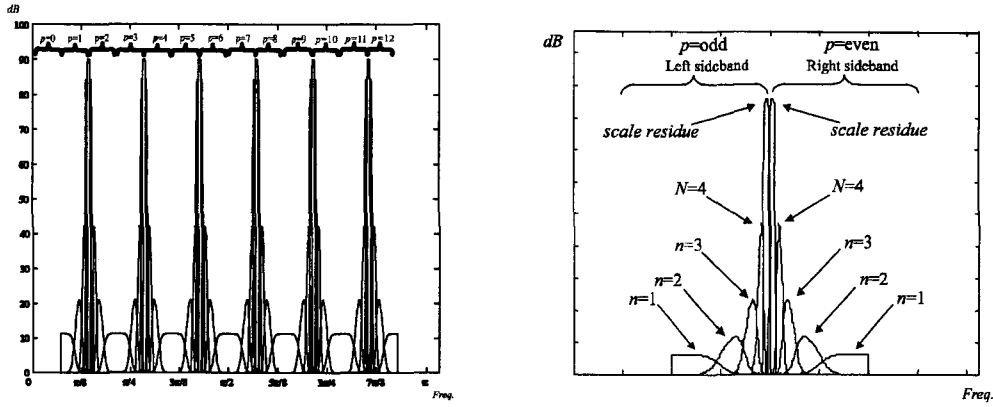


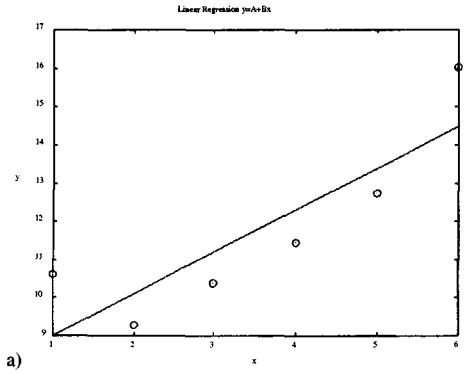
Figure 4.1: Magnitude FT of a HBW basis set, 12 channels a) and a detailed representation of two channels/one harmonic subband decomposition.

meaning of these parameters is intuitively appealing. The parameter σ_k controls the amplitude of the k^{th} harmonic, while the parameters $\gamma_{k,R} = \gamma_{2k} = \gamma_{p_{even}}$ and $\gamma_{k,L} = \gamma_{2k-1} = \gamma_{p_{odd}}$ control the $1/f$–like slopes of the right and left semiband of the k^{th} harmonic, respectively (see Section 3.3.2). The parameters σ_k can be estimated from the frequency spectrum by means of a peak-picking algorithm. The estimation of the parameters γ is based on the results of the HBWT analysis as follows. Each HBWT subband is a piecewise approximation of a $1/f$ spectral curve. Considering the logarithm of the energies of each of the subbands of a single sideband p, we find a linear relationship for the corresponding parameter γ_p. More specifically we perform the following linear regression:

$$\log_2(Var(b_{p,n}[m])) = \gamma_p n + const, \tag{4.2}$$

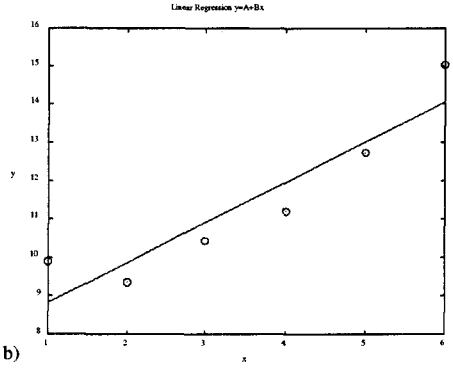
where $k = \lfloor \frac{p+1}{2} \rfloor$ is the k^{th} harmonic and $b_{p,n}[m] = \sum_l x(l)\xi_{n,m,p}(l)$ are the analysis sequences at the different subbands n. The lower values of the parameter γ there correspond higher energies of the stochastic components distributed in the subbands.

The first experiment we performed was a test on the limits of the pseudo-periodic $1/f$ model. We considered different wind instruments (clarinet, trumpet, oboe, bassoon) and bow instruments (cello and violin) and we applied (4.2) to the output coefficients of the HBW analysis. From the experimental results it is clear that not all of the sidebands of the harmonics are representable by a $1/f$–like model. In most cases the first wavelet subbands do not fit the model (see Figure 2.6, dark gray areas). These are the bands containing the extra noise due to the physical device of production of sound and noise from the recording equipment, or, more precisely, the bands where this type of noise is not masked. This type of noise is due, for instance, to breath noise in wind instruments or bow noise in string instruments. The energy of the noise falling in the first level subbands is generally higher than that provided by the $1/f$ slope. This additional noise is masked (but present) in the proximity of the harmonic peaks but it stands out in the first HBWT spectral subbands, where it overlaps the



a)

Figure 4.2: Estimation of the parameter γ: Linear regression result for the subband energies of a left sideband of one of the first harmonics of a trumpet.



b)

Figure 4.3: Estimation of the parameter γ: Linear regression result for the subband energies of a right sideband of one of the first harmonics of a trumpet.

pseudo-periodic $1/f$ spectral behavior. In most cases the first wavelet subband and the subbands of resolution higher than the fifth level are not fitting the $1/f$ model (see Figure 4.2 and 4.3 as an example for the case of a trumpet). The conclusion is that only three or four subbands, depending on the instrument, are well representable by means of the $1/f$ pseudo-periodic model. In fact the high level subbands (the 6^{th} one in Figure 4.2 and 4.3) contain the harmonic part and information concerning the time envelope. From the experimental results shown in Figure 4.4, 4.5 and 4.6 it is possible to see how the pseudo-periodic $1/f$ -like model is well suited for representing the inner subbands, that is, the 2^{nd}, 3^{rd}, 4^{th}, and 5^{th} subbands. These results confirm that the pseudo-periodic $1/f$ model provides a good method for reproducing synthetic voiced sounds with the same power spectra as that of given real-life samples. Figures 4.7 - 4.9 show the results of the resynthesis of an oboe, a trumpet and a flute.

As a further confirmation of the pseudo-periodic $1/f$ model we report the results of the frequency warped version of the method presented in Section 3.6. When the HBW filter bank is replaced by the HB-FWW filter bank in Figure 3.17 an optimization procedure has to be run in order to find an optimal subdivision of the subbands according to the magnitude spectrum of the analyzed sound. The optimization criterion consists of finding the best sequence of parameters $d_1, ..., d_N$ providing sets of HB-FWWT coefficients whose variances fit in an optimal way a straight line. Table 4.1 shows some results for the first 3 harmonics (6 channels) of a trumpet note. As one can see the frequency warping is biased in the sense that the optimized band subdivision is slightly tighter than a pure pseudo-periodic $1/f$ behavior, i.e., the wavelet subbands are larger than the normal dyadic tiling. However the order of the parameters d is very small that it is possible to state that a relevant deviation with respect to a pseudo-periodic $1/f$ model does not occur. Additionally, from listening experiments it appears that the rigid HBW dyadic grid is sufficient for a high quality reproduction of voiced sound with stable pitch. This is not the case for voiced sounds with time-varying pitch, as it will be discussed in Chapter 5.

	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$
$n=1$	0.1801	0.2336	0.0529	0.1668	0.2005	0.0667
$n=2$	0.0616	0.0874	0.0713	0.0239	0.0765	0.1035
$n=3$	0.107	0.0129	0.152	0.0872	0.0716	0.0664
$n=4$	0.0255	0.1833	0.0788	0.0341	0.1466	0.0311

Table 4.1: Frequency warping d_n coefficients optimizing the frequency wavelet tiling for the first 3 harmonics of a trumpet note (347.2 Hz) 4 wavelet scales.

4.3 A model for the stochastic components of voiced sounds

From an acoustic point of view (going beyond the mere magnitude spectrum matching), some refinements are necessary. The white noise coefficient approximation provides a good equilibrium between the energy of the harmonic compo-

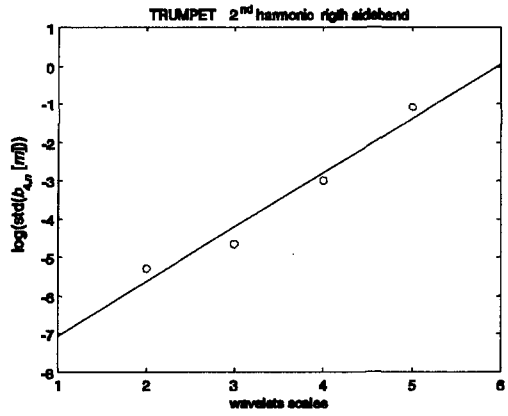


Figure 4.4: Trumpet 2^{nd} harmonic, right sideband, 2^{nd},3^{rd},4^{th}, and 5^{th} subbands, i.e., $p = 4$, $n = 2, 3, 4, 5$. Correlation coefficient: 0.9791

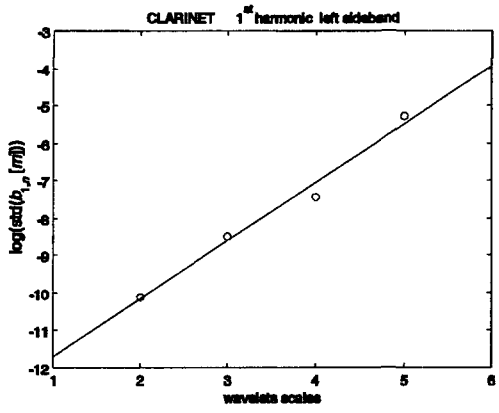


Figure 4.5: Clarinet 1^{st} harmonic, left subband, 2^{nd},3^{rd},4^{th}, and 5^{th} subbands, i.e., $p = 1$, $n = 2, 3, 4, 5$. Correlation coefficient: 0.9911

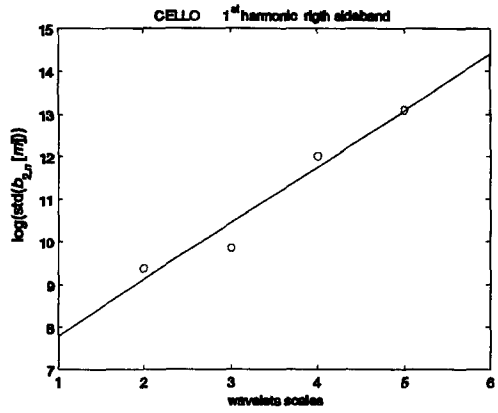


Figure 4.6: Cello 1^{st} harmonic, left sideband, 2^{nd},3^{rd},4^{th}, and 5^{th} subbands, i.e., $p = 2$, $n = 2, 3, 4, 5$. Correlation coefficient: 0.9739.

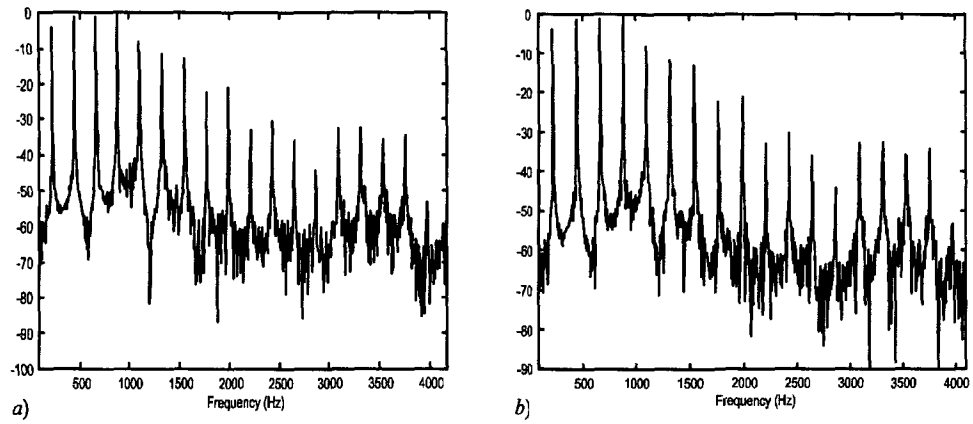


Figure 4.7: *a*) Real-life oboe (287.5 Hz) and *b*) resynthesized version.

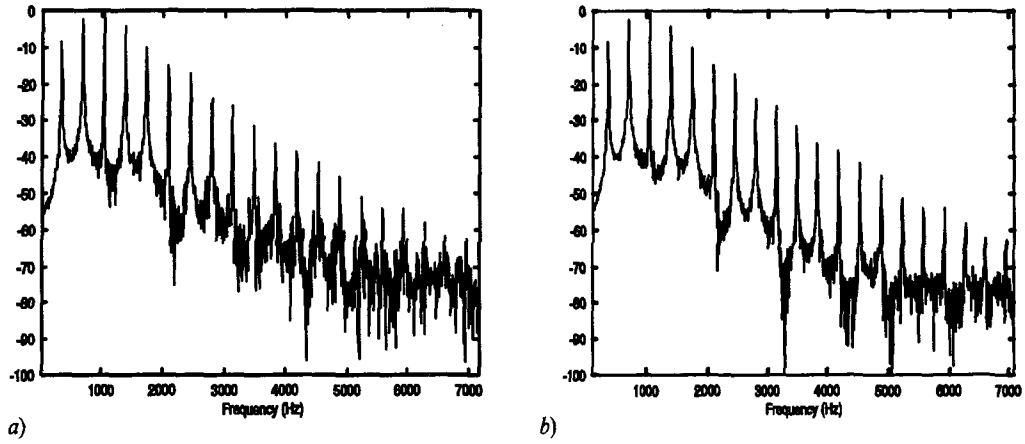


Figure 4.8: *a*) Real-life trumpet (347 Hz) and *b*) resynthesized version.

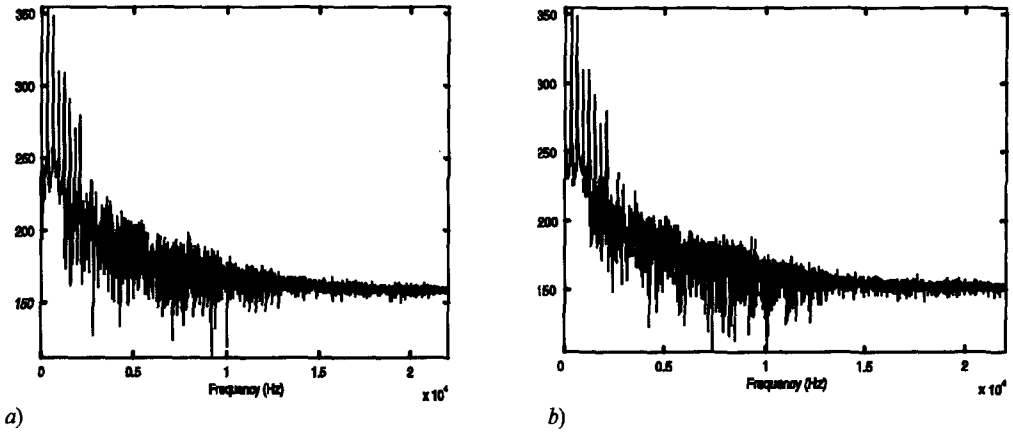


Figure 4.9: *a*) Real-life flute (298 Hz) and *b*) resynthesized version.

nents and the stochastic ones. Nevertheless by means of white noise we obtain something that sounds as "properly energy-scaled white noise" with lack of perceptual fusion with the deterministic components. Some kind of coefficient pre-filtering, i.e., white noise coloring is necessary. The starting point are the results in [99] and [88], which shows that the analytical wavelet coefficients of a $1/f$ process have a small but non-zero correlation. This theoretical result is confirmed by the experimental data. In order to simulate this correlation we perform an LPC analysis of the HBWT coefficients. The resulting AR (autoregressive) filters are employed in the resynthesis, in order to color the raw white noise coefficients (see Figure 4.16). The different "order" of acoustic quality obtained by means of this technique is clearly audible. If we compare each perfectly reconstructed subband with the corresponding synthetic one, the degree of resemblance is very high. This is one of the ingredients of the model for the HBWT coefficients relative to the stochastic components of sounds. The second ingredient of the model is the time envelope of the HBWT analysis coefficient energy. Once extracted, these envelopes are applied to the resynthesis coefficients. As a consequence of these refinements the quality of the reproduced sounds improves significantly at the expense of a larger number of parameters.

In section 4.3.1 a short introduction to LPC is proposed, in which we discuss the role of LPC in terms of a physical model contribution to the spectral modeling based FAS method. In sections 4.3.2 and 4.3.4 we introduce the stochastic model for the HBW coefficient.

4.3.1 Sound modeling via Linear Predictive Coding (LPC)

As any musical instrument, our voice is formed by a resonant system and an excitation system. The resonant system works as a filter and is formed by the cavities of our breathing and oral system. We can partially change the frequency response of this filter by our muscles, closing or deforming some of these cavities. There are basic characteristics making distinct voices sounding different. Various models can be devised in order to face speech-processing problems.

One of the simplest and most appealing one is the exciter-plus-filter system modeling the vocal cords and breath physical behavior and the oral cavity resonances. The model is essentially based on Linear Predictive Coding (LPC), which allows one to extract from a given speech signal a suitable filter approximating the oral cavity response, excited by white noise. LPC is one of the most succesful products of speech processing research [45] [44] [68] [67] [58] [25]. Figure 4.10 represents a model, which is clearly related to the physical structure of our oral system. There are two different types of exciters: an impulse generator for simulating and reproducing the effect of the glottal impulses and a noise generator in order to model the breath noise. Different models of the glottal pulse reproduction, i.e., the "vocal cord vibrations" can be considered. On example is given by the Exponential Model or Rosenberg Model [75]. A possible physical model for simulating the behavior of the oral cavity is the lossless tube model [68]. This model has the advantage of being strictly related to our direct physical experience of speech production.

The LPC model is quite simple and intuitive from a schematic point of view. However it is quite sophisticated and extremely effective from the mathematical and conceptual point of view. In other words it allows a deeper insight in the

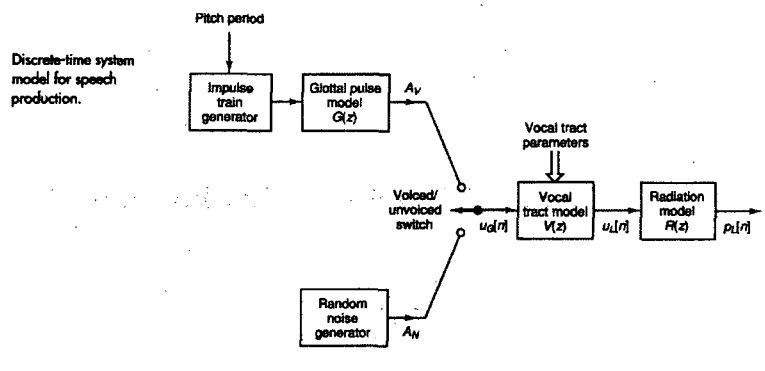


Figure 4.10: A physical model for speech synthesis.

stochastic characteristic of a speech signal itself, while it gives a clear representation of what is going on physically. It can be considered as the prototype of physical modeling techniques. At the same time it provides a very interesting interpretation of what is going on from the DSP point of view, both in time and in frequency domain. In other words it also has a spectral modeling content. This makes LPC one of the most attractive results in terms of audio coding and audio analysis and synthesis.

In our case the AR filters obtained from the LPC analysis of the HBW coefficients correspond somehow to the resonant cavity of the instruments. We said 'somehow', since the HBW coefficients are not the noisy components themselves but a drastically downsampled version of the filtered noisy sidebands. The rough white noise exciter provide a model for the breath or bow noise, different from the sound produced by the vibrating reed/lips or string, respectively. In other words the idea is to reproduce the breath or bow noise by means of white noise and resonant filters modeling the physical behavior of musical instruments.

4.3.2 Autoregressive modeling

The main idea of linear prediction is to model a signal $s[n]$ as a linear combination of its past values and present and past values of a hypothetical input $u[n]$ to a system whose output is the given signal [44]. The previous statement is equivalent to:

$$s[n] = -\sum_{k=1}^{K} a_k s[n-k] + G\sum_{l=0}^{L} b_l u[n-l], b_0 = 1, \tag{4.3}$$

where a_k, b_l and the gain G are the parameters of the hypothetical system, $s[n-k]$ are the past values of the signal and $u[n-k]$ are the samples of the unknown input.

Equation (4.3) is based on the hypothesis that a signal $s[n]$ is predictable from its past and some inputs to a certain system $H(z)$. In particular we are interested in the case where $b_l = 0$, for $1 < l < L$, known as all-pole model or

autoregressive (AR) model. In this case the 4.3 reduces to:

$$s[n] = -\sum_{k=1}^{K} a_k s[n-k] + Gu[n]$$

that in the frequency domain becomes:

$$H(z) = \frac{G}{1 + \sum_{k=1}^{K} a_k z^{-k}}. \tag{4.4}$$

By assuming that the input $u[n]$ is totally unknown and that the signal $s[n]$ can be predicted from a linearly weighted summation of past samples, it is possible to write:

$$\hat{s}[n] = -\sum_{k=1}^{K} a_k s[n-k]$$

where $\hat{s}[n]$, denotes the approximation of $s[n]$. The error or residue is defined as:

$$e[n] = s[n] - \hat{s}[n] = s[n] + \sum_{k=1}^{K} a_k s[n-k]$$

From the minimization of the total squared error

$$E = \sum_{n=-\infty}^{\infty} (e[n])^2 = \sum_{n=-\infty}^{\infty} \left(s[n] + \sum_{k=1}^{K} a_k s[n-k] \right)^2$$

it is possible to derive the well known Yule-Walker equations

$$-R(i) = \sum_{k=1}^{K} a_k R(i-k), 1 \leq i \leq K \tag{4.5}$$

and the minimum average error

$$Err_K = R(0) + \sum_{k=1}^{K} a_k R(k),$$

where

$$R(i) = \sum_{n=-\infty}^{\infty} s[n]\, s[n+i].$$

Equations (4.5) can be solved to obtain the K coefficients defining the predictor error filter (or whitening filter):

$$A(z) = 1 + \sum_{k=1}^{K} a_k z^{-k}$$

and Err_K, corresponding in this case to the variance of the "whitened" output signal. The vocal tract model filter $H(z)$ is given by the inverse filter as in (4.4), with $G = 1$. By defining the input model one has a complete set of parameters characterizing the analyzed signal. The filter $A(z)$ is called whitening filter since it attempts to produce an output signal that is white (with flat spectrum). Conversely, white noise can be used as input to $H(z)$ in order to reproduce $s[n]$.

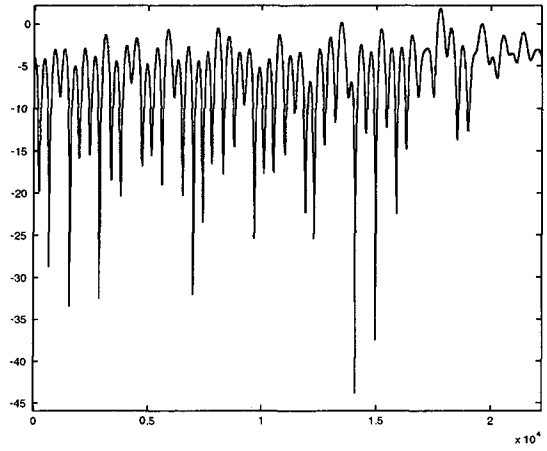


Figure 4.11: Magnitude FT of the HBWT analysis coefficients of a single subband of a trumpet sound: 2^{nd} WT scale.

4.3.3 LPC applied to the HBWT coefficients

In order to achieve an acoustical improvement in the reproduction of the noisy components of sounds we need to consider the little but not zero autocorrelation of the HBWT analysis coefficients [99]. This autocorrelation is very important from a perceptual point of view. We cannot simply use white noise, but we need to perform a spectral shaping of the resynthesis coefficients. In order to reproduce these correlations we perform an LPC analysis of the coefficients $b_{p,n}[m]$. By employing the Yule-Walker equations we compute the AR filter coefficients. The order of the AR filters usually ranges from 10 to 20 according to the musical instrument and to the subband scale level. As represented in Figure 4.15, the analysis is performed for each sideband p subband n. In Figures 4.11-4.14 a comparison between the Magnitude FT of the HBW analysis coefficients of two different subbands of a trumpet and the respective synthetic coefficients is reported. The result is that the reconstructed noisy subbands 4.1 and the synthetic ones are hardly distinguishable in listening tests. This means that, when mixed with the higher energy deterministic components in the final synthesis, the synthetic subbands achieve a complete perceptual fusion with the deterministic part and a transparent reproduction of the original sound noisy components.

4.3.4 Energy time envelope extraction: a refined spectrogram

As shown in Figure 4.15 by means of a short-time evaluation of the variance of the HBW coefficients $b_{p,n}[m]$ we extract an energy envelope of the coefficients themselves. Furthermore a polynomial interpolation of the obtained set of points is performed by means of linear splines. The number of parameters per subband is given by the knots and coefficients of a linear spline interpolation. In the resynthesis process the unitary variance white noise synthesis coefficients are energy-scaled by means of these envelopes (see Figure 4.16). Figures 4.17, 4.18

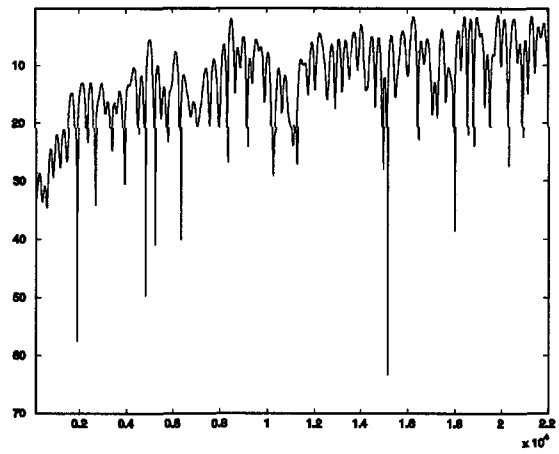


Figure 4.12: Magnitude FT of the HBWT resynthesis coefficients of a single subband of a trumpet sound obtained by means of AR filters: 2^{nd} WT scale.

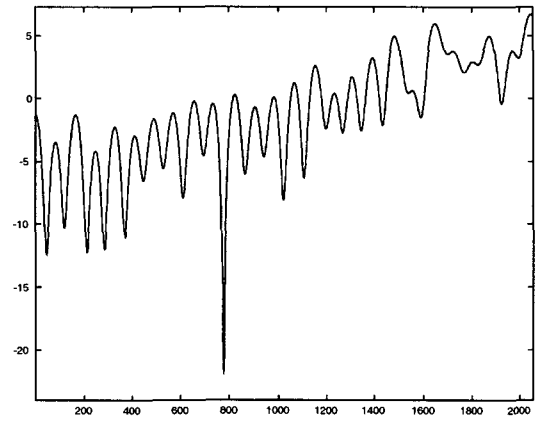


Figure 4.13: Magnitude FT of the HBWT analysis coefficients of a single subband of a trumpet sound: 3^{rd} WT scale.

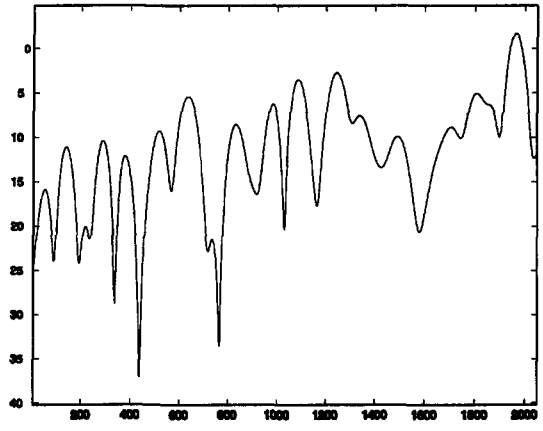


Figure 4.14: Magnitude FT of the HBWT resynthesis coefficients of a single subband of a trumpet sound obtained by means of AR filters: 3^{rd} WT scale.

and 4.19 show the HBW analysis coefficients of the first 3 scales of one sideband of the spectrum of a trumpet with their energy envelopes and their interpolation. Interpolation further reduces the number of parameters necessary to code the stationary part of voiced sounds. The transient segment, which is by the way the most significant in terms of energy, is perfectly reconstructed.

FAS can be thought of as an "intelligent spectrogram", i.e., a spectrogram whose bins are adapted to the characteristics of the signal that has to be analyzed/processed/synthetized. With respect to a STFT coefficient modeling FAS results in a more sophisticated model

Several experiments of reproduction of separated subbands for all of the channels p in the sense of (4.1) were performed and compared with the corresponding perfectly reconstructed subbands. The results are extremely good for all the tested instruments: oboe, clarinet, trumpet, bassoon, french horn, trombone, violin, viola, cello.
Listening sound examples are available at lcavwww.epfl.ch/~pietro/

4.4 A model for the deterministic components of voiced sounds

The second set of parameters we need to define is associated to the HB scale coefficients corresponding to the deterministic components of voiced sounds. This set specifies a model for that part of sound corresponding to the harmonic peaks of the spectrum. Our model recalls somehow another spectral modeling technique: the sinusoidal models. Before introducing the HB scale coefficient model a short review of the sinusoidal modeling techniques is given.

4.4.1 Sinusoidal models

This subsection is a short overview of the sinusoidal modeling techniques. The are two reasons for presenting this material here. The first reason is that our

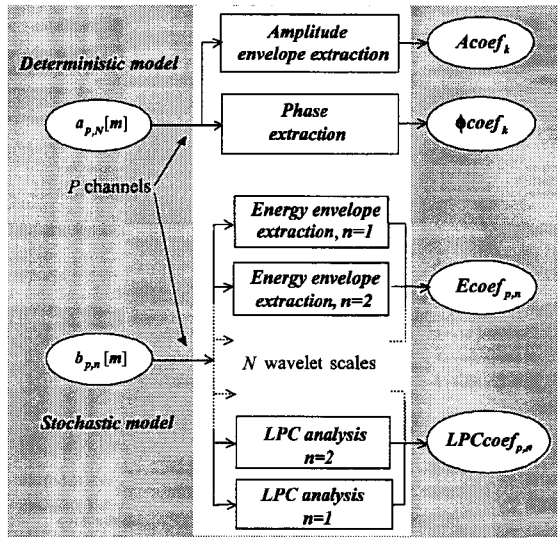


Figure 4.15: Resynthesis parameter extraction from the HBWT analysis coefficients $a_{p,N}[m]$ and $b_{p,n}[m]$. The parameters $Acoef_k$ and $\varphi coef_k$ are the coefficients and knots of the polynomial interpolation of the complexified HB scale coefficients of the k^{th} harmonic. The $Ecoef_{p,n}$ are the interpolation coefficients of the energy envelopes of the HBWT coefficients $b_{p,n}[m]$. The $LPCcoef_{p,n}$ are the filter coefficients resulting from the LPC analysis of the $b_{p,n}[m]$.

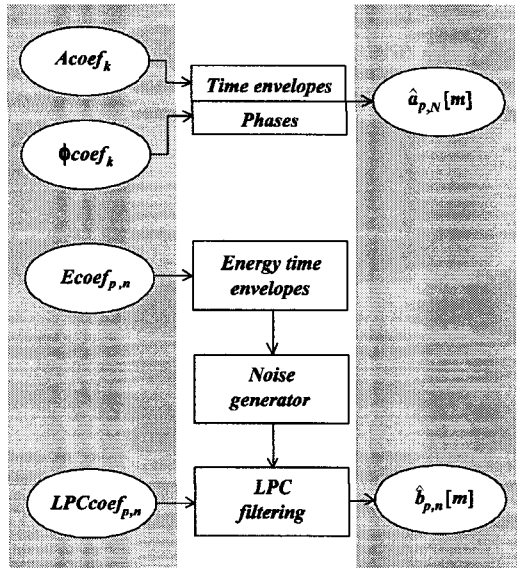


Figure 4.16: Parameteric resynthesis coefficient generation. The same notation as in the previous figure is used.

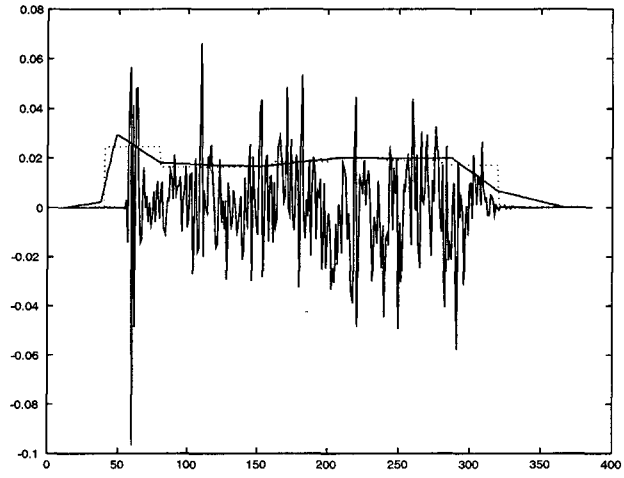


Figure 4.17: 1^{st} scale HBW analysis coefficients of one sideband of a trumpet with their energy envelope (dotted line) and its linear interpolation (continuos line).

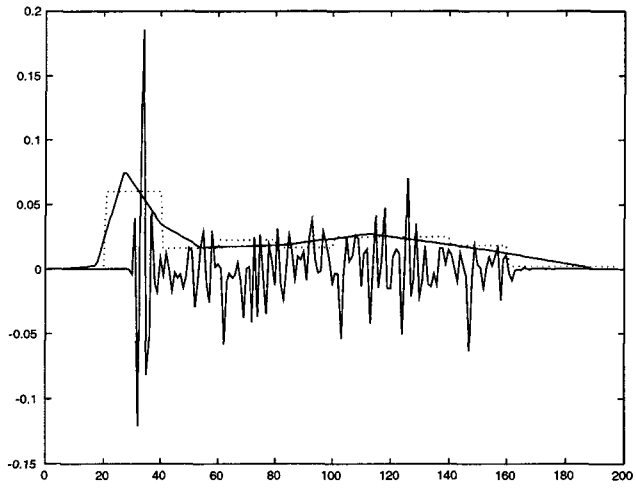


Figure 4.18: 2^{nd} scale HBW analysis coefficients of one sideband of a trumpet with their energy envelope (dotted line) and its linear interpolation (continuos line).

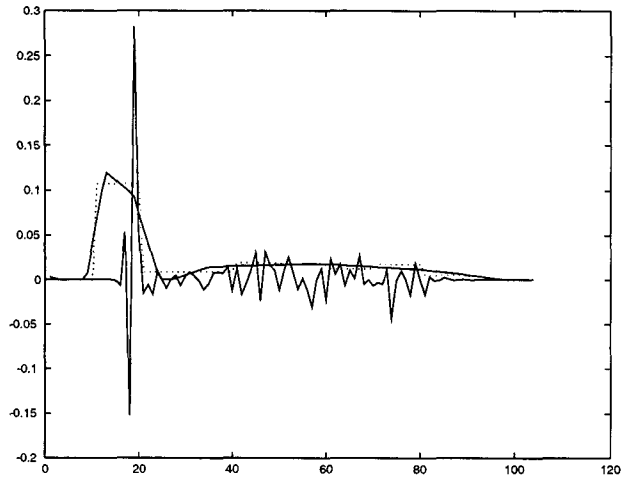


Figure 4.19: 3^{rd} scale HBW analysis coefficients of one sideband of a trumpet with their energy envelope (dotted line) and its linear interpolation (continuos line).

model for the deterministic component HBWT coefficients has some similitudes with sinusoidal models. This will be illustrated in Section 4.4.2. The other reason is to make a comparison between FAS and one of the most effective spectral models of recent research on audio processing and coding, as discussed at the end of this section.

The backbone of sinusoidal model is the STFT. The STFT provides a good sound processing tool in terms of high fidelity reproduction and minimum computation time. Moreover this tool is rather non-flexible and inefficient from a coding point of view. Other models have been proposed as an evolution of the STFT, with the aim of exploiting the computational efficiency of the STFT, while adapting the inner organization of the data to the object of the analysis, i.e., a sound with a spectrum presenting both relevant peaks and noisy bands [14] [15] [72] [73] [93] [30]. One of these models is given by the sinusoidal model introduced in [52]. In order to simplify the notation we will introduce this model in continuous-time, even though its application is mostly performed in discrete-time. The idea is to model and estimate time-varying parameters of sine waves components by means of spectral peaks tracking in the STFT. These parameters are the time-varying envelope $A_r(t)$ and the time-varying phase $\theta_r(t)$. The input sound $s(t)$ is modeled as:

$$s(t) = \sum_{r=1}^{R(t)} A_r(t) \cos[\theta_r(t)]$$

with

$$\theta_r(t) = \int_0^t \omega_r(\tau) d\tau + \phi_r$$

where $\omega_r(t)$ is the instantaneous frequency or frequency track of the r^{th} sine wave. In order to obtain a sinusoidal representation of the sound, a detection of

the spectral peak tracks in the STFT is performed. Finally the peaks and their phases are organized as time-varying sinusoidal tracks. The number of tracks $R(t)$ varies from frame to frame according to the peak tracking algorithm.

An evolution of the straightforward sinusoidal model is the Sinusoidal plus Residual Model or Spectral Modeling Synthesis (SMS) [83] [81] [84] [82] [95] [29] [74]. The introduction of a residual and of a distinction between a deterministic component and a stochastic component of sound makes the model much more flexible and efficient with respect to the previous ones, while maintaining good sound fidelity. The idea is to model a sound $s(t)$ as

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

where $e(t)$ is the residue. This residue is modeled as

$$e(t) = h(t) * w(t)$$

with $w(t)$ is white noise, $h(t)$ some appropriate, possibly time-varying filter and $*$ denotes convolution.

The SMS is a widely developed system for audio synthesis and coding, including time-varying scenarios, through algorithms solving problems as frequency matching criteria from one frame to the next one and decision for birth and death of sinusoid tracks. This makes the SMS a well-suited method for processing and encoding a large gamma of musical sounds. This is also the goal of an extended FAS method. As illustrated in Chapter 5, part of these extensions have already been achieved as a variable pitch version and a development of FAS scheme in order to be able to handle inharmonic spectra. Nevertheless many steps towards a full flexibility has still to be done.

The drawback of SMS is that the stochastic component $e(t)$ is defined simply as the difference between the original signal and the sinusoidal resynthesis. The residue is synthesized as white noise whose spectrum is shaped by means of a filter obtained from an approximation of the whole spectral behavior. This approximation is derived by means of a linear interpolation of the magnitude spectrum of the residue itself. The approach presents some limits in the synthetic sound results. In this sense the FAS stochastic model presented in section 4.3 represents an extremely effective model for the stochastic components of sounds and it can be seen as a significant improvement in the context of Structured Audio (SA) and audio coding for high quality sound reproduction.

4.4.2 A model for the HB scale coefficients

In this section we introduce a model for the HB scale coefficients $a_{p,N}[m]$, presenting itself some analogies with sinusoidal models [64]. The tuned MDCT section of the HBWT in theFAS model provides, in a sense, a broadband sinusoidal model in which each harmonic partial is modeled by two overlapping sidebands. Information on the details of each partial, such as amplitude envelope, bandwidth and center frequency is contained in the MDCT coefficients. In the HBWT these coefficients are further analyzed by means of WT in trend and details. More in particular we resort to a complexification of the HB scale

coefficients of pairs of adjacent channels, corresponding to the two sidebands of one harmonic. In other words, we consider the set of coefficients

$$c_{k,N}[m] = a_{2k-1,N}[m] + j a_{2k,N}[m], \tag{4.6}$$

where the $a_{2k-1,N}[m]$ and the $a_{2k,N}[m]$ are the harmonic-band scale-N coefficients of the left and right sideband of the k^{th} harmonic, respectively corresponding to channels $p = 2k-1$ and $p = 2k$ of Figure 3.10.

Equation (4.6) in polar form becomes

$$\begin{aligned} c_{k,N}[m] &= |c_{k,N}[m]|\, e^{j \arctan\left(\frac{a_{2k,N}[m]}{a_{2k-1,N}[m]}\right)} = \\ &= C_{k,N}[m]\, e^{j\varphi_{k,N}[m]} \end{aligned} \tag{4.7}$$

Additionally, we denote by $a_{p,0}[r]$ the MDCT coefficients at the output of the filters g_p (i.e., at "wavelet scale 0") and by $c_{k,0}[r]$ their complex combination. It is easy to show that these coefficients are constant for a perfectly tuned harmonic input signal with integer period. If the harmonic partials have a slowly time-varying amplitude envelope, the magnitude $C_{k,0}[r] = |c_{k,0}[r]|$ of the complexified coefficients represents a scaled and downsampled version of the envelope of the k^{th} harmonic. The phases $\varphi_{k,0}[r] = \arg(c_{k,0}[r])$ depend on the phase of the sinusoidal components. However, since the assumption on integer periodicity is too strict to be verified by real-life sounds, we need to investigate on their behavior.

It is also easy to show that, when the input signal $s(l)$ is sinusoidal of type

$$\cos(\omega l) = \cos\left(\frac{2\pi k}{P} + \Delta\omega\right),$$

with $|\Delta\omega| \leq \frac{\pi}{P}$, the coefficients $c_{k,0}[r]$ have constant amplitude equal to

$$C_{k,0}[r] = C_{k,0} = \frac{1}{4}\frac{\sin P\Delta\omega}{\sin\frac{\Delta\omega}{2}}. \tag{4.8}$$

Furthermore the phase of the complexified coefficients linearly depends on the index r, according to the following relationship:

$$\varphi_{k,0}[r] = rP\Delta\omega - \theta'_{2k}, \tag{4.9}$$

where θ'_{2k} is some phase depending on k, P and $\Delta\omega$ (according to (4.28) in the Appendix). A proof is given in the Appendix for the case where the window in (3.25) is a sine window (4.16). But we also gave a proof for generic windows within the PR condition of the MDCT.

The result is extendible to the superposition of sinusoidal signals with a time-varying amplitude envelopes $A_k(l) : \sum_k A_k(l)\sin(\frac{2\pi k}{P} + \Delta\omega_l)$. We assume that each amplitude envelope is approximately constant within the MDCT window $w(l)$, i.e., $A_k(l) \approx A_k(rP)$ for $l = rP, ..., rP+P-1$. In this case, (4.8) becomes:

$$C_{k,0}[r] = \frac{1}{4}\frac{\sin P\Delta\omega}{\sin\frac{\Delta\omega}{2}} A_k(r), \tag{4.10}$$

which is a scaled version of the amplitude envelope downsampled by a factor P, where the scaling factor depends on $\Delta\omega$. This property is confirmed in our experimental results.

All the previous results also hold for the coefficients $a_{p,N}[m]$ at the output of the last filters h_0 of the WT filter bank and for their complexification in (4.6), since the $a_{p,N}[m]$ are nothing but a lowpass filtered and downsampled version of the $a_{p,0}[r]$.

The validity of FAS as a method based on an "intelligent spectrogram", is confirmed by the model for the deterministic components. In this spectrogram the "frequency bins" (the HBWT subbands) are adapted to the spectrum of the analyzed sound by tuning P to the period of the sound and by means of the wavelet $1/f$-like frequency subdivision of each sideband of the harmonic peaks. In the resynthesis the coefficients are modulated in amplitude (the deterministic coefficients and the randomly generated stochastic coefficients) and phase (the deterministic coefficients only), according to the analysis results. Also, the spectra of the stochastic coefficients are shaped by means of the results of the LPC analysis of each of the HBWT subband coefficients $b_{p,n}[m]$.

4.4.3 Experimental results

Our experimental results confirm the analytical results of the previous section and show that the HB scale coefficients form smooth and slowly oscillating curves. As a consequence, the amplitudes $C_{k,N}[m]$ and the phases $\varphi_{k,N}[m]$ in (4.7) form smooth curves and nearly linear curves, respectively. These curves can be easily and efficiently approximated by means of a polynomial interpolation. In particular we adopted linear splines. The results in the case of a clarinet note are shown in Figures 4.21 and 4.22. We considered a three-level HBWT analysis of a clarinet sound of length 150984 samples and average pitch $P = 189$ samples (234.5 Hz, at a sr of 44.1 Khz), obtaining 110 HB scale coefficients per sideband. We employed splines of order 2 with 9 knots as interpolating functions for the amplitudes and with 11 knots for the phases. The amplitudes of the $C_{k,N}[m]$, as already remarked in the previous section, are the scaled time envelopes of the k^{th} harmonic partial downsampled by a factor $2^N P$. The phases represent the slow quasi-sinusoidal variation of the coefficients due to the difference between the harmonic partial frequency and the average $\frac{2\pi k}{P}$ of the two central frequencies $\frac{4k-1}{2P}\pi$ and $\frac{4k+1}{2P}\pi$ of the MDCT filters corresponding to the sidebands of the harmonic partial itself. Listening tests confirm that the resynthesis coefficients modeled by means of spline-interpolations provide high quality results. Differences between the original and the synthetic sounds are hardly perceivable.

In our model the transients are perfectly reconstructed from the original analysis coefficients. Perfect reconstruction (instead of a simple cross fade) is performed in order to maintain the deterministic and the stochastic components separate.

An interesting byproduct of the analysis of the harmonic-band scale coefficients is that the second derivatives of the $\varphi_{k,N}[m]$ provide a very efficient transient detector (see Figure 4.23). Where the sound is stationary, the absolute value of second derivative is less than 1. This becomes more than three times larger where the transients occur. We employed this result in order to define automatically the borders of the attack and decay transients.

The method was tested on several notes of different musical instruments: a clarinet, an oboe, a bassoon, a trumpet, a French horn, a violin, a viola and a cello. All of them gave equally good acoustic results.

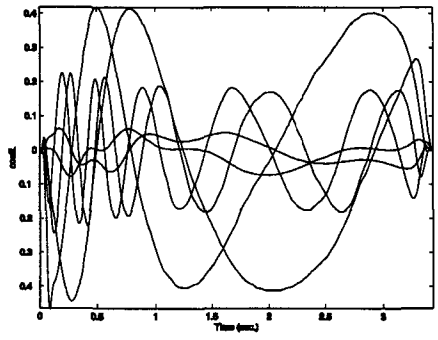


Figure 4.20: HB scale coefficients of a clarinet note at 234.5 Hz (B2). 3^{rd} scale 110 coefficients.

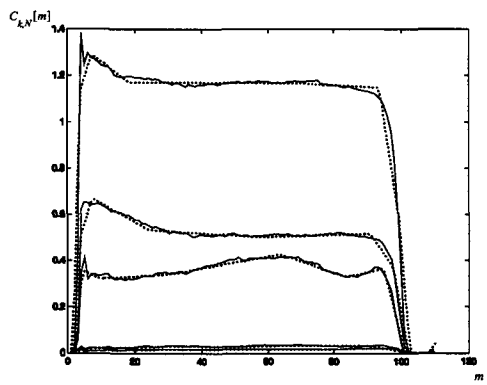


Figure 4.21: Amplitude $C_{k,N}[m]$ of the complex HB scale coefficients of a clarinet sound (continuous line) for $k = 1, ..., 5$ and their spline interpolation (dotted line). These curves are a scaled and downsampled version of the amplitude envelopes of the partials. The polynomial approximation (dotted line) is sufficient in order to make the synthetic sound not distinguishable from the original one.

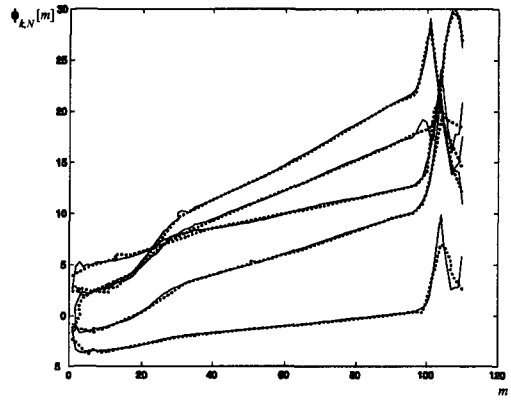


Figure 4.22: Phases $\varphi_{k,N}[m]$ of the complex HB scale coefficients (continuous line) for $k = 1, ..., 5$ and their spline interpolation (dotted line) of a clarinet sound. The behavior is reasonably linear in the stationary part. A temporary slight detuning is remarkable between coefficient 15 and 30. The non-linearity of the beginning and the end of the curves correspond to the attack and the decay transients respectively.

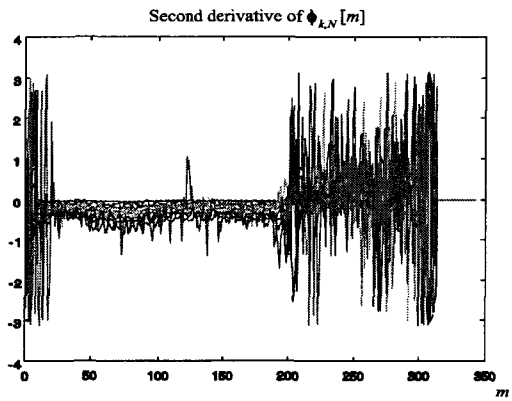


Figure 4.23: The second derivative of the phase of the complex HB scale coefficients of a violin.

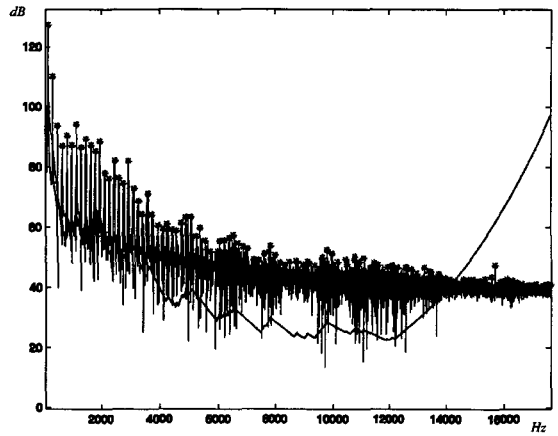


Figure 4.24: Psychoacoustic mask for an E2 legato cello note. The noisy components in the range from 3000 Hz to 14000 Hz are above the masking threshold. In order to provide a high quality sound one needs to reproduce also this part of the sound.

4.5 A psychoacoustic approach

This section discusses the results obtained in terms of data compression, taking into account psychoacoustic criteria and masking effects. The definition of a perceptual masking threshold on the whole frequency range allows one to discard all the HBWT coefficients that are perceptually irrelevant. In the following paragraphs we briefly summarize the adopted psychoacoustic criteria and then we illustrate the experimental results.

The first criterion is based on the following analytical approximation for the absolute hearing threshold [57] [34] [103]:

$$AbTh(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + \\ + 10^{-3}(f/1000)^4 \quad (\text{dB SPL}) \tag{4.11}$$

The second criterion that we adopted is the non-uniform hearing capabilities of the auditory system along the frequency range, i.e., the critical-band subdivision of the frequency domain. This is modeled by means of the cochlear continuous passband filter responses with non-uniform critical bandwidth

$$BW_c(f) = 25 + 75[1 + 1.4(f/1000)^2]^{0.69} \quad (Hz). \tag{4.12}$$

By means of (4.12) it is possible to compute the critical bandwidth corresponding to each partial peak of the spectrum of the analyzed sound. At every partial peak we apply a Signal-to-Mask Ratio (SMR) of 24 dB for the Tone-Masking-Noise (TMN) case. Finally we consider the spread of the masking effect, i.e., of the SMR across the critical bands according to the function:

$$M(b) = 15.81 + 7.5(b + 0.474) - 17.5\sqrt{1 + (b + 0.474)^2}(dB),$$

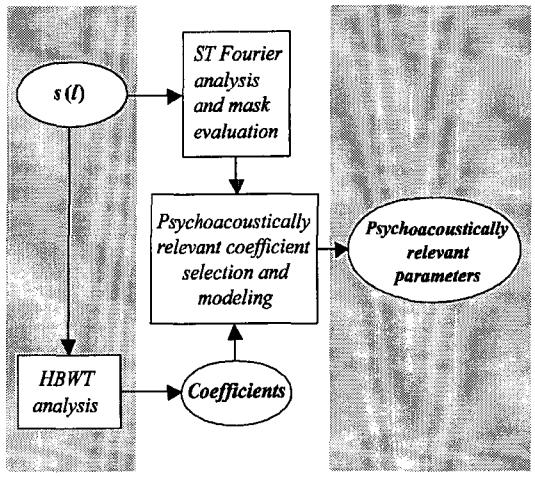


Figure 4.25: FAS in the context of Structured Audio coding methods. Sounds are represented by means of the parameters of the two models, deterministic and stochastic. At the same time a psychoacoustic analysis of sounds themselves establishes which are the perceptually relevant coefficients and which are the coefficients to be discarded. Only the first ones are encoded in terms of psychoacoustic relevant parameters.

where b represents the frequency in bark units [57]:

$$b(f) = [13\arctan(0.00076f) + \\ +3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right]] \text{ (Bark)}$$

The result is a masking threshold corresponding to the Just Noticeable Distortion (JND). Figure 4.24 shows the global masking threshold for a cello note. All of the HBWT coefficients corresponding to the spectral subbands lying underneath the masking threshold are discarded. Figure 4.25 represents a summary scheme of the whole procedure.

4.5.1 Data compression results

Table 4.2 reports the results in terms of data compression for the case of a trumpet. The compression rate is approximately of 20:1 before any adaptive quantization of data (that is of the resynthesis parameters) and entropy coding. Similar results are obtained with the other traditional musical instruments mentioned in the previous section. These results are extremely appealing compared to MPEG coders, where at a digital audio level (with the exception of MPEG-4 HILN) only psychoacoustic criteria are considered. The SA noise model proposed in HILN is the same as that of the Sinusoidal plus Residual model, i.e., as spectral modeling of the whole stochastic component by means of LPC or simple spectral interpolation. In the perspective of SA coders the FAS method can be seen as a transparent and effective tool for encoding and compressing of

the stationary part of voiced-sounds including a highly efficient model for the noisy components. In this sense FAS provides high quality audio reproduction and excellent compression rates at the same time. In this work we do not consider any further coding step. The complete scheme for a coding system would be of the kind shown in Figure 4.26. Higher ratios in terms of coding rates are expected by adding bit allocation and entropy coding to the FAS scheme.

Parameters	Number of parameters	Thr. factor	Param. post Psycoac. anal.
A Knot$_k$	$9{*}P/2{=}567$	$p_t/P \simeq 0.6$	341
A Coef$_k$	$9{*}P/2{=}567$	$p_t/P \simeq 0.6$	341
φ Knot$_k$	$11{*}P/2{=}693$	$p_t/P \simeq 0.6$	417
φ Coef$_k$	$11{*}P/2{=}693$	$p_t/P \simeq 0.6$	417
$Ecoef_{p,n}$	$10{*}N{*}P{=}2540$	0.1980	503
$LPC\text{coef}_{k,n}$	$10{*}N{*}P/2{=}1270$	0.3008	382
Total	**6330**		**2401**

Table 4.2: Data compression results for a 50000 samples long trumpet sound with pitch $P = 127$. In the analysis we adopted a wavelet scale $N = 2$ and a STFT window of length 4096 samples, corresponding at $N = 2$ to $\simeq 8$ HB scale coefficients. p_t denotes the last channel with energy above the masking threshold. All the coefficients of the channels $p_t + 1, ..., P - 1$ are discarded. The parameter notation is the same as in figure 4.15. The knots and the coefficients of the spline interpolation of the complex deterministic coefficients $c_{k,N}[m]$ are reported separately. In this example we performed an order 10 LPC analysis for each harmonic k and scale n. This means that at each scale n we performed only one LPC analysis for both the sidebands of the harmonic. We used ten equispaced coefficients $Ecoef_{p,n}$ to approximate the energy envelope of the random generated stochastic coefficients $\widehat{b}_{p.n,m}$.

4.6 Summary

In this chapter we provided a parametric model for generating the resynthesis coefficients of the stochastic components, i.e., the noisy part of voiced-sounds. The model is based on a parametrization of the stochastic behavior of the HBWT analysis coefficients associated with the noisy components. Then we introduced a model for the coefficients corresponding to the deterministic components of voiced sounds based on a parametrization of the amplitude and phase of a complex combination of the coefficients themselves.

The most attractive feature of the FAS is that it allows one to model both the stochastic and the deterministic part of voiced sounds at a very refined perceptual level and with a minimum amount of parameters controlling the synthesis process. The result is a sound synthesis technique able to provide synthetic sounds with natural timbre dynamics, which is at the same time a powerful data compression method in the sense of Structured Audio algorithms.

Further developments are necessary in order to make FAS flexible enough

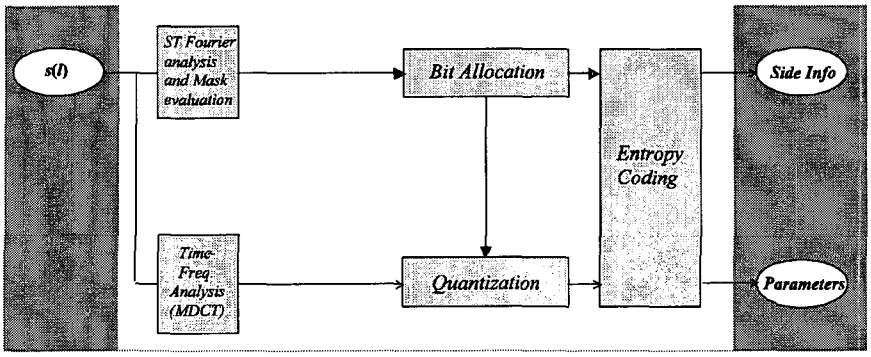


Figure 4.26: A scheme for a complete FAS coder.

and really appealing from a coding point of view. As a first improvement a pitch synchronous version is presented in chapter 5. The goal is to maintain a Perfect Reconstruction (PR) structure, being able to follow the pitch deviation of a vibrato note or more relevant change of pitch as, for instance, in a glissando. This implies the necessity of building a PR P-channel filter bank where one can change the number of channels P at each period. The requirement is that the HB scale coefficients $a_{p,N}[m]$ have to form smooth curves as in the time-invariant case.

A further degree of flexibility in the design of the P-channel filter bank is then necessary in order to achieve an arbitrary subdivision of the whole frequency range. This is the second subject of Chapter 5. An arbitrary band multichannel filtered bank has to be implemented in order to deal, for instance, with non-harmonic or polyphonic sounds. This can be obtained by giving up the PR constrains of the HBWT scheme or by means of some generalized technique of frequency warping [21] [20] [85], at the expense of an increased computational complexity and number of parameters. The same wavelet spectral tiling as in the harmonic case for modeling the noise and the peaks of the partial would be still available. In this case the goal is to enlarge the class of instruments that can be processed by means of FAS, including percussive instruments and, in general, instruments with non-harmonic spectra.

Appendix

In this appendix we examine the behavior of the output coefficients of the MDCT filter bank, to provide analytical support to the complex-sinusoidal model for the HB scale coefficients. In the first subsection, we consider the case of a $P-$periodic signal, where P is also the number of channels of the MDCT filter bank. In the second subsection, we handle the case of a sinusoidal signal with frequency slightly "out of tune" with respect to the number of channels P.

P–periodic signal

1) Let $s(l)$ be a P–periodic signal. Then, if

$$f_{p,r}(l) = f_p(l - rP),$$

with

$$p = 0, ..., P-1; \quad l = 0, ..., 2P-1; \quad r \in \mathbb{Z}$$

we obtain:

$$a_{p,0}[r] = \langle s, f_{p,r} \rangle = \sum_l s(l) f_p(l - rP) = \sum_{l'} s(l') f_p(l'), \tag{4.13}$$

i.e., the result is independent from r. It is easy to show that by a simple change of the index $l' = l - rP$ in the sum and exploiting the periodicity of $s(l)$. In particular this is true if $s(l)$ is represented by a Fourier series (additive synthesis with amplitude $A_k(l) = constant$).
2) Let $s(l)$ be one of the sinusoidal components of a P–periodic signal, i.e.,

$$s(l) = \cos\left(\frac{2\pi}{P} kl\right), \; k = 0, \pm 1, ..., \pm \left\lfloor \frac{P}{2} \right\rfloor$$

(the amplitude and the phase can be omitted, since these are constant and can be factored) and let be

$$f_p(l) = w(l) \cos\left(\frac{l(2p+1)\pi}{2P} + \theta_p\right),$$

where

$$\theta_p = \frac{(1-P)(2p+1)\pi}{4P}. \tag{4.14}$$

After some algebra we obtain:

$$a_{p,0}[r] = \langle s, f_{p,r} \rangle = \sum_l s(l) f_p(l) \tag{4.15}$$

$$= \mathrm{Re}\left\{ \frac{e^{j\theta_p}}{2} W\left(\frac{\pi}{2P}(2(2k+p)+1)\right) + \frac{e^{-j\theta_p}}{2} W\left(\frac{\pi}{2P}(2(2k-p-1)+1)\right) \right\},$$

where

$$W(\omega) = \sum_{l=0}^{2P-1} w(l) e^{-j\omega l}$$

is the DTFT of the window w.

In the specific case of a sine window:

$$w(l) = \sin\left(\frac{\pi}{2P}(l + \tfrac{1}{2})\right) \tag{4.16}$$

we have (geometric sum):

$$W(\omega) = \frac{e^{j\frac{\pi}{4P}}}{2j}\frac{e^{-jP\left(\omega-\frac{\pi}{2P}\right)}}{e^{-j\frac{1}{2}\left(\omega-\frac{\pi}{2P}\right)}}\frac{\sin P\left(\omega-\frac{\pi}{2P}\right)}{\sin\frac{1}{2}\left(\omega-\frac{\pi}{2P}\right)} - \frac{e^{-j\frac{\pi}{4P}}}{2j}\frac{e^{-jP\left(\omega+\frac{\pi}{2P}\right)}}{e^{-j\frac{1}{2}\left(\omega+\frac{\pi}{2P}\right)}}\frac{\sin P\left(\omega+\frac{\pi}{2P}\right)}{\sin\frac{1}{2}\left(\omega+\frac{\pi}{2P}\right)}. \tag{4.17}$$

It is easy to show that:

$$W\left(\frac{(2m+1)\pi}{2P}\right) = 0,\ m \neq 2nP,\ m \neq 2nP-1.$$

It follows that, in the particular case of the sine window, (4.15) is non-zero only for:

$$\begin{aligned} 2k+p &= 2nP \\ 2k+p &= 2nP-1 \\ 2k-p-1 &= 2nP \\ 2k-p-1 &= 2nP-1 \end{aligned} \tag{4.18}$$

It is easy to show that the only values of k that satisfy these conditions are

$$|k| = \left\lfloor \frac{p+1}{2} \right\rfloor. \tag{4.19}$$

In the case of a periodic signal this means that for each harmonic component k only the coefficients corresponding to the two sidebands of the harmonic itself are non-zero. These properties have to be approximately true for the other windows.

Generic sinusoidal signal

We consider now the more general case of a sinusoidal signal

$$s(l) = \cos(\omega l),$$

with frequency ω not necessarily multiple integer of $\frac{2\pi}{P}$, i.e., where the number of channels P does not correspond to the period of the signal. In this case

$$\begin{aligned} a_{p,0}[r] &= \langle s, f_{p,r}\rangle = \mathrm{Re}\left\{\sum_l e^{-jl\omega} f_p(l-rP)\right\} = \\ &= \mathrm{Re}\left\{e^{-jrP\omega}F_p(\omega)\right\}, \end{aligned}$$

where

$$\begin{aligned} F_p(\omega) &= \frac{e^{j\theta_p}}{2}W\left(\omega-\frac{(2p+1)\pi}{2P}\right) + \\ &\quad + \frac{e^{-j\theta_p}}{2}W\left(\omega+\frac{(2p+1)\pi}{2P}\right), \end{aligned}$$

and θ_p is given in (4.14).

If we write ω as $\omega = \frac{2\pi k}{P} + \Delta\omega$, for $|\Delta\omega| \leq \frac{\pi}{P}$, then:

$$a_{p,0}[r] = \mathrm{Re}\left\{e^{-jrP\Delta\omega} F_p\left(\frac{2\pi k}{P} + \Delta\omega\right)\right\} \tag{4.20}$$

From the discussion in the previous subsection (4.19) and supposing $k > 0$, it follows that only the elements for $p = 2k-1$ and $p = 2k$ are essentially non-zero and

$$\begin{aligned} F_{2k-1}\left(\frac{2\pi k}{P} + \Delta\omega\right) &\approx \frac{e^{j\theta_{2k-1}}}{2} W\left(\Delta\omega + \frac{\pi}{2P}\right) \\ F_{2k}\left(\frac{2\pi k}{P} + \Delta\omega\right) &\approx \frac{e^{j\theta_{2k}}}{2} W\left(\Delta\omega - \frac{\pi}{2P}\right). \end{aligned} \tag{4.21}$$

Furthermore, for $\Delta\omega \approx 0$ from the (4.17) we have

$$\begin{aligned} W\left(\Delta\omega + \frac{\pi}{2P}\right) &\approx \frac{e^{j\frac{\pi}{4P}}}{2j} e^{-j\left(P-\frac{1}{2}\right)\Delta\omega} \frac{\sin P\Delta\omega}{\sin\frac{\Delta\omega}{2}} \\ W\left(\Delta\omega - \frac{\pi}{2P}\right) &\approx -\frac{e^{-j\frac{\pi}{4P}}}{2j} e^{-j\left(P-\frac{1}{2}\right)\Delta\omega} \frac{\sin P\Delta\omega}{\sin\frac{\Delta\omega}{2}}, \end{aligned} \tag{4.22}$$

since the other terms in

$$\frac{\sin\left(P\Delta\omega \pm \pi\right)}{\sin\left(\frac{\Delta\omega}{2} \pm \frac{\pi}{2}\right)}$$

are not significant.

Thus we have:

$$a_{2k-1,0}[r] \approx \tag{4.23}$$
$$\mathrm{Re}\left\{e^{-jrP\Delta\omega} \frac{e^{j\theta_{2k-1}}}{2} \frac{e^{j\frac{\pi}{4P}}}{2j} e^{-j\left(P-\frac{1}{2}\right)\Delta\omega} \frac{\sin P\Delta\omega}{\sin\frac{\Delta\omega}{2}}\right\}$$

$$a_{2k,0}[r] \approx \tag{4.24}$$
$$-\mathrm{Re}\left\{e^{-jrP\Delta\omega} \frac{e^{j\theta_{2k}}}{2} \frac{e^{-j\frac{\pi}{4P}}}{2j} e^{-j\left(P-\frac{1}{2}\right)\Delta\omega} \frac{\sin P\Delta\omega}{\sin\frac{\Delta\omega}{2}}\right\}.$$

Therefore, *the only non-zero analysis coefficients of a sinusoidal component with a frequency deviation* $|\Delta\omega| \leq \frac{\pi}{P}$ *with respect to the analysis frequency* $\frac{2\pi k}{P}$ *are the* $a_{2k-1,0}[r]$ *and the* $a_{2k,0}[r]$ *as given in (4.23) and (4.24), respectively. These analysis coefficients oscillate sinusoidally with frequency* $P\Delta\omega$ *and a certain phase also depending on* $\Delta\omega$.

Showing that the phase of the complexified linear combination of the coefficients

$$\begin{aligned} c_{k,0}[r] &= \langle s, f_{2k-1,r}\rangle + j\langle s, f_{2k,r}\rangle = \\ &= a_{2k-1,0}[r] + j a_{2k,0}[r] \end{aligned} \tag{4.25}$$

is linear in r becomes now quite straightforward. From the (4.23) and the (4.24) we can write:

$$\begin{aligned} a_{2k-1,0}[r] \approx\ & -\frac{1}{4}\frac{\sin P\Delta\omega}{\sin\frac{\Delta\omega}{2}} * \\ & * \sin\left(rP\Delta\omega - \theta_{2k-1} + \left(P - \frac{1}{2}\right)\Delta\omega - \frac{\pi}{4P}\right) \end{aligned} \tag{4.26}$$

$$a_{2k,0}[r] \approx \frac{1}{4}\frac{\sin P\Delta\omega}{\sin\frac{\Delta\omega}{2}} * \\ * \sin\left(rP\Delta\omega - \theta_{2k} + \left(P - \frac{1}{2}\right)\Delta\omega + \frac{\pi}{4P}\right). \tag{4.27}$$

Since

$$\theta_{2k-1} = \theta_{2k} + \frac{\pi}{2} - \frac{\pi}{2P},$$

(4.26) becomes

$$a_{2k-1,0}[r] \approx \frac{1}{4}\frac{\sin P\Delta\omega}{\sin\frac{\Delta\omega}{2}} * \cos\left(rP\Delta\omega - \theta'_{2k}\right),$$

with

$$\theta'_{2k} = \theta_{2k} - \left(P - \frac{1}{2}\right)\Delta\omega - \frac{\pi}{4P}. \tag{4.28}$$

Writing $c_{k,0}(r)$ in polar form results in

$$\begin{aligned} c_{k,0}[r] &= C_{k,0}[r]\, e^{j\varphi_{k,0}(r)} = \\ &= \left(a_{2k-1,0}[r]^2 + a_{2k,0}[r]^2\right) e^{j\arctan\left(\frac{a_{2k,0}(r)}{a_{2k-1,0}(r)}\right)} \end{aligned}$$

and

$$\varphi_{k,0}[r] = \arctan\frac{\sin\left(rP\Delta\omega - \theta'_{2k}\right)}{\cos\left(rP\Delta\omega - \theta'_{2k}\right)} = rP\Delta\omega - \theta'_{2k}, \tag{4.29}$$

which proves the linear dependency on r of the phase $\varphi_{k,0}[r]$.

On the other hand

$$\begin{aligned} C_{k,0}[r] &= \frac{1}{4}\frac{\sin P\Delta\omega}{\sin\frac{\Delta\omega}{2}}\sqrt{\sin^2\left(\varphi_{k,0}[r]\right) + \cos^2\left(\varphi_{k,0}[r]\right)} \\ &= \frac{1}{4}\frac{\sin P\Delta\omega}{\sin\frac{\Delta\omega}{2}} \end{aligned} \tag{4.30}$$

If we consider a sinusoid with a time-varying amplitude envelope $A_k(l)\cos(\omega l)$ and supposing that this is approximately constant within a window $w(l)$, i.e., $A_k(l) \approx A_k(r)$, then we obtain

$$C_{k,0}[r] = \frac{1}{4}\frac{\sin P\Delta\omega}{\sin\frac{\Delta\omega}{2}} A_k(r),$$

i.e., the result obtained experimentally that is a scaled version of the amplitude envelope downsampled by a factor P, where the scaling factor depends on $\Delta\omega$.

Chapter 5

Towards a Flexible Method

In this chapter, we introduce two generalizations of Fractal Additive Synthesis (FAS). The purpose is to extend the class of sounds that can be processed and encoded by means of FAS beyond voiced sounds with stable pitch. The first extension aims at handling voiced sounds with variable pitch. In order to achieve this result we consider two alternate paths:

a) Modify the P-channel filter bank design in order to obtain a time-varying structure able to follow the pitch variations of the analyzed sound.

b) Process the sound in order to make its pitch stable and analyze it with the ordinary Harmonic-Band Wavelet Transform (HBWT) scheme.

The second extension allows us to employ FAS also for inharmonic sounds, i.e. sounds with non-harmonically distributed spectral peaks as those of a gong or of a tubular bell.

In the last section we present a real-time DSP module, implementing the synthesis section of FAS on the basis of either artificially generated input parameters or parameter extracted from an off-line HBWT analysis of a real-life sound.

5.1 A pitch synchronous version

In this section, we discuss an extension of the FAS technique able to follow the natural evolution of the pitch of real-life voiced sounds, as well as more significant pitch modulations as in the case of vibrato effects. The main challenge of this task is the design of a perfect reconstruction (PR) filter bank with time-varying number of channels $P(r)$, where $P(r)$ is tuned to the pitch of the sound at the r^{th} period. Our previous HBW model was constrained to the case of fixed harmonic spectra. In the HBW analysis, P was tuned to the average pitch of the analyzed voiced sound. In order to set our method free from this limit, we introduce a PR pitch-synchronous P-channel filter bank design technique, obtained by modifying an already existing filter design method [79] and adapting this filter bank design to the FAS scheme. The output of the time-varying P-channel filter bank is processed in a similar way as in the previous FAS scheme. Thus, every output of the filter bank is analyzed by means of wavelets and the analysis coefficients are modeled by means of set of parameters with the same meaning as in the fixed-pitch version of the method. This method is illustrated

in Sections 5.1.1 and 5.1.2.

As a second solution for the varying-pitch problem we employed Time-Varying Frequency Warping (TVFW). TVFW allows for the stabilization of the pitch. It reverts the pitch-synchronous case to the fixed pitch case. Better results from a perceptual point of view are obtained at the cost of an increased number of parameters. This method is illustrated in Section 5.1.3.

5.1.1 Efficient cosine modulated filter banks

We introduce a method for designing a time-varying P_{max}-band filter bank able to switch to an arbitrary number of bands $P(r) \leq P_{max}$, while maintaining PR and critical sampling during the transitions. More precisely, we are able to switch to a number of bands $P(R) \leq P_{max}$ at any time l_R with $l_R = \sum_{r=0}^{R-1} P(r)$ for some sequence $P(0), P(1), \ldots, P(R-1) \leq P_{max}$. The $P(r)$'s form the time sequence of number of bands of the time-varying filter bank. Our goal is to obtain a pitch synchronous version of the HBWT scheme that is a scheme able to deal with voiced sounds with varying pitch as in the flute segment shown in Figure 2.21. The sequence of channels $P(0), P(1), \ldots, P(R-1) \leq P_{max}$ will be tuned to the pitch variations of the type shown in Figure 2.22.

In order to do that, we consider a new class of cosine-modulated filter banks recently introduced in [79] [80] [78]. The whole design method is based on the polyphase representation of filter banks [89] [87]. The design technique consists in a factorization of the polyphase matrices representing the filter bank. This factorization decomposes the polyphase matrix into a set of elementary sparse matrices, a diagonal matrix and a cosine-modulation matrix.

The great and general advantage of modulated filter banks is that the set of filters can be designed by means of a simple modulation of an FIR baseband prototype. From a computational point of view, this means that one needs to run an optimization only for the prototype frequency response, in order to approach as much as possible to the case of an ideal low-pass filter. Moreover the design method that we are going to introduce allows one to deduce the ensemble of filters banks for any $P(r) \leq P_{max}$ number of channels from the single prototype for the P_{max} case. For each $P(r)$ the filter banks are biorthogonal.

We start by illustrating the case of time-invariant number of bands as presented in [79]. In the next section, we introduce the time-varying case. The time-varying P-channel filter bank that we realized allows us to define the Pitch-Synchronous Harmonic-Band Wavelet Transform (PS-HBWT) as discussed in Section 5.1.2. We first subdivide the input sound $s[l]$ into length-P vectors, where P is the number of channels of the analysis and synthesis filter bank. In this way, we obtain a polyphase representation of $s[l]$:

$$\mathbf{s}[r] = (s_0[r], \cdots, s_{P-1}[r])^T$$

with

$$s_i[r] = s[rP + i].$$

In the z-domain, this can be written as:

$$\mathbf{S}(z) = [S_0(z), \cdots, S_{P-1}(z)]^T \tag{5.1}$$

where $S_i(z)$ is the z-transform of the signal sections $s_i[m]$.

We then consider a type-2 polyphase representation of an analysis P-channel filter bank [89] [87]. That is, we design a matrix $\mathbf{A}(z)$ whose elements are given by:

$$[\mathbf{A}(z)]_{p,l} = \sum_{r=0}^{\infty} g_p\,(rP + P - 1 - l)\,z^{-r}, \qquad p, l = 0, ..., P-1 \tag{5.2}$$

where the g_p are the impulse responses of the p^{th} filter of the P-channel filter bank. A type-1 polyphase representation provides the inverse P-channel cosine-modulated filter bank. That is, the synthesis polyphase matrix $\mathbf{R}(z)$ can be written as:

$$[\mathbf{R}(z)]_{p,l} = \sum_{r=0}^{\infty} g_p\,(rP + l)\,z^{-r}, \qquad p, l = 0, ..., P-1.$$

The analysis/resynthesis polyphase matrices $\mathbf{A}(z)$ and $\mathbf{R}(z)$ satisfy the perfect reconstruction relationship:

$$\mathbf{R}\,(z) \cdot \mathbf{A}\,(z) = \mathbf{I}.$$

The great advantage of the formulation introduced in this section is the extreme simplicity of the design consisting, as we will see, of a simple product of elementary matrices. Also, it is well known that modulated filter banks provide an efficient implementation based on the polyphase components of the prototypes and a fast transform. This is a further advantage from a computational point of view. Finally, as we will see, the formalism and implementation of filter banks with time-varying number of channels will be extremely easy and straightforward. First we write $\mathbf{A}(z)$ as:

$$\mathbf{A}\,(z) = \mathbf{C} \cdot \mathbf{F}\,(z),$$

where $\mathbf{C}$ is the $P \times P$ Discrete Cosine Transform (DCT) type IV matrix, whose elements are given by:

$$\mathbf{C}_{p,l} = \cos\left(\frac{\pi}{P}\,(p{+}0.5)\,(l + 0.5)\right), \qquad 0 \le p, l \le P-1 \tag{5.3}$$

and the filter matrix $\mathbf{F}(z)$ is a matrix with a sparse "bi-diagonal form", containing the polyphase components of the prototype

$$\mathbf{F} = \begin{bmatrix} f_{0,0} & & 0 & & f_{0,P} \\ & \ddots & & \iddots & \\ 0 & & \ddots & & 0 \\ & \iddots & & \ddots & \\ f_{P,0} & & 0 & & f_{P,P} \end{bmatrix}.$$

By means of the factorization introduced in [79] $\mathbf{F}(z)$ can be written as:

$$\mathbf{F}\,(z) = \prod_{i=\nu-1}^{0} \mathbf{L}_i\,(z) \cdot \mathbf{D}\,,$$

where the matrices $\mathbf{L}_i(z,m)$ have the following form:

$$\mathbf{L}_i(z) = \mathbf{J} + diag\left(l_0^i, \cdots, l_{\lceil P/2 \rceil - 1}^i, 0, \cdots, 0\right) \cdot z^{-1}$$

and where the l_j^i are some arbitrary coefficients and $\mathbf{D}$ is the diagonal matrix

$$\mathbf{D} = diag(d_0, ..., d_{P-1}). \tag{5.4}$$

The number of matrices ν is directly related to the length of the impulse response of the filters, i.e. it is related to the number of parameters l_j^i at disposal to optimize the frequency responses of the low-pass prototype filter. The inverse or resynthesis polyphase matrix is given by

$$\mathbf{R}(z) = \mathbf{F}^{-1}(z) \cdot \mathbf{C}^{-1},$$

where $\mathbf{C}^{-1}$ is the inverse of the $P \times P$ DCT type IV matrix (5.3) and $\mathbf{F}^{-1}$ can be easily computed as:

$$\mathbf{F}^{-1}(z) = \mathbf{D}^{-1} \cdot \prod_{i=\nu-1}^{0} \mathbf{L}_i^{-1}(z),$$

where

$$\mathbf{L}_i^{-1}(z) = \mathbf{J} - diag\left(0, \cdots, 0, l_{\lceil P/2 \rceil - 1}^i, \cdots, l_0^i\right) \cdot z^{-1}$$

and $\mathbf{D}^{-1}$ is the inverse of (5.4). In this way, we obtain a PR analysis/resynthesis scheme, where the analysis and resynthesis operations are given by:

$$\mathbf{Y}(z) = \mathbf{A}(z) \cdot \mathbf{S}(z) \qquad \text{and} \qquad \hat{\mathbf{S}}(z) = \mathbf{R}(z) \cdot \mathbf{Y}(z),$$

respectively.

5.1.2 The time-varying case: PS-HBWT

The main goal of this section is to provide a pitch-synchronous extension of FAS, introducing a new class of wavelets: the Pitch-Synchronous Harmonic-Band Wavelet Transform (PS-HBWT). The PS-HBWT can be viewed as an extension of the pitch-synchronous wavelet transform [18]. We first need to design a perfect reconstruction time-varying cosine modulated filter bank and then we adapt it to the structure of fractal additive synthesis. An extension of the P-channel filter design of the previous section to the case of filter banks with time-varying number of bands is possible and it will provide the necessary tool for the implementation of the PS-HBWT. In order to do this, we have to modify the DCT matrix in order to make it suitable for modulating a matrix $\mathbf{F}(z)$ of size $P_{max} \times P_{max}$, while generating a number $P(r) < P_{max}$ of bands. We can obtain this by splitting the modulation matrices into two symmetric parts (horizontally the analysis one and vertically the synthesis one) and inserting in between the two parts a zeros matrix of size $P(r) \times (P_{max} - P(r))$ in the analysis case and of size $(P_{max} - P(r)) \times P(r)$ in the synthesis case. We obtain the following matrices:

$$\mathbf{C}_a(r) = [\mathbf{C}_{left}(r) \mid 0 \mid \mathbf{C}_{right}(r)]$$

for the analysis case and

$$\mathbf{C}_s(r) = \begin{bmatrix} \mathbf{C}_{up}(r) \\ 0 \\ \mathbf{C}_{down}(r) \end{bmatrix}$$

for the synthesis case, with

$$\mathbf{C}_s(r) \cdot \mathbf{C}_a(r) = \begin{bmatrix} \mathbf{I} & & 0 \\ & 0 & \\ 0 & & \mathbf{I} \end{bmatrix}$$

The matrix $\mathbf{F}(z)$ has also to be adapted to the $P(r)$ case. The time varying case is constrained to using only one matrix $\mathbf{L}(z, r)$ at every period r. The matrices $\mathbf{L}(z, r)$ change period by period in the following way:

$$\mathbf{L}(z,r) = \mathbf{J} + diag\left(l_0, \cdots, l_{\lceil P(r)/2 \rceil - 1}, 0, \cdots, 0\right) \cdot z^{-1}$$

The reason of this restriction to only one matrix, as better explained later, is due to the necessity of a short impulse response in order to obtain time domain aliasing cancellation (TDAC) in the transition from one pitch to the following one. The matrix $\mathbf{D}$ is also changed by means of an insertion of ones according to:

$$\mathbf{D} = diag\left(d_0, \cdots, d_{\lceil P(r)/2 \rceil - 1}, 1, \cdots, 1, d_{P_{\max} - \lceil P(r)/2 \rceil + 1}, \cdots, d_{P_{\max}-1}\right).$$

Notice that the matrices $\mathbf{L}(z,r)$ (and thus the resulting matrices $\mathbf{F}(z,r)$) are of size $P_{max} \times P_{max}$ also for the $P(r)$ case. In the time-varying case, the optimization is performed only for the prototype of the P_{max} filter bank.

Finally we modify the input sound in order to subdivide it into vectors of length P_{max} in order to fit the $P_{max} \times P_{max}$ $\mathbf{F}(z,r)$ matrices, while using only $P(r)$ input samples at each period r. In order to do this we simply insert $P_{max} - P(r)$ zeros in the middle of the time-varying polyphase vector of the input sound $s[l]$, i.e. we build the following vectors:

$$\mathbf{s}[r] = \left(s_0[r], \cdots, s_{\lfloor P(r)/2 \rfloor - 1}[r], 0, \cdots, 0, s_{\lfloor P(r)/2 \rfloor}[r], \cdots, s_{P(r)-1}[r]\right)^T$$

where

$$s_i[r] = s\left[\sum_{j=0}^{r-1} P_j + i\right], \quad \text{with} \quad i = 0, 1, ..., P(r) - 1$$

Since the factorization at each period r includes only one matrix $\mathbf{L}(z, r)$ and the diagonal matrix $\mathbf{D}(r)$, only the vector $\mathbf{s}[r]$ and the first half of the vector $\mathbf{s}[r-1]$ are relevant in the polyphase convolution. Thus, the zero-insertion (switching to a lower number of channels) and "removal" (switching to a higher number of channels) in the modulation matrices does not generate artifacts during the transitions from one pitch to a different one. The same holds for the removal (switching to a lower number of channels) and restoration (switching to a higher number of channels) of the coefficients l_i and d_i in the matrices $\mathbf{L}(z,r)$ and $\mathbf{D}(r)$. The output vectors $\mathbf{y}[r]$ of the analysis filter bank $\mathbf{A}(z,r)$ (see Figure 5.1) have length $P(r)$ and can be written as:

$$\mathbf{y}[r] = \left(y_0[r], \cdots, y_{P_m - 1}[r]\right)^T$$

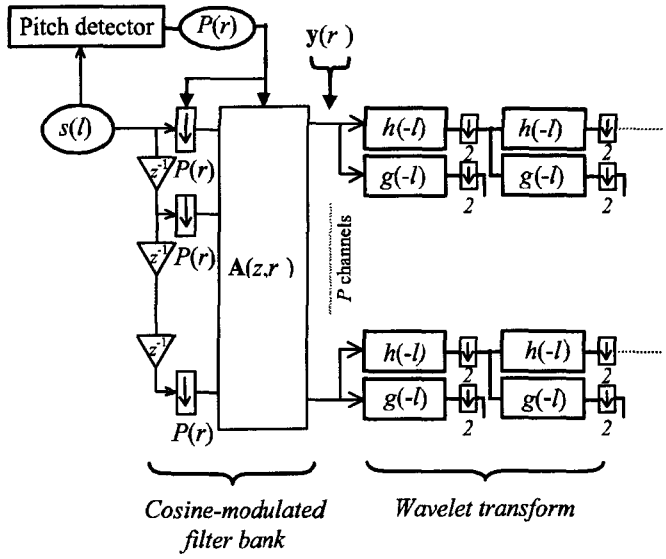


Figure 5.1: Polyphase representation of a cosine-modulated filter bank with time-varying number of channels. The scheme includes also the wavelet transformation of each channel. The whole structure implements the PS-HBWT.

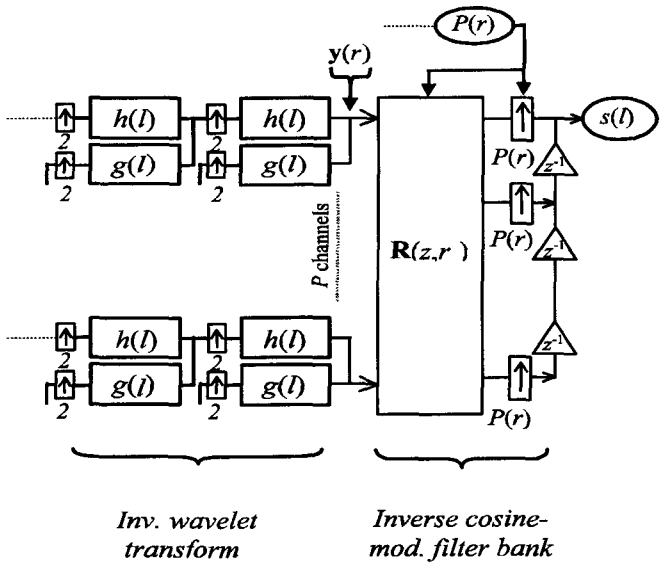


Figure 5.2: Polyphase representation of an inverse cosine-modulated filter bank with time-varying number of channels with inverse wavelet transformation of each channel. The whole structure implements the inverse PS-HBWT.

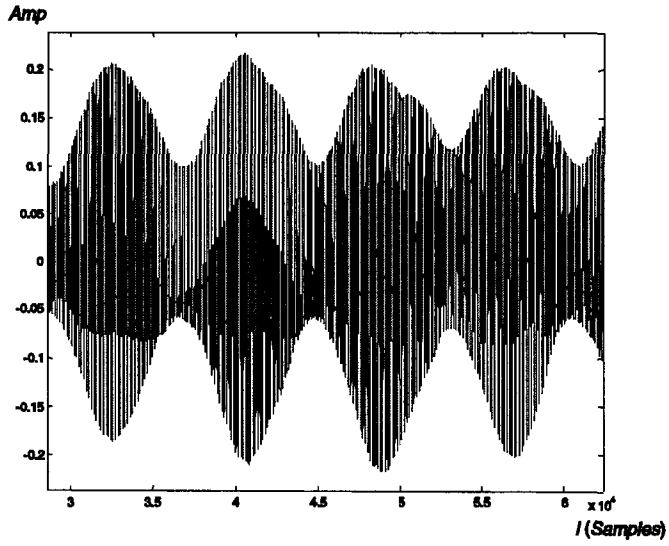


Figure 5.3: A segment of a flute sound with vibrato. Average pitch 298 Hz

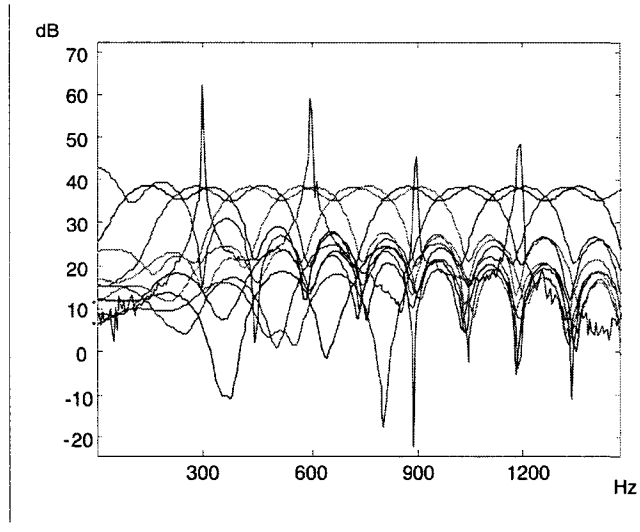


Figure 5.4: Magnitude frequency response of the PS-CMFB for the analysis of flute of Figure 5.3 and, superposed, 4 harmonic peaks of the flute itself.

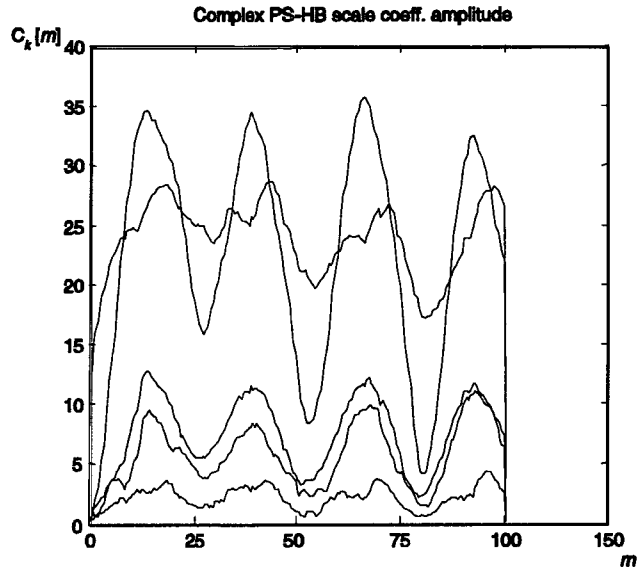


Figure 5.5: Magnitude of the complex PS-HB scale coefficients of the analysis of the flute of Figure 5.3.

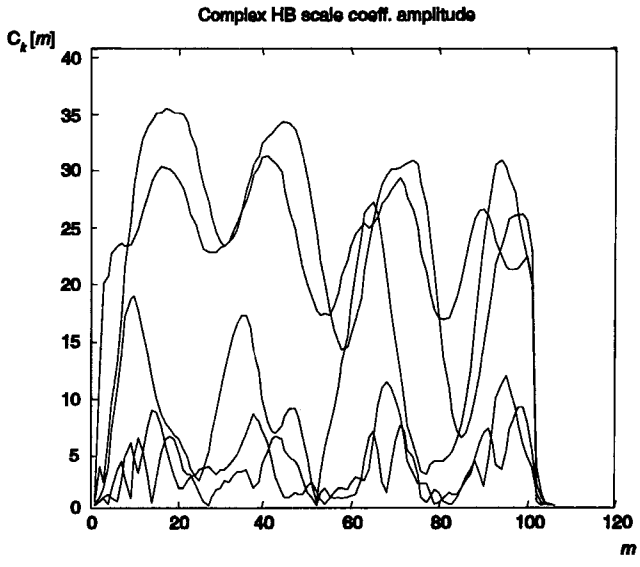


Figure 5.6: Magnitude of the complex HB scale coefficients of the analysis of the flute of Figure 5.3.

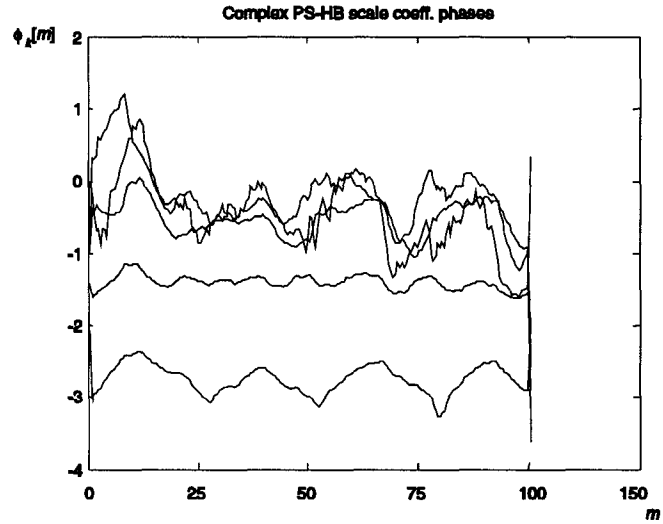


Figure 5.7: Phase of the complex PS-HB scale coefficients of the analysis of the flute of Figure 5.3.

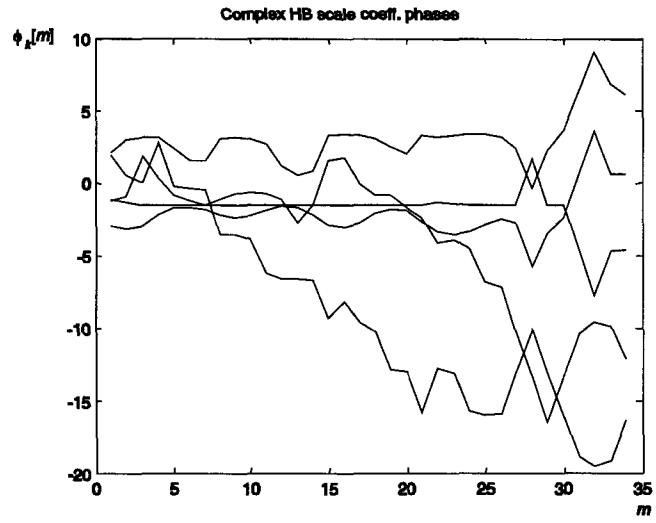


Figure 5.8: Phase of the complex HB scale coefficients of the analysis of the flute of Figure 5.3.

By means of simple zero-padding, we can make them all of the same length P_{max}. In this way, the P_{max} sequences of analysis coefficients can be injected into the wavelet filter banks as in the fixed-P HBW analysis. Finally the same parameter extraction as in the fractal additive synthesis can be performed. Due to the pitch-synchronous filter bank, the PS-HB scale coefficients now represent the period-by-period-time-varying harmonic peaks of the analyzed sound. In a similar way, the PS-HBW coefficients corresponding to the stochastic components represent the time-varying noisy sidebands of the harmonic peaks.

Figure 5.4 shows the first 4 harmonic spectral peaks of the flute in Figure 5.3 and the magnitude frequency response of a pitch-synchronous Cosine Modulated Filter Bank (CMFB) tuned on the instantaneous pitch of the flute itself. The overlapping of the adjacent filter bands is significant. This negative aspect is due to the short length of the impulse responses, i.e. to the low number of parameters at disposal for the optimization. Nevertheless, the relevant sidelobes of each filter embrace either the second sideband of the corresponding harmonic, or low energy spectral regions between the harmonics. Therefore, in spite of the strong overlap between the filter bank frequency bands, the acoustic results of the PS synthesis are still acceptable. This can be justified by the fact that since the sidebands of each harmonics usually have similar shape, their independent synthesis is often not necessary.

The introduction of the PS-HBWT allows one to deal with sounds with varying pitch of the kind of the flute with vibrato of Figure 2.21, whose pitch variations are shown in Figure 2.22. Experiments have been performed with many instruments, confirming the efficacy of the method. The main point of this filter design technique is that the PS-HB scale coefficients show a regular behavior through the transitions from one pitch to the other. Figures 5.5 and 5.7 show the amplitudes and phases of the complex version of the PS-HB scale coefficients of five harmonics of the flute with vibrato of Figure 5.3. The phase presents a slight "vibrato" behavior. Nevertheless, a piecewise linear interpolation is sufficient in order to obtain acoustic results of the same quality as in the fixed-pitch case. The envelopes in Figure 5.5 clearly follow the amplitude envelope original sound in Figure 5.3. Figure 5.6 and 5.8 show what happens to the amplitudes and phases, when the same flute is analyzed by means of a HBW scheme with P equal to the average pitch of the flute itself. The envelope distortions due to the variations of the flute pitch with respect to P are evident. In conclusion the coefficients at the output of the time-varying $P(r)$-channel filter bank provide smooth curves and, as a consequence, a robust sound model as in the constant pitch case.

An additional problem has to be solved for the pitch synchronous synthesis. While the PS-HBWT filter bank achieves TDAC, as soon as the resynthesis coefficients are approximated by means of the FAS model, dicontinuities occur in the synthetic signal at the transition from one period to the next one. These discontinuities embrace normally 4-6 samples. From the sequence $P(r)$, one has knowledge on the time position of the discontinuities are. By means of polynomial interpolations it is then easy to smooth them in order to cancel their effect on sound reproduction. Figure 5.9 shows two periods of a synthetic viola sound with vibrato, where two irregular regions at the borders of the periods are clearly visible. Figures 5.10 and 5.11 represent a detail of one period of the same viola sound before and after smoothing, respectively.

It is necessary to note that the number of coding parameters increases, i.e.

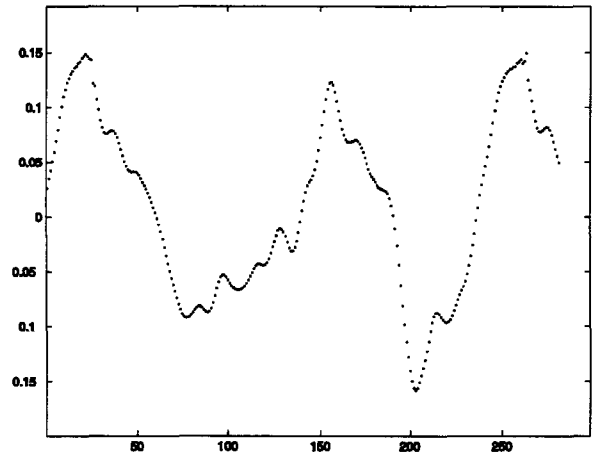


Figure 5.9: Two periods of a viola sound with vibrato analyzed and resynthesized by means of pitch-synchronous FAS.

the sequence of the $P(r)$'s has to be sent in terms of coding rate. This means one parameter per period in the worst case. Since pitch variations do not usually occur period by period, the number of parameters can be reduced by means of run-length pitch encoding.

5.1.3 Pitch-synchronous FAS by means of the Laguerre transform

In this section, we introduce a way to face the varying pitch case, which is dual to the PS-HBWT. In other words, instead of acting on the P-channel filter bank in order to make it able to follow the pitch variations, we act on the input sound. More precisely, we modify the varying-pitch input sound in order to obtain a sound with stable pitch P. The usual HBW scheme is sufficient in order to run the FAS coding procedure. The tool to obtain this is Frequency Warping (FW). FW was briefly reviewed in Section 3.6.1. Recently, a time-varying version of FW was introduced in [22] and [23]. This powerful technique allows to perform a FW sample by sample, i.e. to modify the "istantaneous" (defined sample by sample) pitch according to an arbitrary sequence of distortion coefficients $d(l)$, $l = 0, 1, ..., L-1$, where L is the length of the analyzed sound. Our goal is to obtain a sound with stable pitch equal to the average pitch P. From the instantaneous pitch and by means of the (3.32) it is possible to compute the sequence that stabilizes the pitch to the average pitch P of the input sound. The usual HBW analysis, coefficient modeling and resynthesis can be performed with stable pitch P. The pitch sequence of a viola sound with vibrato is shown in Figure 5.13 (before the TVFW) and 5.14, where it was stabilized by means of a TVFW. Figure 5.15 and 5.16 reproduce the phases of the complex HB scale coefficients of 5 harmonics of the same viola sound and the effect of the TVFW on the same HB scale coefficients, respectively. Even if a slight vibrato effect is still present, the pitch regularization effect of the TVFW is evident.

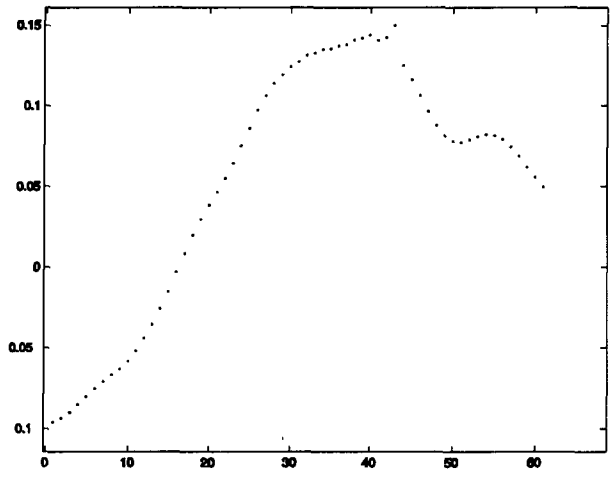


Figure 5.10: One period from Figure 5.9, showing in detail the discontinuities occurring at the junction of two periods.

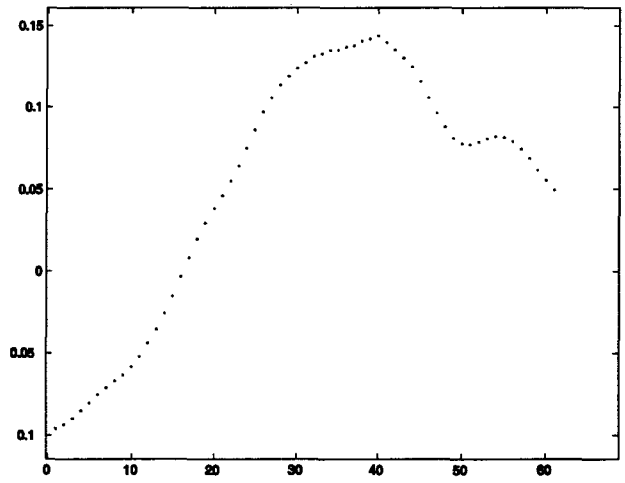


Figure 5.11: Same period as in Figure 5.10. The discontinuities have been smoothed by means of a polynomial interpolation.

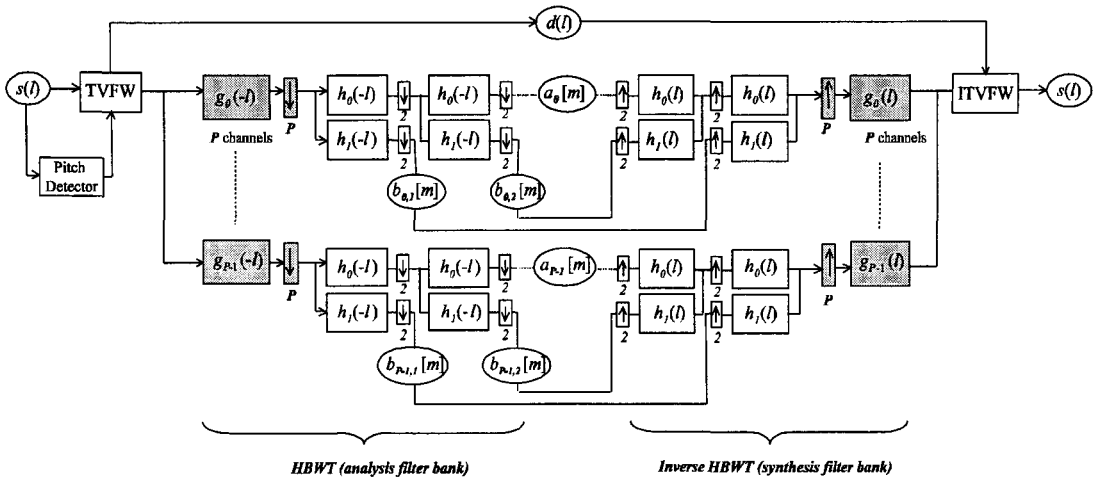


Figure 5.12: Pitch synchronous scheme realized by means of TVFW and HBWT.

In this case, the number of additional parameters $d(l)$ is equal to the number of samples itself. However, $d(l)$ follow the vibrato of the sound as shown in Figure 5.18 for the case of the flute. The curve in Figure 5.19 is the lowpass filtered version of the curve in Figure 5.18 and can be easily modeled by means of polynomial interpolation. This reduces drastically the number of parameters to the set of interpolation knots and coefficients equal to two times the number of oscillations. In conclusion, the number of additional parameters necessary to code the $d(l)$'s is of the same order as the other FAS parameters and can be included in the FAS scheme.

5.2 Inharmonic extension of the method

In the previous chapters, we demonstrated that the harmonic version of the wavelet transform, i.e. the HBWT is a "natural" tool to separate, decompose and resynthesize both the deterministic components and the noisy sidebands of the harmonic spectral peaks. This decomposition allows one to represent the different components of sound by means of a restricted set of parameters. These parameters, different for the deterministic and the stochastic parts, correspond to the two models that make the FAS an interesting method both for data compression in the context of structured audio and for sound synthesis/processing.

In this section, we present the extension of the method to the inharmonic case. The HBWT and PS-HBWT models are confined to the case of harmonic spectra. The time-frequency plane tiling is strictly harmonic. This is a major limitation and makes the method not usable for a large class of sounds, e.g., for sounds produced by percussion instruments. The spectra of many of these instruments show relevant peaks (see Figure 5.20). These peaks are the so called partials or deterministic components of the sound and can be sinusoidally modeled. These partials also show an approximately $1/f$ spectral behavior around the peak as in the harmonic case. The $1/f$-shaped, spectral sidebands

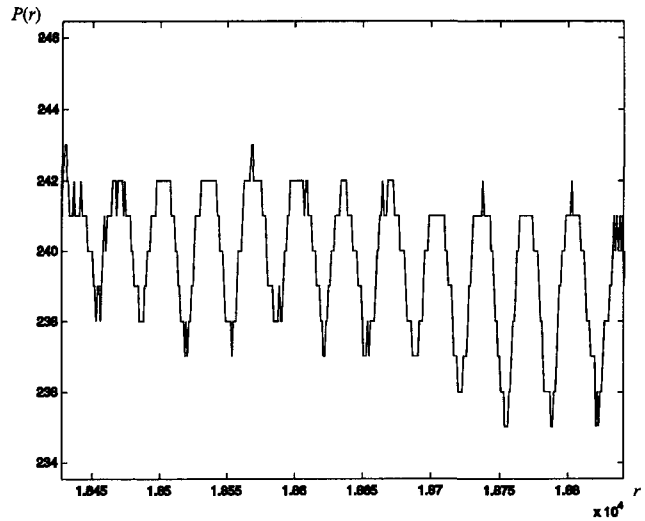


Figure 5.13: Pitch variations of a viola sound with vibrato. Average pitch 239.6 Hz.

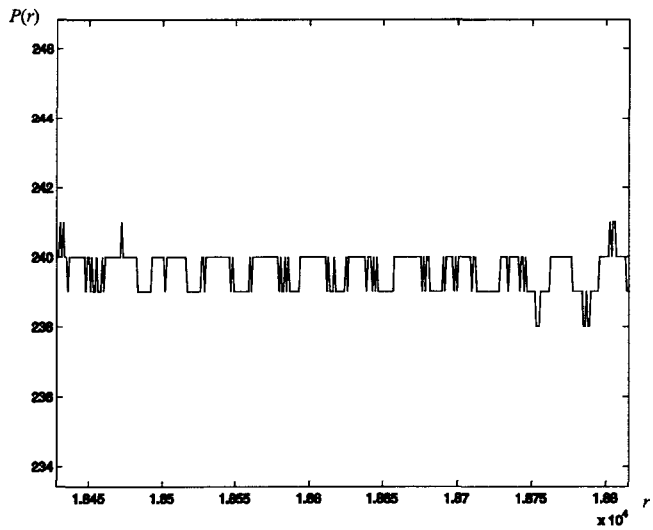


Figure 5.14: Stabilized pitch of a viola sound with vibrato after pitch stabilization via TVFW.

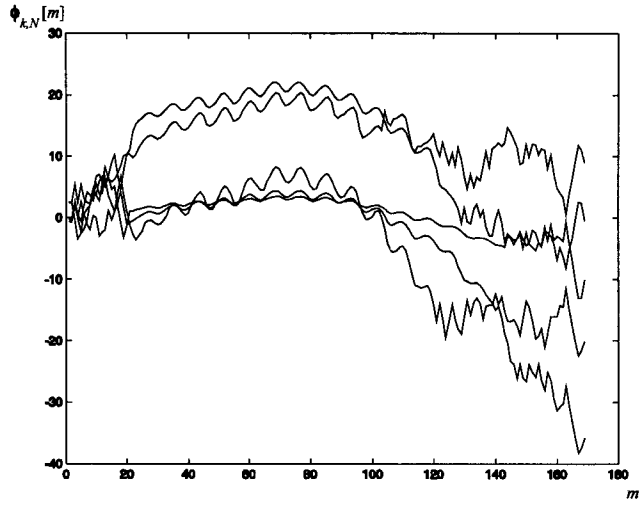


Figure 5.15: Phase of the complex HB scale coefficients of a viola sound with vibrato. Average pitch 239.6 Hz.

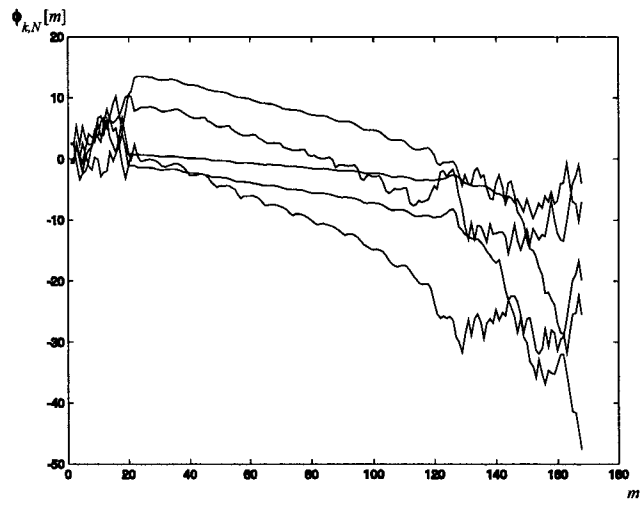


Figure 5.16: Phase of the complex HB scale coefficients of a viola sound with vibrato after TVFW pitch stabilization.

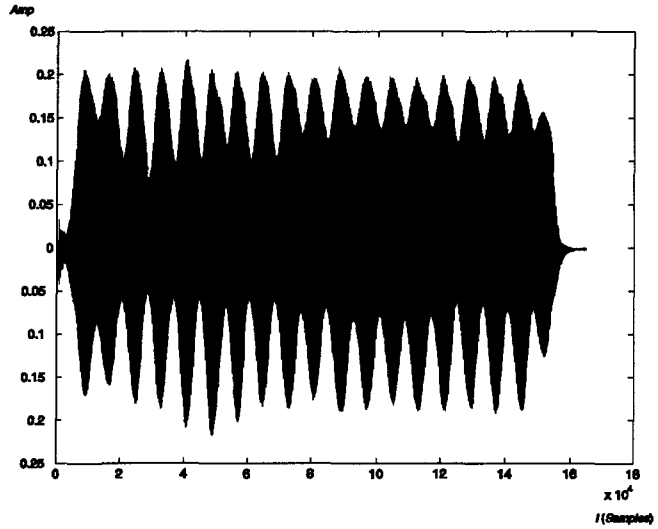


Figure 5.17: Flauto sound with vibrato. Average pitch 298 Hz.

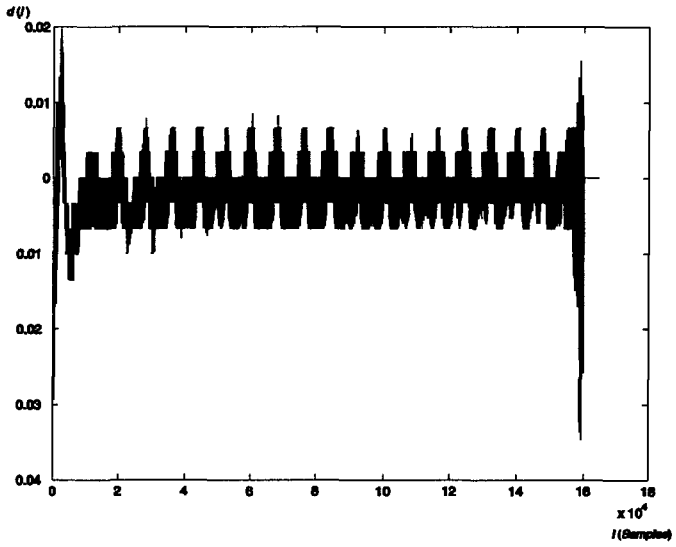


Figure 5.18: $d(l)$ sequence for the stabilization of the pitch of the flute sound with vibrato of Figure 5.17.

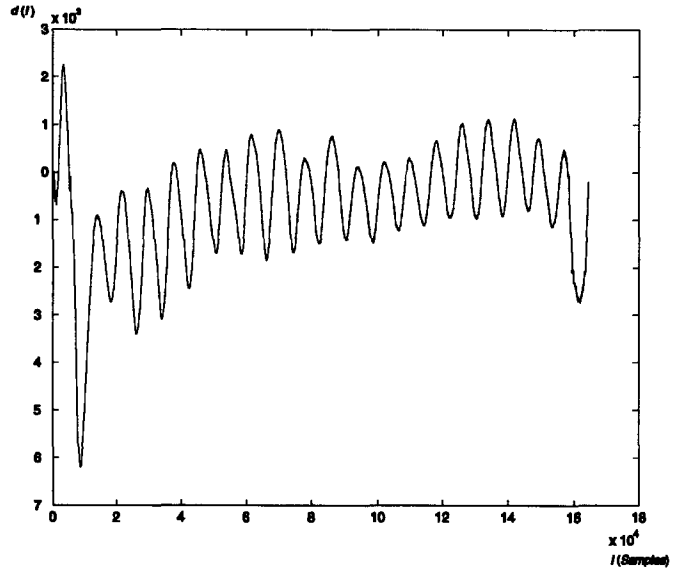


Figure 5.19: Lowpass filtered version of the $d(l)$ of Figure 5.18.

are the stochastic components. Thus, the same stochastic model used in the harmonic case can be employed. It is therefore reasonable to find a way to extend the FAS method to sounds with spectra of the kind shown in Figure 5.20.

The main problem addressed in this section is to provide a more flexible analysis/synthesis structure, extending the FAS model to inharmonic sounds. In order to do this, we abandon the PR structure provided by the HBWT and resort to a non-PR scheme, which is able to deal with aperiodic spectra like the one in Figure 5.20. A non-PR structure leads to aliasing problems and artifacts in the resynthesis. These artifacts are minimized by a careful filter design procedure and optimization. This part is discussed in detail in Section 5.2.2.

The peaks of inharmonic sounds correspond to deterministic physical events as, for instance, vibrational modes of membranes, metallic or wooden surfaces and bars occurring in many instruments as gongs, tam-tams, tympani, bells, vibraphones, marimbas and other percussion instruments. These peaks are not harmonically spaced but they show an approximately $\frac{1}{|f-f_n|}$ shape as in the harmonic case. Here, f_n denotes the frequency of the n^{th} partial. The idea is to use the same two models as in the harmonic case in order to control the resynthesis coefficients of the partials peaks and of their sidebands, respectively. The principle of the wavelet subband subdivision is therefore maintained. What we need to change is the MDCT section of the method, which is limited to a uniformly spaced subdivision of the frequency domain. For this purpose we design a non-uniform cosine-modulated filter bank (CMFB), where the band-pass filters are adaptively tuned to the non-equally-spaced spectral peaks of an inharmonic sound. The system is not PR, in the sense that only the spectral

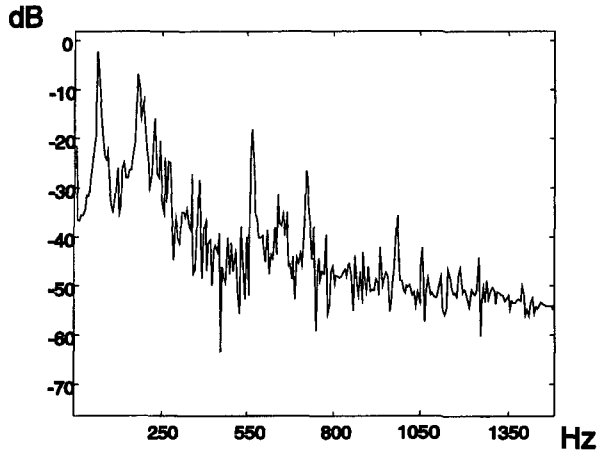


Figure 5.20: Magnitude FT of a Gong sound.

regions corresponding to the main peaks are analyzed and the overlap of the filter passbands is chosen empirically according to the spectral peak distribution. As a result of the analysis we obtain sets of analysis coefficients plus some more or less relevant residue. The residue energy can be arbitrarily reduced at the expense of an increasing computational time by means of recursive analysis of the residue itself. The "main body" of the sound, including both the partial peaks and their noisy spectral sidebands, is analyzed in a similar way as in the case of the HBWT.

5.2.1 Spectral peak picking

A preliminary and fundamental step in our technique is the implementation of a good spectral peak estimation algorithm. The task of this algorithm is not trivial, due to the variety of spectra produced by inharmonic sounds. Often it is hard to distinguish between noise and low energy partials. Also, resolution problems due to the proximity of two or more partials occur. The goal of the partial detection algorithm should be to find only the 'significant' peaks in the magnitude Fourier transform (FT) of a sound, where to be significant or not finally depends only on perceptual criteria. Once the peak picking algorithm has defined the partial frequencies, we can design the associated filter bank. With respect to an ordinary peak detection, we also need to define the optimal bandwidth of the filters that will subdivide the spectral range. For this purpose we need to take into account not only the partial position but also the position of its two neighbors (see Figure 5.21).

As a first step we consider the average of the spectrogram frames of the sound under analysis. This is an easy and effective way to make the partials more distinct and to get rid of the noise in the spectrum. For this purpose, we used the Blackman-Tukey method [86]. In this 'cleaned' spectrum, we perform a peak detection. The basic principle of the algorithm used in this work consists in comparing the magnitude of the candidate peak with a linear combination

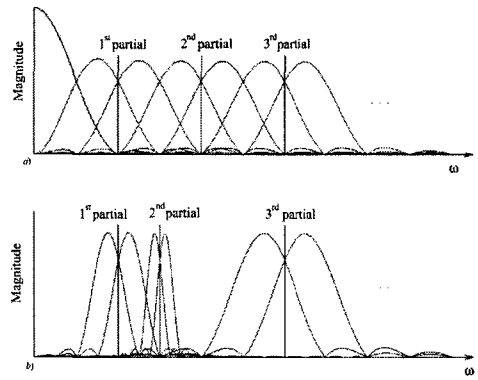


Figure 5.21: Magnitude FT of a CMFB a) Harmonic case b) Inharmonic case.

of the mean and standard deviation of a certain region R of the magnitude FT. The region R is chosen in different ways (Figures 5.22 and 5.23) in the neighborhood of the candidate peak itself. In order to make the algorithm more robust, different values of the coefficients of the linear combination and different criteria of definition of the region R are considered and compared before designating a peak.

5.2.2 Optimized band subdivision and filter design

As a second step we need to define the bandwidth of the filters. A preliminary estimate of the bandwidth is obtained by considering the distance of the peak from the left and right neighbor peak positions (d_{left} and d_{right}, respectively)

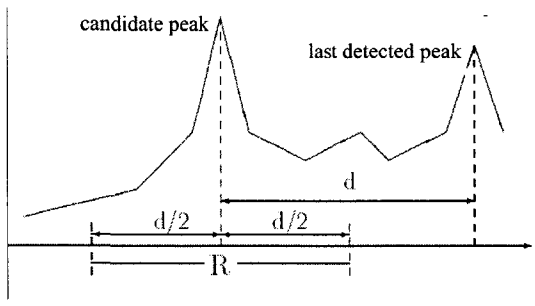


Figure 5.22: Example of choice of region R for peak searching. d is the distance between two consecutive peaks.

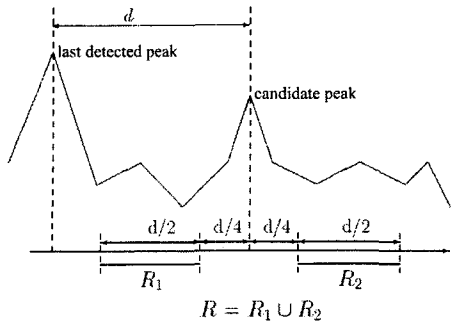


Figure 5.23: Example of choice of region $R = R_1 \cup R_2$ for peak searching. R_1 and R_2 are the frequency intervals $f_n \pm \left[\frac{d}{4}, \frac{3}{4}d\right]$, where f_n is the n^{th} peak and d the distance between two consecutive peaks.

and the approximately $\frac{1}{|f-f_n|}$ shape of the sidebands of the n^{th} partial. As a simple evaluation parameter of the shape we take the length $d_{threshold}$ of an interval around the peak where the magnitude spectrum is above a certain threshold, which depends on the spectral characteristics of the sound (see Figure 5.24). The chosen bandwidth is the minimum among $d_{left}/2$, $d_{right}/2$ and $d_{threshold}/2$. The parameter $d_{threshold}$ is important in order to maintain an analogy with the pseudo-periodic $1/f$ model, especially in the case of isolated spectral peaks.

The design of an inharmonic CMFB requires the definition of the most appropriate "hypothetical pitch" for each detected partial peak. As an example, we consider the first spectral partial of an inharmonic sound. If the first partial could correspond to a certain "harmonic" k_1 of a "hypothetical pitch" P_1, then we consider a P_1-channel MDCT filter bank of filters $g_p(l)$ as given in (3.25). We implement only two filters out of the whole filter set, corresponding to the indexes $p = 2k_1 - 1$ and $p = 2k_1$. These filters cover the two sidebands of the partial. As said in the previous subsection, the definition of the parameters P_1 and k_1 depends not only on the position of the partial but also on the position of the neighbor peaks. This provides a preliminary estimate of the bandwidth π/P_1 of the P_1-channel filter bank. However, these are not the only criteria for the choice of P_1 and k_1. Additionally, it is worthwhile to define the filter design procedure in a way to reduce the aliasing occurring in the analysis of each partial.

As shown in the Appendix of Chapter 4, when a sinusoid at frequency $k\pi/P$ is analyzed by a P-channel cosine modulated filter bank, only the outputs of the $2k^{th}$ and $(2k-1)^{th}$ channels are different from zero. Therefore, due to the fact that the P-channel filter bank is PR, this sinusoid can be reconstructed without aliasing using these two bands only. This means that, if the crossover frequency of the two filters is well centered on the peak frequency, we can achieve a nearly aliasing-free reconstruction of the deterministic part of the sound, i.e. of the part where the aliasing effects are more relevant. Additionally, increasing the parameter P provides filters with narrower passbands. This obviously means

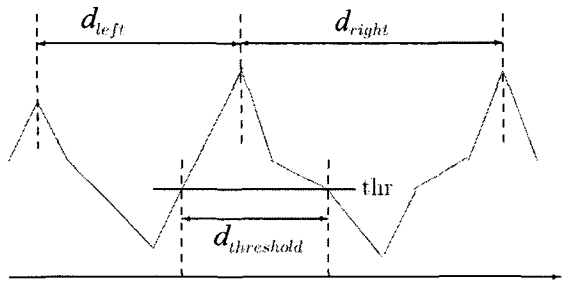


Figure 5.24: The parameters for the definition of the first estimate of the bandwidth (before the optimization) of the filters relative to one partial peak.

a higher resolution in terms of distribution of the cross over frequencies of the filters and the possibility of getting closer to the partial frequency. The goal of the optimization algorithm is to achieve a trade-off between the 'tuning' of the filters around the partial peak and a large enough bandwidth to include the sidebands of the partial. In order to do this, the following parameters are considered: the frequency of the partial, the preliminary estimate of the bandwidth, an interval for the variation of the bandwidth and an upper bound for the deviation of the crossover frequency of the two tested filters from the frequency of the partial. The algorithm calculates all of the filters with bandwidths in the given interval and selects the one differing the least from the desired frequency. If the difference does not fulfill the upper bound condition, then the algorithm reduces the bandwidth until the condition is fulfilled. The same criteria are applied to define the parameters P_n and k_n of the other pairs of filters, for the analysis of the other partials.

Once all the filters are defined, the inharmonic CMFB is implemented as in Figure 5.25. The structure of Figure 5.25 has the advantage of being PR at the condition of keeping the overall residue and adding it back to the reconstructed sound. From a coding point of view the residue is not considered and the goal is to reduce its perceptual relevance as much as possible.

In more detail, the filters separate the sidebands of the partials. Each n^{th} partial is processed as the k_n^{th} harmonic of a hypothetic voiced sound with pitch P_n. Each sideband undergoes a wavelet transformation, which decomposes it into subbands as in the harmonic case. The meaning of the upsampling of order P_n and of the inverse filters $g_p(l)$ is to reconstruct both the n^{th} partial and the aliasing due to downsampling of order P_n. In this way we keep track of the aliasing through the following partial analysis steps. When we reconstruct the partials and add them up together with the overall residue, we are able to achieve a time domain aliasing cancellation. As already mentioned, the overall residue can be arbitrarily reduced in energy by means of a recursive analysis at the cost of an increasing number of parameters.

Figure 5.26 represents some results of the peak detection algorithm applied to a gong (a), the resulting CMFB coefficients (b) and the resulting scale coef-

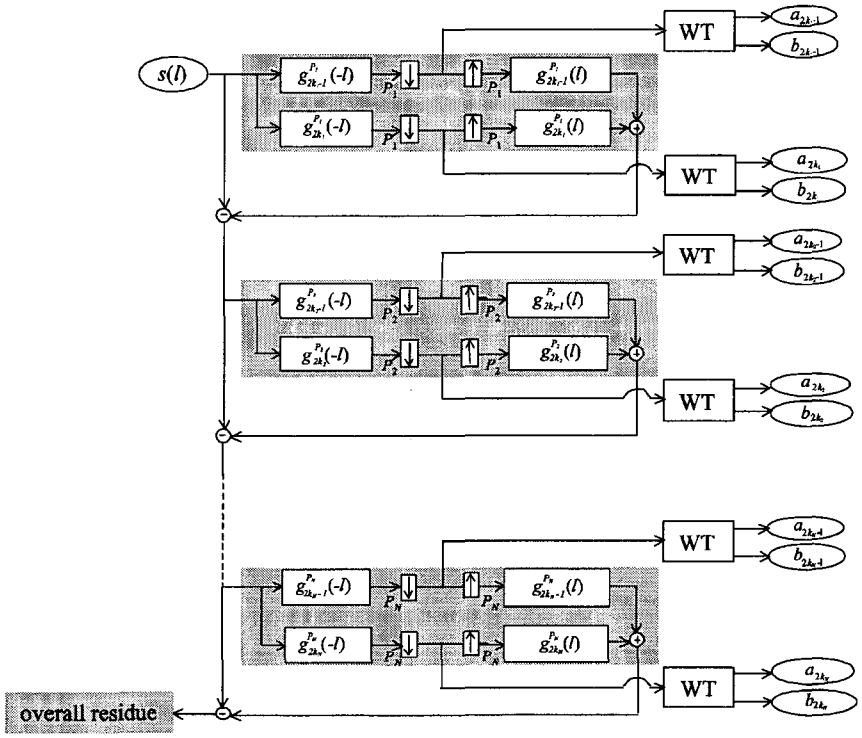


Figure 5.25: Analysis scheme. The index P_n refers to the P_n-channel filter bank chosen to analyze the n^{th} partial, $n = 1, \ldots, N$. The indexes k_n refers to the couple of filters selected from the P_n-channel FB "embracing" the partial peak. 'WT' denotes a wavelet transform block.

ficients of the analysis of the first partial (c). From the last figure it is evident how the scale coefficients form smooth and slowly oscillating curves as in the harmonic case. The pseudo-sinusoidal model can thus be successfully applied even in the inharmonic case. The stochastic model for the $1/f$-shaped sidebands of the partials holds from both a numerical and a listening point of view.

5.2.3 Experimental results

We applied our method successfully to instruments with different degrees of inharmonicity, ranging from very inharmonic sounds to quasi-harmonic sounds: gongs, tympani, tubular bells, and a piano. All sounds have been partially reconstructed (without the residue) by means of the analysis coefficients and resynthesized by means of parametrically controlled coefficients. In some cases (gong and tubular bells) the synthetic sounds are hardly distinguishable from the original ones. Furthermore, the deterministic part and the different wavelet scale (stochastic) components were synthesized separately.

As illustrated in Section 4.5 for the harmonic case, by taking into account psychoacoustic criteria, compression ratios of the order of 1/30 can be easily obtained. This result is achieved only by means of a parametric representation of sounds, i.e. without any quantization and coding optimization. In the inharmonic case one has also to take into account the parameters P_n and k_n. If the sound spectrum is sufficiently stable in time, then the number of parameters increases of $2N$ with respect to the harmonic case. In the example of the gong we considered 35 partials in order to have a sufficiently good result. In this case we needed only 70 additional parameters for encoding a sound of 350,000 samples.

The inharmonic extension also enlarges FAS possibilities as a new synthesis technique. The idea of FAS as a sort of augmented additive synthesis is confirmed and consolidated. We can add an arbitrary number of partials, arbitrarily distributed in the frequency range. Furthermore, we can parametrically control the shape of the partial sidebands, i.e. we can control the amount of timbre dynamics and noisiness. A possible 'minimal parameter scenario' for a FAS module could be to have two parameters per partial: one for the amplitude and one for the 'noisiness', where the latter parameter would control the slope of both spectral sidebands of the partial. Inner parameters (also editable) could be the amplitude envelopes, the phase of the complex scale coefficients, the central frequencies and bandwidths of the inharmonic CMFB. In the next section we describe a more articulated module realized in Pd (Pure data), a software environment for audio processing in real-time.

Digital audio effects as pitch shifting, time stretching, noise-to-harmonic component ratio modifications are easily obtainable by means of interpolation and modulations of the parameters controlling the resynthesis coefficients generation. As for the harmonic case, the most interesting feature is the possibility of processing the stochastic and the deterministic components separately by means of two distinct and appropriate models. The time-stretched and pitch-shifted samples of the gong and of the tubular bell, which we reproduced, sound very realistic. Another effect we implemented was to transform a inharmonic sound into a harmonic one, i.e. to keep the partials and their natural behavior and resynthesize them by means of an artificial harmonic CMFB. We realized different versions of harmonized tympani and gongs, with different noisy sidebands

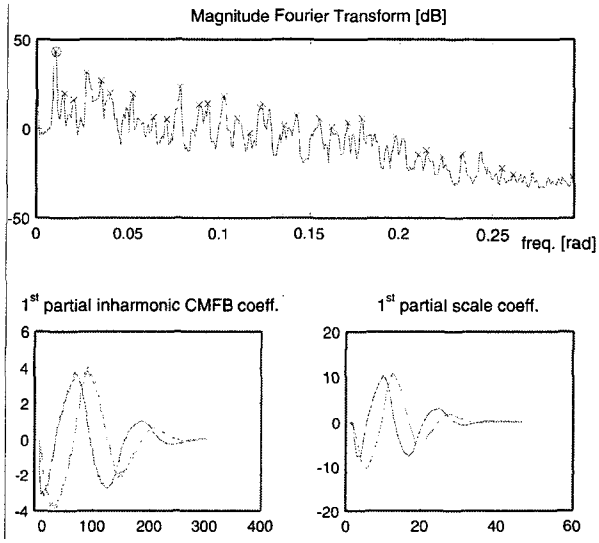


Figure 5.26: a) Magnitude Fourier transform of a gong. The 'x's denote the detected partials. b) The output of the two channels of the inharmonic CMFB corresponding to the first partials (the circled peak of figure a). c) The scale coefficients resulting from the wavelet analysis of the coefficients of figure b.

widths. In conclusion, FAS can be also seen as a flexible sound processor allowing one to manipulate the spectrum of a sound in a perceptually meaningful way.

5.3 Real-time implementation of the method

A real-time (RT) implementation of the synthesis section of FAS was developed by means of Pure data (Pd) [66], [65], a software environment for RT sound processing, running under Linux and Windows operating systems. Pd is based on a minimal graphical interface for visual programming. Data flows are represented by means of strings interconnecting one block to the next one. Figures 5.27, 5.28 and 5.29 give an example of how Pd patches look like. The DSP engine is the computer CPU itself.

We had to implement the IWT and IMDCT blocks. No multirate processing is implemented in Pd. However, a control rate allows one to perform data processing at lower rates than the current sampling rate. Therefore, in order to implement the multirate HBWT scheme, we implemented the whole WT section as if it were a control section, i.e. working at control rate. Having at disposal the IWT and IMDCT blocks, it was possible to develop a complete FAS module for RT generation and control of the resynthesis parameters for the general case of inharmonic sounds. Figure 5.27 shows the main control panel of the resynthesis process. The light grey sliders control the subband energies of the left and right sidebands of all of the partials. We considered 3 scale levels analysis and resynthesis. The 2 dark grey sliders in the center control the global energy of

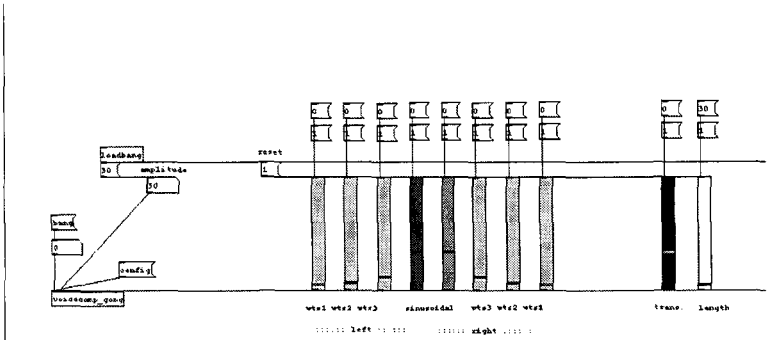


Figure 5.27: Main interface for playing the FAS in RT. The levels of the faders reproduce the $\frac{1}{|f-f_n|}$ spectral behavior around the partials f_n, $n = 1, ..., N$.

the harmonic components. The 2 sliders at the far right of the picture control the position of the crossfade between the original transient and the synthetic sound and the RT time stretching, respectively.

Figure 5.28 represents the subpatch for a single partial. The set of sliders has the same function as those of the main interface, but they are applied to the single partial. These sliders are directly controlled by the sliders in the main interface. Furthermore, the time envelopes of the harmonic components and of the noisy subbands are graphically displayed. These plots are editable by means of the mouse. Also, an editable display for the phase of the harmonic part is available. Figure 5.29 shows how it is possible to edit the envelopes. The same scheme is repeated for all of the implemented partials. On a Pentium III processor with a 1GHz clock rate it is possible to implement a RT synthesis of at least 30 partials.

The analysis is not implemented in RT. It is possible to load the parameters of any previously performed FAS analysis and then run the RT synthesis and processing. The analysis parameters are loaded in the following format per each partial i: Pitch P_i, harmonic k_i, complex HB scale coefficients C_{k_i} and ϕ_{k_i}, $LPCcoef_{k_i,n}$ for the 3 subband levels and distinct sets $Ecoef_{p_i,n}$ for the 6 subbands.

As further developments of the FAS Pd module, we envisage to make the parameters P_i controllable in RT in order to have a RT pitch shifter. Another objective is to implement RT analysis and parameter extraction of an input sound. This would provide a full RT sound processor based on a powerful and articulated spectral modeling technique as that provided by FAS. At the same time we would have a RT implementation of what is the main goal of our thesis, i.e. a prototype of a low rate coding system for high quality audio transmission and reproduction.

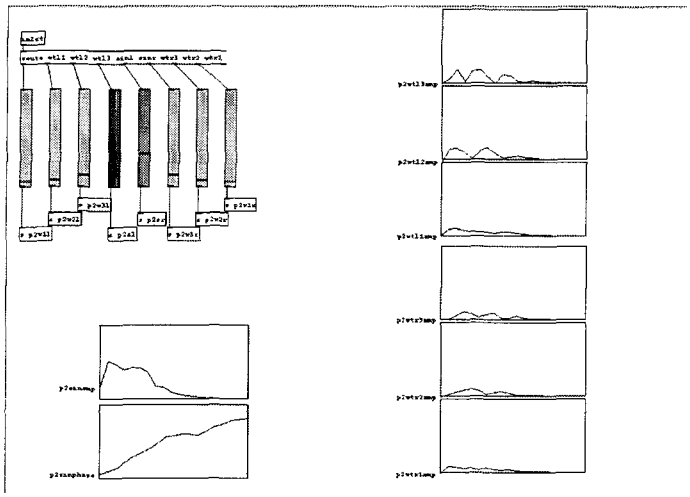


Figure 5.28: Graphical editing for each partial.

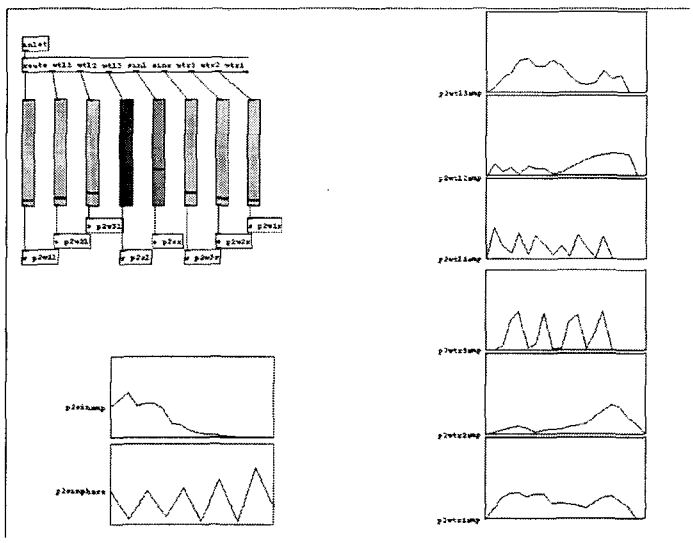


Figure 5.29: Same partial as in Figure 5.28. The envelopes have been edited graphically.

5.4 Summary

In this chapter, we described an extension of FAS for pitch-synchronous analysis and synthesis of voiced sound. FAS allows one to control and reproduce the microfluctuations present in real-life sounds so important from a perceptual point of view in order to perceive a sound as a natural one. By means of the filter design technique introduced in Section 5.1.2 we extended the FAS method to the more general case of voiced sounds with time-varying pitch. More precisely, we adapted a new type of time-varying CMFB to the HBWT scheme, obtaining what we called the Pitch-Synchronous Harmonic-Band Wavelet Transform (PS-HBWT). We showed how the FAS parametric model can be extended to PS-HBWT analysis/synthesis scheme. The most appealing feature of this method is that it allows one to reproduce both the deterministic and the stochastic components of voiced sounds with varying pitch by means of a very restricted number of parameters.

An extension of the FAS to the case of inharmonic sounds has been introduced. The new method is not PR. Nevertheless, an almost alias free reconstruction of sounds can be achieved. Analysis and resynthesis by means of the inharmonic FAS produce very good results for various percussion instruments. We obtained excellent results even for sounds that are generally difficult to model, such as gongs. However, further refinements are necessary in order to obtain equally good results for a wider and more general class of instruments with inharmonic spectra. As in the harmonic case, not only the deterministic components of sounds are modeled and reproduced, but also the noisy components, which are important in order to perceive a sound as realistic.

The method provides great flexibility in terms of sound processing. Experiments on the modifications of the synthesis parameters show that there are interesting applications in the fields of sound synthesis and digital audio effects for electronic music.

Finally a real-time software implementation of FAS was realized and illustrated.

Chapter 6

Conclusions and Future Work

The goal of this work has been the definition of a complete synthesis by analysis method for sound, providing two integrated models for representing both the deterministic and the stochastic components of sound.

In Chapter 3 we introduced a general model for describing the behavior of voiced sounds, i.e. of sounds whose spectra present harmonically-spaced peaks. These peaks represent the deterministic components of sound, while the sidebands in between two peaks correspond to the noisy/stochastic components of sound. Our work was based on the assumption, experimentally verified, that the energy distribution of the sidebands of voiced-sounds is approximately shaped as powers of the inverse of the distance from the closest partial. In other words we assumed that the spectral behavior in the neighborhood of the partial is of the kind $1/(f - f_k)$, where f_k is the k^{th} harmonic. The fact that the sidebands have an approximately $1/f$ behavior around the harmonic peak recalls the fractal/self-similar characteristics of the $1/f$-noise. Our idea is that the $1/(f - f_k)$-like spectral behavior is related to a sort of band-limited self-similar property. This is what we called the pseudo-periodic $1/f$-like model and self similarity is the reason for the name Fractal Additive Synthesis (FAS), selected for our technique. In Chapter 3 we also introduced the Harmonic-Band Wavelet Transform (HBWT) and we designed and implemented a scheme for the analysis and the resynthesis of voiced sounds by means of the HBWT. The HBWT is a harmonic extension of the ordinary Wavelet Transform (WT). Due to its time-scale character, the WT provides a natural tool for the analysis and synthesis of signals with $1/f$-like spectra. Similarly, the HBWT is a natural tool for extracting, decomposing into subbands and resynthesizing the noisy $1/(f - f_n)$-like sidebands of the harmonic spectral peaks.

In Chapter 4 we introduced two distinct models for the HBWT coefficients of the two types of components of voiced sounds, defining FAS as a complete method for the analysis, encoding, decoding and resynthesis of voiced sounds. In other words we defined a complete model for the representation of both the deterministic and stochastic components of voiced sounds in terms of high-level/perceptually-relevant parametric representation. The method provides an efficient data compression tool, while maintaining a high quality sound reproduction. In section 4.5 we discussed the compression results taking into consideration psychoacoustic criteria as well. Perceptually transparent audio coding with compression ratios of the order of 30:1 are easily attainable.

In chapter 5 we extended FAS in order to obtain a method usable for a sufficiently large class of instrumental sounds. The achieved flexibility allows FAS to deal with voiced sounds with varying pitch and with sounds with inharmonic spectra. A prototype of a real-time implementation of the decoding and resytnhesis section of FAS was also realized and illustrated in Section 5.3. A complete real-time implementation of FAS including analysis and parameter extraction is still under construction.

Refinements of the FAS technique are still necessary as well as improvements and optimization of data compression results. Redundancies in the parameters such as those present in Table 4.2 could be avoided by means of a deeper psychoacoustic evaluation of their effects. Moreover, the inharmonic extension of FAS needs improvements in the filter design in order to reach homogeneous results on a larger class of percussion instruments.

Finally, as already mentioned, FAS does not concern transients. The attack transient of the sound is a fundamental element for timbre perception. Our ear is extremely sensitive to transient stimula. Due to their non-stationarity and short duration modeling transients is not an easy task. There are various interesting research directions towards a definition of an effective transient modeling technique, as the DCT domain transient representation introduced in [94] [92] [91] or the Exponentially Damped Sinusoidal (EDS) model [8] [9] [1]. Other possibilities of investigation are provided by matching pursuit modeling [31] [32]. A high quality model for transients, integrated with FAS, would provide a global method for voiced and inharmonic sound representation. Also, it would supply a model for those sounds, whose main content is given by transients. This class of sounds contains staccato sounds, pizzicatos in string instruments and fast percussive sounds, such as castanettes.

In conclusion, we have defined, implemented and tested a method for the analysis, coding, transmission and resynthesis of high quality audio at low bit rate, which we called FAS. This method has been designed in the perspective of MPEG-4 Structured Audio approach to audio and music coding. With respect to the available synthesis algorithms, FAS claims to be a transparent coding method for a large class of sounds. Furthermore, FAS provides a powerful tool for sound synthesis and sound processing in the sense of digital audio effects.

Bibliography

[1] R. Badeau, R. Boyer, and B. David. Eds parametric modeling and tracking of audio signals. *Proceedings of the DAFx-02 Digital Audio Effects Workshop*, pages 139–144, 2002.

[2] J. A. Barnes and D. W. Allan. A statistical model of flicker noise. *Proc. IEEE*, 54:176–178, February 1966.

[3] M. Barnsley. *Fractals Everywhere*. Academic Press Inc., 1990.

[4] R. J. Barton and V. H. Poor. Fractional brownian motion, fractional noisees and applications. *IEEE Trans. Inform. Theory*, 34:943–959, September 1988.

[5] H. Berlioz. *Instrumentationslehre. Ergänzt und revidiert von R. Strauss.* Peters, Leipzig, 1904.

[6] J. Bernamont. *Ann. Phys.*, 71(7), 1937.

[7] T. Blu. Iterated filter banks with rational rate changes connection with discrete wavelet transform. *IEEE Trans. Signal Processing*, 41:3232Ű–3244, December 1993.

[8] R. Boyer and K. Abed-Meraim. Efficient parametric modeling for audio transients. *Proceedings of the DAFx-02 Digital Audio Effects Workshop*, pages 97–100, 2002.

[9] R. Boyer and J. Rosier. Iterative method for harmonic and exponentially damped sinusoidal models. *Proceedings of the DAFx-02 Digital Audio Effects Workshop*, pages 145–150, 2002.

[10] J. Chowning. The synthesis of complex audio spectra by means of frequency modulation. *Journal of the Audio Engineering Society*, 21(7):526–534, 1973.

[11] A. Cohen and J. Kovacevic. Wavelets: The mathematical background. *Proc. IEEE*, 84(4):514–522, 1996.

[12] P. Cosi, G. De Poli, and G. Lauzzana. Auditory modelling and self-organizing neural networks for timbre classification. *Journal of New Music Research*, 23(1), March 1994.

[13] I. Daubechies. Ten lectures on wavelets. *SIAM CBMS series*, 1992.

[14] Ph. Depalle, G. Garcia, and X. Rodet. Analysis of sound for additive synthesis: tracking of partials using hidden markov models. *Proceedings of the International Computer Music Conference*, pages 94–Ú97, 1993.

[15] Ph. Depalle and T. Helie. Extraction of spectral parameters using a short-time fourier transform modeling and no sidelobe windows. *Proceedings of the 1997 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 228Ú–231, 1997.

[16] M. Deriche and A.H. Tewfik. Signal modeling with filtered discrete fractional noise processes. *Signal Processing, IEEE Transactions on*, 41(9):2839–2849, September 1993.

[17] G. Evangelista. Wavelet transforms that we can play. In A. Piccialli, G. De Poli, and C. Roads, editors, *Representations of Musical Signals*, pages 119–136. MIT Press, Cambridge, Ma, 1991.

[18] G. Evangelista. Pitch synchronous wavelet representations of speech and music signals. *IEEE Trans. on Signal Processing Special Issue on Wavelets and Signal Processing*, 41(12):3313–3330, December 1993.

[19] G. Evangelista. Comb and multiplexed wavelet transforms and their applications to signal processing. *IEEE Trans. on Signal Processing*, 42(2):292–303, February 1994.

[20] G. Evangelista and S. Cavaliere. Discrete frequency warped wavelets: Theory and applications. *IEEE Transaction on Signal Processing*, 46(4):874–885, April 1998.

[21] G. Evangelista and S. Cavaliere. Frequency-warped filter banks and wavelet transforms: A discrete time approach via laguerre expansion. *IEEE Transaction on Signal Processing*, 46(10):2638–2650, October 1998.

[22] G. Evangelista and S. Cavaliere. Time-varying frequency warping: results and experiments. *Proceedings of the DAFx-99 Digital Audio Effects Workshop*, pages 13–16, 1999.

[23] G. Evangelista and S. Cavaliere. Audio effects based on biorthogonal time-varying frequency warping. *EURASIP Journal on Applied Signal Processing*, 2001(1):27–35, March 2001.

[24] K. Fitz, L. Haken, S. Lefvert, and M. OŠDonnell. Sound morphing using loris and the reassigned bandwdith-enhanced additive sound model:practice and applications. *ICMC 02 Proceeding*, 2002.

[25] J. L. Flanagan. *Speech Analysis Synthesis and Perception.* Springer-Verlag, New York, 1972.

[26] P. Flandrin. Wavelet analysis and synthesis of fractional brownian motion. *Information Theory, IEEE Transactions on*, 38(2):910–917, March 1992.

[27] D. Gabor. Theory of communication. *Journal IEE*, 93:429–457, 1946.

[28] D. T. Gillespie. The mathematics of brownian motion and johnson noise. *Am. J. Phys.*, 64:225–240, March 1996.

[29] M. Goodwin and M. Vetterli. Residual modeling in music analysis-synthesis. *Acoustics, Speech, and Signal Processing, ICASSP-96 1996 IEEE International Conference on. Conference Proceedings.*, 2:1005–1008, 1996.

[30] M. Goodwin and M. Vetterli. Time-frequency signal models for music analysis, transformation, and synthesis. *Time-Frequency and Time-Scale Analysis, 1996., Proceedings of the IEEE-SP International Symposium on*, pages 133–136, 1996.

[31] M. Goodwin and M. Vetterli. Atomic decompositions of audio signals. *Applications of Signal Processing to Audio and Acoustics, IEEE ASSP Workshop on*, 1997.

[32] M. Goodwin and M. Vetterli. Matching pursuit and atomic signal models based on recursive filter banks. *Signal Processing, IEEE Transactions on*, 47(7):1890–1902, July 1999.

[33] J. M. Grey. Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, 61(5):1270–1277, 1976.

[34] M. Kahrs and K. Brandenburg. *Applicatons of Digital Signal Processing to Audio and Acoustic.* Kluwer Academic Publisher, Boston, 1998.

[35] O. Keiji. Nono's prometeo and akiyoshidai international art village. *Nagata Acoustics News, Nagata Acoustics Inc., Tokyo, Japan*, 98(9):128–142, September 1998.

[36] M. S. Keshner. 1/f noise. *Proc. IEEE*, 70(3):212–218, March 1982.

[37] J. Kovacevic and I. Daubechies Eds. *Proceedings of the IEEE, Special Issue on Wavelets*, 84(4), 1996.

[38] J. Kovacevic and M. Vetterli. Perfect reconstruction filter banks with arbitrary rational sampling rates. *IEEE Transactions on Signal Processing*, 41(6):2047–2066, June 1993.

[39] R. Kronland-Martinet. The wavelet transform for analysis synthesis and processing of speech and music sounds. *Computer Music Journal*, 12(4):11–20, 1988.

[40] P. E. Kudumakis and M. B. Sandler. Synthesis of audio signals using the wavelet transform. *IEE Colloquium on Audio DSP - Circuits and Systems*, pages 41–45, 1993.

[41] H. Lachenmann, A. Isozaki, A. Asada, and S. Choki. Luigi nono and prometeo. *InterCommunication, NTT Publishing Co., Tokyo, Japan*, 27:128–142, 1999.

[42] B. Leslie and M. B. Sandler. Audio compression using wavelets. *IEE Colloquium on Audio and Music Technology: The Challenge of Creative DSP*, 2:1–7, 1998.

[43] G. Ligeti. *Artikulation, for tape.* B. Schott's Soehne, Mainz, 1958.

[44] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63:561Ű–580, April 1975.

[45] J. Makhoul and J. J. Wolf. Linear prediction and the spectral analysis of speech. Technical Report 2800, Bolt Beranek and Newman Inc., Cambridge, Mass., April 1974.

[46] S. Mallat. Multiresolution approximations and wavelet orthonormal bases of l2(r). *Trans. Amer. Math. Soc.*, 315:69–87, 1992.

[47] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, Cambridge Ma, 1997.

[48] B. B. Mandelbrot. *Multifractals and 1/f noise.* Springer Verlag, New York, 1999.

[49] B. B. Mandelbrot and H. W. Van Ness. Fractional brownian motion, fractional noises and applications. *SIAM Rev.*, 10:422–436, October 1968.

[50] M. V. Mathews. The digital computer as an instrument. *Science*, 142, 1963.

[51] M. V. Mathews and J. R. Pierce. *Current Directions in Computer Music Research.* M.I.T. Press, Cambridge, MA, 1989.

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

[53] T. Q. Nguyen and R. D. Koilpillai. The theory and design of arbitrary-length cosine-modulated filter banks and wavelets, satisfying perfect reconstruction. *IEEE Trans. on Signal Processing*, 44(3):473–483, March 1996.

[54] L. Nono. *Das atmende Klarsein, for small choir, bass flute, live electronics and tape.* Ed. Ricordi, Milano, 1983.

[55] L. Nono. *Scritti e colloqui.* Ed. Ricordi, Milano, 2001.

[56] N. Osaaka. Timbre interpolation of sounds using a sinusoidal model. *ICMC 95 Proceeding*, 1995.

[57] T. Painter and A. Spanias. Perceptual coding of digital audio. *Proc. IEEE*, 88(4):441–513, April 2000.

[58] T. Parson. *Voice and Speech Processing.* McGraw-Hill, New York, 1986.

[59] W. J. Pielemeier, G. H. Wakefield, and M. H. Simoni. Time-frequency analysis of musical signals. *Proceedings of the IEEE*, 84(9):1216–1230, April 1996.

[60] J. R. Pierce. *The science of musical sound.* Scientific American Books, Paris, 1983.

[61] P. Polotti and G. Evangelista. Harmonic-band wavelet coefficient modeling for pseudo-periodic sounds processing. *DAFx-00 Proceedings*, pages 103–108, December 2000.

[62] P. Polotti and G. Evangelista. Analysis and synthesis of pseudo-periodic 1/f-like noise by means of wavelets with applications to digital audio. *EURASIP Journal on Applied Signal Processing*, 2001(1):1–14, March 2001.

[63] P. Polotti and G. Evangelista. Fractal additive synthesis by means of harmonic-band wavelets. *Computer Music Journal*, 25(3):22–37, 2001.

[64] P. Polotti and G. Evangelista. Multiresolution sinusoidal/stochastic model for voiced-sounds. *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-01)*, December 2001.

[65] M. Puckette. Pure data. *Proceedings, International Computer Music Conference*, pages 269–272, 1996.

[66] M. Puckette. Pure data: another integrated computer music environment. *Proceedings, Second Intercollege Computer Music Concerts*, pages 37–41, 1996.

[67] L. R. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*, volume I. Prentice Hall, New York, 1993.

[68] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals.* Prentice Hall, 1978.

[69] O. Rioul and M. Vetterli. Wavelets and signal processing. *IEEE Signal Processing Magazine*, 84:14–38, 1991.

[70] J. C. Risset and M. V. Matthews. Analysis of musical-instrument tones. *Physics Today*, 22(2):23–30, 1969.

[71] C. Roads. *The Computer Music Tutorial.* MIT press, Cambridge Ma, 1996.

[72] X. Rodet and P. Depalle. A new additive synthesis method using inverse fourier transform and spectral envelopes. *Proceedings of ICMC*, pages 265–272, May 1992.

[73] X. Rodet and P. Depalle. Spectral envelopes and inverse fft synthesis. *Proceedings of the AES conference 1992*, 1992.

[74] Xavier Rodet. Musical sound signal analysis/synthesis: Sinusoidal+residual and elementary waveform models. *IEEE Time-Frequency and Time-Scale Workshop, Coventry, Grande Bretagne*, 1997.

[75] A. E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *Journal of Acoustic Society of America*, 49(2):583Ű–590, February 1971.

[76] P. Schaeffer. *Traite' des objets musicaux.* Seuil, Paris, 1966.

[77] P. Schaeffer. *La musique concrete.* Presses universitaires de France, Paris, 1967.

[78] G. Schuller. Time-varying filter banks with variable system delay. *Acoustics, Speech, and Signal Processing 1997 IEEE International Conference on, ICASSP-97.*, 3:2469–2472, August 1997.

[79] G. D. T. Schuller and T. Karp. Modulated filter banks with arbitrary system delay: Efficient implementations and the time-varying case. *IEEE Transaction on Signal Processing*, 48(63), March 2000.

[80] G.D.T. Schuller and M.J.T. Smith. New framework for modulated perfect reconstruction filter banks. *Signal Processing, IEEE Transactions on*, 44(8):1941–1954, August 1996.

[81] X. Serra. Musical sound modeling with sinusoids plus noise. In A. Piccialli, G. De Poli, C. Roads, and S. T. Pope, editors, *Musical Signal Processing.* Swets and Zeitlinger, Amsterdam, 1993.

[82] X. Serra. Sound hybridization techniques based on a deterministic plus stochastic decomposition model. *Proceedings of the International Computer Music Conference*, pages 384Ű–351, 1994.

[83] X. Serra and J. Smith. Parshl: an analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation. *Proceedings of the International Computer Music Conference*, pages 290Ű–297, 1987.

[84] X. Serra and J. O. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):14–24, 1990.

[85] B. M. R. Shankar and A. Makur. Allpass delay chain-based iir pr filterbank and its application to multiple description subband coding. *IEEE Transaction on Signal Processing*, 50(4):814–823, April 2002.

[86] P. Stoica and R. Moses. *Introduction to Spectral Analysis.* Prentice Hall, Upper Saddle River, New Jersey, 1997.

[87] G. Strang and T. Nguyen. *Wavelets and Filter Banks.* Cambridge Press, Wellesley, MA, 1996.

[88] A.H. Tewfik and M. Kim. Correlation structure of the discrete wavelet coefficients of fractional brownian motion. *Information Theory, IEEE Transactions on*, 38(2):904–909, March 1992.

[89] P.P. Vaidyanathan. *Multirate Systems and Filter Banks.* Prentice Hall Signal Processing Series, Upper Saddle River NJ, 1993.

[90] E. Varese. *Ionisation, for Percussion Ensemble of 13 Players.* Colfranc Music Publishing Corporation, New York, 1934.

[91] T. Verma, S. Levine, and T. Meng. Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals. *1997 International Computer Music Conference (ICMC'97) Proceedings of, Thessaloniki, Greece*, pages 164–167, September 1997.

[92] T. Verma and T. Meng. An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio. *1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-98) Proceedings of*, May 1998.

[93] T. Verma and T. Meng. Sinusoidal modeling using frame-based perceptually weighted matching pursuits. *Acoustics, Speech, and Signal Processing, 1999 IEEE International Conference on. Conference Proceedings.*, 2:981–984, March 1999.

[94] T. Verma and T. Meng. Extending spectral modeling synthesis with transient modeling synthesis. *Computer Music Journal*, 24(2):47–59, 2000.

[95] T.S. Verma and T.H.Y. Meng. Time scale modification using a sines+transients+noise signal model. *Proceedings DAFx-98 Digital Audio Effects Workshop*, pages 49Ü–52, 1998.

[96] M. Vetterli and J. Kovacevic. *Wavelets and Subband Coding.* Prentice Hall, 1995.

[97] A. Vidolin. Il suono mobile. In La Biennale di Venezia, editor, *Con Luigi Nono*, pages 42–47. ed. Ricordi, Milano, 1993.

[98] R. F. Voss and J. Clarke. 1/f noise in music: Music from 1/f noise. *J. Acoust. Soc. Am.*, 63(1), 1978.

[99] G. W. Wornell. Wavelet-based representations for the 1/f family of fractal processes. *Proc. IEEE*, 81(10):1428–1450, October 1993.

[100] G. W. Wornell and A. V. Oppenheim. Wavelet-based representations for a class of self-similar signals with applications to fractal modulation. *IEEE Trans. Inform. Theory*, 38(2):785–800, March 1992.

[101] A.L. Mc Worther. 1/f noise and related surface effects in germanium. *Report of the Massachusetts Institute of Tecnology*, (89), 1955.

[102] S. Yadegari. Using self-similarity for sound/music synthesis. *Proceedings of ICMC, Montreal, International Computer Music Association, San Francisco*, pages 423–424, 1991.

[103] U. Zoelzer. *Digital Audio Signal Processing.* J. Wiley and Sons, Chichester, 1997.

PIETRO POLOTTI

Professional experience

March 2001 - Now: Assistant professor, teaching digital signal processing at the Conservatory G. Tartini, Trieste, Italy.
November 1999 - Now: Research assistant at the EPFL (École Polytechnique Fédérale de Lausanne), Switzerland, with the Audio-Visual Communications Laboratory (LCAV).
October 1999 - Now: Teaching electronic music composition at the LaSDIM (Laboratory for Experimentation and Didactics of Computer Music), Sezione di Musica Contemporanea, Civica Scuola di Musica di Milano, Italy.

Scientific education

2002: PhD in Communication Systems, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland.
1999: Doctoral school in Communication Systems, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland.
1997: Laurea degree in Physics, Università degli Studi, Trieste, Italy.

Musical education

1998: Diploma in Electronic Music at the Conservatory B. Marcello, Venice, Italy. Supervisor professor A. Vidolin.
1993: Diploma in Composition at the Conservatory G. Verdi, Milan, Italy. Supervisor professor A. Corghi.
1989: Diploma in Piano at the Conservatory G. Tartini, Trieste, Italy.
1988: First grade exam of Clarinet (5th year) at the Conservatory G. Tartini, Trieste, Italy.

Publications

Journals and Manuscripts

- P. Polotti, G. Evangelista, "Fractal Additive Synthesis: a Sinusoidal/Stochastic Model for Voiced Sounds". To be submitted. *IEEE Transactions on Speech and Audio Processing.*
- P. Polotti, G. Evangelista, "Fractal Additive Synthesis by means of Harmonic-Band Wavelets", *Computer Music Journal*, 25(3), pp. 22-37, Fall 2001.

- P. Polotti, G. Evangelista, "Analysis and Synthesis of Pseudo-Periodic 1/f-like Noise by means of Wavelets with Applications to Digital Audio", *EURASIP Journal on Applied Signal Processing*, Hindawi Publishing Corporation, Vol. 1, pp. 1-14, March 20

Conferences

- P. Polotti, G. Evangelista, "Inharmonic Sound Spectral Modelling by Means of Fractal Additive Synthesis", *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, Sep. 2002.
- P. Polotti, "Fractal Additive Synthesis: A Pitch-Synchronous Extension of the Method for the Analysis and Synthesis of Natural Voiced-Sounds", *Proceedings of the ICMC 2002*, Göteborg, Sweden, Aug. 2002.
- P. Polotti, G. Evangelista, "Multiresolution Sinusoidal/Stochastic Model for Voiced-Sounds", *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-01)*, Limerick, Ireland, Dec. 2001.
- P. Polotti, G. Evangelista, "Harmonic-Band Wavelet Coefficient Modeling for Pseudo-Periodic Sound Processing", *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-00)*, Verona, Italy, Dec. 2000.
- P. Polotti, G. Evangelista "Sound Modeling by means of Harmonic-Band Wavelets: New Results and Experiments", *Proc. of XIII CIM*, pp. 43-46, L'Aquila, Italy, Sept. 2000.
- P. Polotti, G. Evangelista, "Time-Spectral Modeling of Sounds by Means of Harmonic-Band Wavelets", *Proceedings of the ICMC 2000*, pp. 388-391, Berlin, Germany, Aug. 2000.
- P. Polotti, G. Evangelista, "Dynamic Models of Pseudo-Periodicity", *Proc. of DAFx99*, Trondheim, Norway, Dec. 1999.
- G. Evangelista, P. Polotti, " Analysis and Synthesis of Pseudoperiodic $1/f$ -like noise by means of Multiband Wavelets ", *Proceedings of CIM 98*, 12th international meeting of computer music 1998, Gorizia, Italy.

Compositions

Composition of several chamber music and electronic music pieces performed in various contemporary music festivals and competitions:

- *1948* for tape solo (1998), performed in the Festival "Musica e Scienza", Rome, Italy (20/6/01), and in the IX International Electroacoustic Music Festival, "Primavera en La Habana", (5/3/02), La Habana, Cuba.
- *A-Tom* for tape solo (1998) performed in the Festival "La Terra Fertile", L'Aquila, Italy (4/9/98) and in the International Festival of Trento and Rovereto, Italy (19/10/1998). Edited on CD by Ars Publica, 1998.
- *Intrecci* for tape solo (1994), performed on the 22/2/94 at the Auditorium Fenzi, Conegliano Veneto, Italy, in a concert organised by the Teatro la Fenice of Venice.
- *Permutazioni auree* for five flutes (1991), winner of the third prize at the composition competition Castello di Belveglio, Asti, Italy (1991).
- *Pèntacha* for voice, clarinet, violin, cello and piano (1988), broadcast on the 24/6/90 by the RAI 2nd channel.

Other research experiences

October-November 2001: Host researcher at the MTG (Music Technology Group) of the Universitat Pompeu Fabra, Barcelona, Spain.
Collaboration to a project for "A New Musical Interface Using Acoustic Tap Tracking". A prototype of interface was presented at the *Mosart Workshop on Current Research Directions in Computer Music*, Barcelona, November 15-17, 2001.

[92] T. Verma and T. Meng. An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio. *1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-98) Proceedings of*, May 1998.

[93] T. Verma and T. Meng. Sinusoidal modeling using frame-based perceptually weighted matching pursuits. *Acoustics, Speech, and Signal Processing, 1999 IEEE International Conference on. Conference Proceedings.*, 2:981–984, March 1999.

[94] T. Verma and T. Meng. Extending spectral modeling synthesis with transient modeling synthesis. *Computer Music Journal*, 24(2):47–59, 2000.

[95] T.S. Verma and T.H.Y. Meng. Time scale modification using a sines+transients+noise signal model. *Proceedings DAFx-98 Digital Audio Effects Workshop*, pages 49Ű–52, 1998.

[96] M. Vetterli and J. Kovacevic. *Wavelets and Subband Coding.* Prentice Hall, 1995.

[97] A. Vidolin. Il suono mobile. In La Biennale di Venezia, editor, *Con Luigi Nono*, pages 42–47. ed. Ricordi, Milano, 1993.

[98] R. F. Voss and J. Clarke. 1/f noise in music: Music from 1/f noise. *J. Acoust. Soc. Am.*, 63(1), 1978.

[99] G. W. Wornell. Wavelet-based representations for the 1/f family of fractal processes. *Proc. IEEE*, 81(10):1428–1450, October 1993.

[100] G. W. Wornell and A. V. Oppenheim. Wavelet-based representations for a class of self-similar signals with applications to fractal modulation. *IEEE Trans. Inform. Theory*, 38(2):785–800, March 1992.

[101] A.L. Mc Worther. 1/f noise and related surface effects in germanium. *Report of the Massachusetts Institute of Tecnology*, (89), 1955.

[102] S. Yadegari. Using self-similarity for sound/music synthesis. *Proceedings of ICMC, Montreal, International Computer Music Association, San Francisco*, pages 423–424, 1991.

[103] U. Zoelzer. *Digital Audio Signal Processing.* J. Wiley and Sons, Chichester, 1997.

PIETRO POLOTTI

Professional experience

March 2001 - Now: Assistant professor, teaching digital signal processing at the Conservatory G. Tartini, Trieste, Italy.
November 1999 - Now: Research assistant at the EPFL (École Polytechnique Fédérale de Lausanne), Switzerland, with the Audio-Visual Communications Laboratory (LCAV).
October 1999 - Now: Teaching electronic music composition at the LaSDIM (Laboratory for Experimentation and Didactics of Computer Music), Sezione di Musica Contemporanea, Civica Scuola di Musica di Milano, Italy.

Scientific education

2002: PhD in Communication Systems, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland.
1999: Doctoral school in Communication Systems, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland.
1997: Laurea degree in Physics, Università degli Studi, Trieste, Italy.

Musical education

1998: Diploma in Electronic Music at the Conservatory B. Marcello, Venice, Italy. Supervisor professor A. Vidolin.
1993: Diploma in Composition at the Conservatory G. Verdi, Milan, Italy. Supervisor professor A. Corghi.
1989: Diploma in Piano at the Conservatory G. Tartini, Trieste, Italy.
1988: First grade exam of Clarinet (5th year) at the Conservatory G. Tartini, Trieste, Italy.

Publications

Journals and Manuscripts

- P. Polotti, G. Evangelista, "Fractal Additive Synthesis: a Sinusoidal/Stochastic Model for Voiced Sounds". To be submitted. *IEEE Transactions on Speech and Audio Processing.*
- P. Polotti, G. Evangelista, "Fractal Additive Synthesis by means of Harmonic-Band Wavelets", *Computer Music Journal*, 25(3), pp. 22-37, Fall 2001.

- P. Polotti, G. Evangelista, "Analysis and Synthesis of Pseudo-Periodic 1/f-like Noise by means of Wavelets with Applications to Digital Audio", *EURASIP Journal on Applied Signal Processing*, Hindawi Publishing Corporation, Vol. 1, pp. 1-14, March 20

Conferences

- P. Polotti, G. Evangelista, "Inharmonic Sound Spectral Modelling by Means of Fractal Additive Synthesis", *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, Sep. 2002.
- P. Polotti, "Fractal Additive Synthesis: A Pitch-Synchronous Extension of the Method for the Analysis and Synthesis of Natural Voiced-Sounds", *Proceedings of the ICMC 2002*, Göteborg, Sweden, Aug. 2002.
- P. Polotti, G. Evangelista, "Multiresolution Sinusoidal/Stochastic Model for Voiced-Sounds", *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-01)*, Limerick, Ireland, Dec. 2001.
- P. Polotti, G. Evangelista, "Harmonic-Band Wavelet Coefficient Modeling for Pseudo-Periodic Sound Processing", *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-00)*, Verona, Italy, Dec. 2000.
- P. Polotti, G. Evangelista "Sound Modeling by means of Harmonic-Band Wavelets: New Results and Experiments", *Proc. of XIII CIM*, pp. 43-46, L'Aquila, Italy, Sept. 2000.
- P. Polotti, G. Evangelista, "Time-Spectral Modeling of Sounds by Means of Harmonic-Band Wavelets", *Proceedings of the ICMC 2000*, pp. 388-391, Berlin, Germany, Aug. 2000.
- P. Polotti, G. Evangelista, "Dynamic Models of Pseudo-Periodicity", *Proc. of DAFx99*, Trondheim, Norway, Dec. 1999.
- G. Evangelista, P. Polotti, " Analysis and Synthesis of Pseudoperiodic $1/f$ -like noise by means of Multiband Wavelets ", *Proceedings of CIM 98*, 12th international meeting of computer music 1998, Gorizia, Italy.

Compositions

Composition of several chamber music and electronic music pieces performed in various contemporary music festivals and competitions:

- *1948* for tape solo (1998), performed in the Festival "Musica e Scienza", Rome, Italy (20/6/01), and in the IX International Electroacoustic Music Festival, "Primavera en La Habana", (5/3/02), La Habana, Cuba.
- *A-Tom* for tape solo (1998) performed in the Festival "La Terra Fertile", L'Aquila, Italy (4/9/98) and in the International Festival of Trento and Rovereto, Italy (19/10/1998). Edited on CD by Ars Publica, 1998.
- *Intrecci* for tape solo (1994), performed on the 22/2/94 at the Auditorium Fenzi, Conegliano Veneto, Italy, in a concert organised by the Teatro la Fenice of Venice.
- *Permutazioni auree* for five flutes (1991), winner of the third prize at the composition competition Castello di Belveglio, Asti, Italy (1991).
- *Pèntacha* for voice, clarinet, violin, cello and piano (1988), broadcast on the 24/6/90 by the RAI 2nd channel.

Other research experiences

October-November 2001: Host researcher at the MTG (Music Technology Group) of the Universitat Pompeu Fabra, Barcelona, Spain.
Collaboration to a project for "A New Musical Interface Using Acoustic Tap Tracking". A prototype of interface was presented at the *Mosart Workshop on Current Research Directions in Computer Music*, Barcelona, November 15-17, 2001.